A survey on datasets for fairness-aware machine learning

Tai Le Quy*[1], Arjun Roy†[12], Vasileios Iosifidis‡[1], Wenbin Zhang§[3], and Eirini Ntoutsi¶[2]

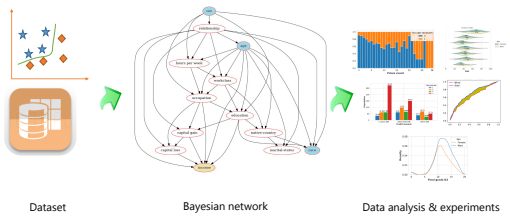
[1]L3S Research Center, Leibniz University Hannover, Germany
[2]Institute of Computer Science, Free University Berlin, Germany
[3]Carnegie Mellon University, United States

Article category: Overview

Conflict of interest: The authors have declared no conflicts of interest for this article.

Abstract

As decision-making increasingly relies on Machine Learning (ML) and (big) data, the issue of fairness in data-driven Artificial Intelligence (AI) systems is receiving increasing attention from both research and industry. A large variety of fairness-aware machine learning solutions have been proposed which involve fairness-related interventions in the data, learning algorithms and/or model outputs. However, a vital part of proposing new approaches is evaluating them empirically on benchmark datasets that represent realistic and diverse settings. Therefore, in this paper, we overview real-world datasets used for fairness-aware machine learning. We focus on tabular data as the most common data representation for fairness-aware machine learning. We start our analysis by identifying relationships between the different attributes, particularly w.r.t. protected attributes and class attribute, using a Bayesian network. For a deeper understanding of bias in the datasets, we investigate the interesting relationships using exploratory analysis.



A workflow of the survey on datasets for fairness-aware machine learning

*Corresponding author. tai@l3s.de 0000-0001-8512-5854
†arjun.roy@fu-berlin.de 0000-0002-4279-9442
‡iosifidis@l3s.de 0000-0002-3005-4507
§wenbinzhang@cmu.edu 0000-0003-3024-5415
¶eirini.ntoutsi@fu-berlin.de 0000-0001-5729-1003

1 Introduction

Artificial Intelligence and Machine Learning are widely employed nowadays by businesses, governments and other organizations to improve their operational quality and assist in decision-making in areas such as loan approval (Mukerjee, Biswas, Deb, & Mathur, 2002), recruiting (Faliagka, Ramantas, Tsakalidis, & Tzimas, 2012), school admission (Moore, 1998), risk prediction (Yeh & Lien, 2009). There are many advantages of using algorithmic decision-making as computers can efficiently analyze large amounts of data with high accuracy. Along with the advantages, unfortunately, there is plenty of evidence regarding the discriminative impact of ML-based decision-making on individuals and groups of people on the basis of *protected attributes* such as gender or race. As an example, *racial-bias* was observed in COMPAS (Angwin, Larson, Mattu, & Kirchner, 2016), a software used by the U.S. courts to assess the risk of recidivism; in particular, it has been found that black defendants were predicted with a higher risk of recidivism than their actual risk compared to white defendants. Another example refers to search algorithms in job search websites; it has been found that such algorithms exhibit *gender-bias* as they display higher-paying jobs to male applicants compared to female ones (Simonite, 2015; Datta, Tschantz, & Datta, 2015).

Data are an essential part of machine learning. Usage of sensitive information during the learning process is undesirable but hard to guarantee even if known protected attributes are omitted from the analysis. The reason is the causal effects (Madras, Creager, Pitassi, & Zemel, 2019) of such attributes, including observable "proxy" attributes. As an example, the non-protected attribute "zip-code" was found to be a proxy for the protected attribute "race" (Datta, Fredrikson, Ko, Mardziel, & Sen, 2017) or the "credit rating" can be used as a proxy for "safe driving" (Warner & Sloan, 2021). Hence, even if the protected attributes like race or gender are not used, the resulting ML models can still be biased (Angwin et al., 2016) due to the causal effects of such attributes. Although methods for detecting proxy attributes exist, e.g., (Yeom, Datta, & Fredrikson, 2018) detects proxies in linear regression models by using a convex optimization procedure, eliminating all the correlated features might drastically reduce the utility of the data for the learning problem.

The domain of bias and fairness in machine learning has attracted much interest in recent years, and as a result, several surveys exist that provide a broad overview of the area, its technical challenges and solutions (Ntoutsi et al., 2020; Mehrabi, Morstatter, Saxena, Lerman, & Galstyan, 2021; Chhabra, Masalkovaitė, & Mohapatra, 2021; Pitoura, Stefanidis, & Koutrika, 2021; Xivuri & Twinomurinzi, 2021). However, an overview of the datasets used for fairness-aware machine learning evaluation is still missing. As data are a vital part of ML and benchmark datasets a decisive factor for the success of AI research[1], we believe our survey is serving to fill a gap in the extant research.

In this survey, we overview the different datasets used in the domain of fairness-aware machine learning, and we characterize them according to their application domain, protected attributes and other learning characteristics like cardinality, dimensionality and class (im)balance. For each dataset, we provide an exploratory analysis by first using a Bayesian network to identify the relationships among attributes. Based on the Bayesian network, we provide a graphical analysis of the attributes for a deeper understanding of bias in the dataset. The Bayesian network illustrates the conditional (in)dependence between the protected attribute(s) and the class attribute; thus, it reduces the space and complexity of data analysis that needs to be performed to discover and

[1] https://qz.com/1034972/the-data-that-changed-the-direction-of-ai-research-and-possibly-the-world/

clarify the fairness-related problems in the dataset. We then focus our exploratory analysis on features having a direct or indirect relationship with the protected attributes. We accompany our exploratory analysis with a quantitative evaluation of measures related to predictive and fairness performance.

We believe that our survey is useful as it gathers many fairness-related datasets scattered around the web and organizes them in terms of different principles (application domain, learning challenges like dimensionality and class imbalance, fairness-aware related challenges like the number of protected attributes, etc.). As such, we except that it will help researchers to easily select the most appropriate datasets for their application domain (e.g., learning analytics vs recidivism), learning challenges (e.g., balanced vs imbalanced classification), classification task (e.g., binary classification vs multiclass learning), fairness-related challenges (e.g., single protected vs multiple protected attributes etc.).

As datasets have played a foundational role in the advancement of machine learning research (Paullada, Raji, Bender, Denton, & Hanna, 2021), our survey also indicates the need for more open benchmark datasets that would reflect different application domains (from education and healthcare to recruitment and logistics), different contexts (e.g., spatial, temporal, etc.), various (machine) learning challenges (dimensionality, imbalance, number of classes, etc.) as well as different notions of fairness (multi-discrimination, temporal fairness, distributional fairness, etc.). We advocate that the community should also pay attention to benchmark datasets in parallel to new methods and algorithms. The area of fairness-aware machine learning will undoubtedly benefit from having benchmark datasets for various tasks.

The rest of the paper is structured as follows: In Section 2, we describe our methodology for dataset collection and evaluation. The most commonly used datasets for fairness are presented in Section 3 together with the results of their exploratory analysis. Section 4 demonstrates a quantitative evaluation of a classification model on the different datasets w.r.t. predictive performance and fairness. We summarize several open issues on datasets for fairness-aware machine learning in Section 5. Finally, the conclusion and outlook are summarized in Section 6.

2 Methodology of the survey process

In this section, we describe our dataset collection strategy and introduce Bayesian networks as a tool for learning the structure from the data. In addition, we provide a summary of fairness measures we will use for the quantitative evaluation.

2.1 Strategy for collecting datasets

To identify the relevant datasets, we use Google Scholar[2] with "fairness datasets" as the primary query term along with other terms like "bias", "discrimination", "public" to narrow down the search. After identifying the related datasets, we use Google Scholar to find the related papers which satisfy the following conditions: 1) The public dataset is used in the experiments, and 2) The learning tasks, i.e., classification, clustering, are related to fairness problems. To restrict the investigation of the related work, we consider only important works as assessed by the number of citations, quality of publication venue, i.e., published in ranked conferences, journals. We consider datasets that have been used in at least three fairness-related papers. Datasets that are not publicly available via some known repository like the UCI machine learning repository [3],

[2] https://scholar.google.com/
[3] https://archive.ics.uci.edu

Kaggle [4], etc., are not taken into consideration.

2.2 Bayesian network

A Bayesian network (BN) (Holmes & Jain, 2008) is a directed and acyclic probabilistic graphical model which provides a graphical representation to understand the complex relationships between a set of random variables. In the case of a dataset, random variables corresponding to the attributes of the feature space in which the data are represented. The graphical structure $\mathcal{M} : \{\mathcal{V}, \mathcal{E}\}$ of a BN contains a set of nodes $\mathcal{V}$ (random variables/attributes) and a set of directed edges $\mathcal{E}$. Let $X_1, X_2, \cdots, X_d$ be the attributes defining the feature space $\mathcal{X}$ of a dataset $\mathcal{D}$, such that $\mathcal{X} \in \mathbb{R}^d$. For two attributes $X_i, X_j \in \mathcal{X}$, if there is a directed edge from X_i to X_j, then X_i is called the parent of X_j. The edges indicate conditional dependence relations, i.e., if we denote X_{pa_i} as the parents of X_i, the probability of X_i is conditionally dependent on the probability of X_{pa_i}. If we know the outcome (value) of X_{pa_i}, then the probability of X_i is conditionally independent of any other ancestor node. The structure of a BN describes the relationships between given attributes, i.e., the joint probability distribution of the attributes in the form of conditional independence relations. Formally:

$$P(X_1, X_2, \cdots, X_d) = \prod_{i=1}^{d} P(X_i \mid X_{pa_i}) \tag{1}$$

Learning the structure of a BN from the dataset $\mathcal{D}$ is an optimization problem (Husmeier, Dybowski, & Roberts, 2006), namely to learn an optimal BN model $\mathcal{M}^\star$ which maximizes the likelihood of generating $\mathcal{D}$. A set of parameters of any BN model $\mathcal{M}$, denoted by $\widehat{\mathcal{M}}$, is the set of edges $\mathcal{E}$ which represents the conditional independence relationship between the attribute set $\mathcal{V}$. Moreover, between the possible models M, the less complex one, i.e., the one with the least $\widehat{\mathcal{M}}$, should be selected.

Note that in a learned BN model $\mathcal{M}$, the position of the class attribute y can be in any position (root-, internal- or leaf-node), since the objective is to maximize $P(\mathcal{D} \mid \mathcal{M})$. However, we aim to investigate the factors (protected/non-protected attributes) that determine the class attribute's prediction probability. Therefore, we also employ a constraint on the class attribute to be a leaf node in our learning objective. Formally the problem is defined as:

$$\begin{aligned} &\max_{\mathcal{M}^\star} \{P(\mathcal{D} \mid \mathcal{M}) - \gamma \widehat{\mathcal{M}}\} \\ &\text{subject to } y \in \mathcal{L} \end{aligned} \tag{2}$$

where $y \in \mathcal{X}$ is the class attribute, $\mathcal{L}$ is the set of leaf nodes and γ is a penalty hyperparameter controlling the effect of the model's complexity in the final model selection. The aim of the learned model is to maximize $P(X_i \mid X_{pa_i})$ for each $X_i \in \mathcal{X}$ (Eq. 1 and Eq. 2).

A high conditional probability often refers to a strong correlation (Daniel, 2017). Attribute X_i is strongly correlated with X_j if there exists a *direct edge* between X_i and X_j, for any pair of attributes $X_i, X_j \in \mathcal{X}$. Intuitively, the correlation is comparatively weaker with ancestors that are not immediate parents, i.e., *indirect edges*. In addition, the attributes which do not have any incoming or outgoing edge (direct/indirect connection) with X_i, the correlation between them will be negligible. As a consequence, if we find any direct/indirect edge from any protected

[4] https://www.kaggle.com

attribute to the class attribute in our learned BN structure $\mathcal{M}^{\star}$ then we may infer that the dataset is biased w.r.t. the specific protected attribute.

When learning a BN, the continuous variables are often discretized because many BN learning algorithms cannot efficiently handle continuous variables (Chen, Wheeler, & Kochenderfer, 2017). Therefore, we need to discretize the continuous numeric data attributes into meaningful categorical attributes to keep the complexity of learning the BN model in a polynomial time. We describe the discretization procedure for each dataset in Section 3.

2.3 Fairness metrics

Measuring bias in ML models comprises the first step to bias elimination. Fairness depends on context; thus, a large variety of fairness measures exists. Only in the computer science research area, more than 20 measures of fairness have been introduced thus far (Žliobaitė, 2017; Verma & Rubin, 2018). Nevertheless, there is no fairness measure that is universally suitable (Foster, Ghani, Jarmin, Kreuter, & Lane, 2016; Verma & Rubin, 2018). Therefore, to make the experimental results more diverse, we report on three prevalent fairness measures: *statistical parity*, *equalized odds* and *Absolute Between-ROC Area (ABROCA)*. In which, *statistical parity* (Dwork, Hardt, Pitassi, Reingold, & Zemel, 2012) is one of the earliest and most popular discrimination measures in the fairness-aware ML literature. Statistical parity is also considered as the statistical counterpart of the legal doctrine of disparate impact (Krop, 1981). However, one main disadvantage of statistical parity is that it does not require compliance to the ground truth labels; hence, in many ML scenarios might not be ideal (Hardt, Price, & Srebro, 2016). *Equalized odds* introduced by (Hardt et al., 2016) countered this problem by considering the ground truth of both positive and negative class instances and grew to be one of the most promising fairness notion, being used in the leading edge methods (Zafar, Valera, Gomez Rodriguez, & Gummadi, 2017; Krasanakis, Spyromitros-Xioufis, Papadopoulos, & Kompatsiaris, 2018; Iosifidis & Ntoutsi, 2019). Later, (Gardner, Brooks, & Baker, 2019) argued that *equalized odds* does not consider any formal strategy such as slicing analysis to identify the prevalent biases, which might be a necessity in particular domains such as education. ABROCA measure introduced by (Gardner et al., 2019) tackles such an analysis issue and is argued to be an illustratively efficient method of representing the divergence of values of a protected attribute. Although, as mentioned earlier, there is a rich literature of fairness notions to follow, in this work, we limit our study to the above-mentioned notions, as these notions together cover a diverse area of the fairness concepts currently followed in the state-of-the-art fairness-aware ML practices.

The measures are presented hereafter assuming the following problem formulation: Let $\mathcal{D}$ be a binary classification dataset with class attribute $y = \{+, -\}$. Let S be a binary protected attribute with $S \in \{s, \overline{s}\}$, in which s is the discriminated group (referred to as *protected group*), and $\overline{s}$ is the non-discriminated group (referred to as *non-protected group*). For example, let S = "Sex" $\in$ {Female, Male} be the protected attribute; s = "Female" could be the protected group and $\overline{s}$ = "Male" could be the non-protected group. We use the notation s_+ (s_-), $\overline{s}_+$ ($\overline{s}_-$) to denote the protected and non-protected groups for the positive (negative, respectively) class.

2.3.1 Statistical parity

Statistical parity (SP) introduced by (Dwork et al., 2012) states that the output of any classifier satisfies SP if the difference (bias) in predicted outcome ($\hat{y}$) between any two groups under study

(i.e., s and $\overline{s}$) is up to a predefined tolerance threshold ϵ. Formally:

$$P(\hat{y}|S = s) - P(\hat{y}|S = \overline{s}) \leq \epsilon \tag{3}$$

Using the definition in Eq. 3 to measure the bias of a classifier, various measuring notions (Simoiu, Corbett-Davies, & Goel, 2017; Žliobaitė, 2015) have been proposed. The violation of statistical parity can be measured as:

$$SP = P(\hat{y} = +|S = \overline{s}) - P(\hat{y} = +|S = s) \tag{4}$$

The value domain is: $SP \in [-1, 1]$, with $SP = 0$ standing for no discrimination, $SP \in (0, 1]$ indicating that the protected group is discriminated, and $SP \in [-1, 0)$ meaning that the non-protected group is discriminated (*reverse discrimination*).

2.3.2 Equalized odds

Equalized odds (shortly *Eq.Odds*) (Hardt et al., 2016) is preserved when the predictions $\hat{y}$ conditional on the ground truth y is equal for both the groups s and $\overline{s}$ defined by S. Formally:

$$Eq.Odds : P(\hat{y} = +|S = s, Y = y) = P(\hat{y} = +|S = \overline{s}, Y = y) \tag{5}$$

where y is the ground truth class label, $\hat{y}$ is the predicted label.

Using Eq. 5 we can measure the prevalent bias as:

$$Eq.Odds_{viol} = \sum_{y \in \{+,-\}} |P(\hat{y} = +|S = s, Y = y) - P(\hat{y} = +|S = \overline{s}, Y = y)| \tag{6}$$

The value domain is: $Eq.Odds_{viol} \in [0, 2]$, with 0 standing for no discrimination and 2 indicating the maximum discrimination.

2.3.3 Absolute Between-ROC Area (ABROCA)

This is a fairness measure introduced by the research of (Gardner et al., 2019). It is based on the Receiver Operating Characteristics (ROC) curve. ABROCA measures the divergence between the protected (ROC_s) and non-protected group ($ROC_{\overline{s}}$) curves across all possible thresholds $t \in [0, 1]$ of *false positive rates* and *true positive rates*. In particular, it measures the absolute difference between the two curves in order to capture the case that the curves may cross each other and is defined as:

$$\int_0^1 \mid ROC_s(t) - ROC_{\overline{s}}(t) \mid \, dt \tag{7}$$

ABROCA takes values in the $[0, 1]$ range. The higher value indicates a higher difference in the predictions between the two groups and therefore, a more unfair model.

3 Datasets for fairness

In this section, we provide a detailed overview of real-world datasets used frequently in fairness-aware learning. We organize the datasets in terms of the application domain, namely: financial datasets (Section 3.1), criminological datasets (Section 3.2), healthcare and social datasets

Table 1: Overview of real-world datasets for fairness.

Dataset	#Instances	#Instances (cleaned)	#Attributes (cat./bin./num.)	Class	Domain	Class ratio (+:-)	Protected attributes	Target class	Collection period	Collection location
Adult	48,842	45,222	7/2/6	Binary	Finance	1:3.03	Sex, race, age	Income	1994	US
KDD Census-Income	299,285	284,556	32/2/7	Binary	Finance	1:15.30	Sex, race	Income	1994-1995	US
German credit	1,000	1,000	13/1/7	Binary	Finance	2.33:1	Sex, age	Credit score	1973-1975	Germany
Dutch census	60,420	60,420	10/2/0	Binary	Finance	1:1.10	Sex	Occupation	2001	Netherlands
Bank marketing	45,211	45,211	6/4/7	Binary	Finance	1:7.55	Age, marital	Deposit subscription	2008-2013	Portugal
Credit card clients	30,000	30,000	8/2/14	Binary	Finance	1:3.52	Sex, marriage, education	Default payment	2005	Taiwan
COMPAS recid.	7,214	6,172	31/6/14	Binary	Criminology	1:1.20	Race, sex	Two year recidivism	2013-2014	US
COMPAS viol. recid.	4,743	4,020	31/6/14	Binary	Criminology	1:5.17	Race, sex	Two year violent recid.	2013-2014	US
Communities&Crime	1,994	1,994	4/0/123	Multi	Criminology	-	Black	Violent crimes rate	1995	US
Diabetes	101,766	45,715	33/7/10	Binary	Healthcare	1:3.13	Gender	Readmit in 30 days	1999–2008	US
Ricci	118	118	0/3/3	Binary	Society	1:1.11	Race	Promotion	2003	US
Student-Mathematics	649	649	4/13/16	Binary	Education	1:2.04	Sex, age	Final grade	2005-2006	Portugal
Student-Portuguese	649	649	4/13/16	Binary	Education	1:5.49	Sex, age	Final grade	2005-2006	Portugal
OULAD	32,593	21,562	7/2/3	Multi	Education	-	Gender	Outcome	2013-2014	England
Law School	20,798	20,798	3/3/6	Binary	Education	8.07:1	Male, Race	Pass the bar exam	1991	US

(Section 3.3) and educational datasets (Section 3.4). A summary of the statistics of the different datasets[5] is provided in Table 1.

For each dataset, we discuss the basic characteristics like cardinality, dimensionality and class imbalance as well as typically used protected attributes in the literature. When available, we also provide temporal information regarding the data collection and the timespan of the datasets.

We start our analysis with the BN structure learned from the data (see Section 2.2), which can help us to understand the relationships between attributes of the dataset. In addition, the BN visualization already provides interesting insights on the dependencies between non-protected and protected attributes and their conditional dependencies in predicting the class attribute. We further provide an exploratory analysis of interesting correlations from the Bayesian graph (for both direct- and indirect- edges), particularly those related to the fairness problem (paths to and from protected attributes).

3.1 Financial datasets

3.1.1 Adult dataset

The adult dataset (Kohavi, 1996) (also known as "Census Income" dataset[6]) is one of the most popular datasets for fairness-aware classification studies (Appendix A). The classification task is to decide whether the annual income of a person exceeds 50,000 US dollars based on demographic characteristics.

Dataset characteristics: The dataset consists of 48,842 instances, each described via 15 attributes, of which 6 are numerical, 7 are categorical and 2 are binary attributes. An overview of attribute characteristics is shown in Table 2. We discard the attribute *fnlwgt* (final weight) as the suggestions of related work (B. H. Zhang, Lemoine, & Mitchell, 2018; Kamiran & Calders, 2012; Calders, Kamiran, & Pechenizkiy, 2009; Calders & Kamiran, 2010).

Missing values exist in 3,620 (7.41%) records. Many researchers remove the instances containing missing values (Zafar, Valera, Rogriguez, & Gummadi, 2017; Iosifidis & Ntoutsi, 2018, 2019; Choi, Farnadi, Babaki, & Van den Broeck, 2020) in their experiments; other researches consider the whole dataset or do not clarify how the missing values are handled. To avoid the effect of missing values on the analysis, we remove the missing data and obtain a clean dataset with 45,222 instances.

[5] We use the names of the protected attributes given in the original datasets, i.e. *sex, gender* are used with the same meaning. We do not present the *class ratio* (denoted by '-') in several datasets because their *class label* is 'multiple'.

[6] https://archive.ics.uci.edu/ml/datasets/adult

Table 2: Adult: attributes characteristics

Attributes	Type	Values	#Missing values	Description
age	Numerical	[17 - 90]	0	The age of an individual
workclass	Categorical	7	2,799	The employment status (Private, State-gov, etc.)
fnlwgt	Numerical	[13,492 - 1,490,400]	0	The final weight
education	Categorical	16	0	The highest level of education
educational-num	Numerical	1 - 16	0	The highest level of education achieved in numerical form
marital-status	Categorical	7	0	The marital status
occupation	Categorical	14	2,809	The general type of occupation
relationship	Categorical	6	0	Represents what this individual is relative to others
race	Categorical	5	0	Race
sex	Binary	{Male, Female}	9	The biological sex of the individual
capital-gain	Numerical	[0 - 99,999]	0	The capital gains for an individual
capital-loss	Numerical	[0 - 4,356]	0	The capital loss for an individual
hours-per-week	Numerical	[1 - 99]	0	The hours an individual has reported to work per week
native-country	Categorical	41	857	The country of origin for an individual
income	Binary	{≤50K, >50K}	0	Whether or not an individual makes more than $50,000 annually

Protected attributes: Typically the following attributes have been used as bias triggers in the literature[7]:

- *sex* = {*male, female*}: the dataset is dominated by male instances. The ratio of *male:female* is 32,650:16,192 (66.9%:33.1%).
- *race* = {*white, black, asian-pac-islander, amer-indian-eskimo, other*}. Typically, *race* is used as a binary attribute in the related work (Luong, Ruggieri, & Turini, 2011; Chakraborty, Peng, & Menzies, 2020; Zafar, Valera, Rogriguez, & Gummadi, 2017): *race* = {*white, non-white*}. The dataset is dominated by *white* people, the *white:non-white* ratio is 38,903:6,319 (86%:14%). In our analysis we also encode *race* as a binary attribute.
- *age* = [17-90]. Typically, *age* is used as a categorical attribute in the related work. In our analysis, we also discretize *age* as (Zafar, Valera, Rogriguez, & Gummadi, 2017): *age* = {25-60, <25 or >60}. The dataset is dominated by the $[25-60]$ years old group, the ratio is 35,066:10,156 (77.5%:22.5%).

In the research of (Deepak & Abraham, 2020), *marital-status* and *native-country* are considered as the protected attributes. However, due to missing information on their pre-processing method on these attributes, we will not consider those as the protected attributes in our survey.

Class attribute: The class attribute is *income* $\in \{\leq 50K, > 50K\}$ indicating whether an individual makes less or more than 50K. The positive class is "$> 50K$". The dataset is imbalanced with an imbalance ratio (IR) $1:3.03$ (positive:negative).

Bayesian network: Figure 1 illustrates the Bayesian network learned from the dataset. The class label *income* is the leaf node, i.e., there are no outgoing edges. To generate the Bayesian network, we discretize four numerical attributes (*age, capital gain, capital loss, hours per week*) as follows: *age* = {25-60, <25 or >60}; *capital gain* = {≤5000, >5000}, *capital loss* = {≤40, >40}; *hours per week* = {<40, 40-60, >60}. To reduce the computation space of the BN generator, we also transform seven categorical attributes as follows: *workclass* = {*private, non-private*}; *education* = {*high, low*}; *marital-status* = {*married, other*}; *relationship* = {*married, other*}; *native-country* = {*US, non-US*}; *race* = {*white, non-white*}; *occupation* = {*office, heavy-work, other*}.

[7] Please note that the majority of fairness-aware ML methods can handle only single protected attribures. The problem of multi-fairness has only recently been addressed (Hébert-Johnson, Kim, Reingold, & Rothblum, 2018; Martinez, Bertran, & Sapiro, 2020; Abraham, Sundaram, & Deepak, 2020)

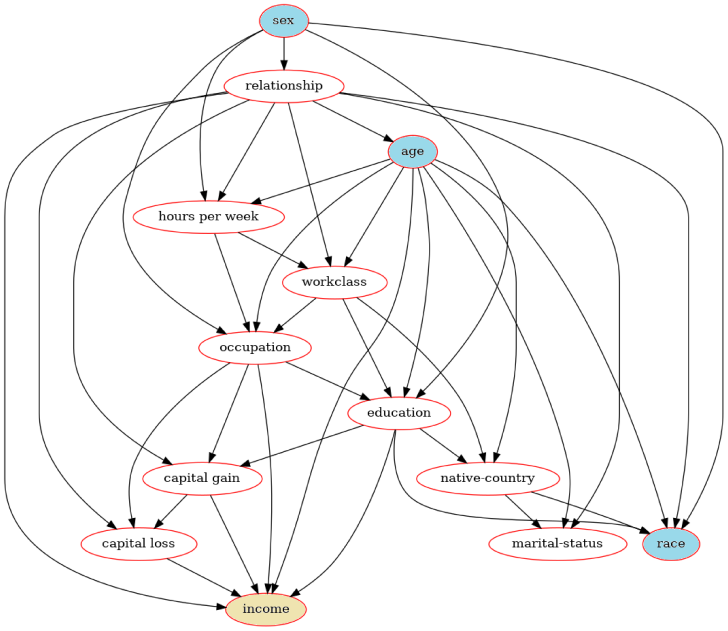
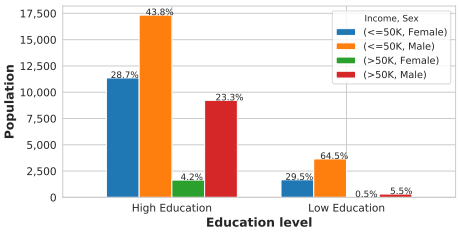
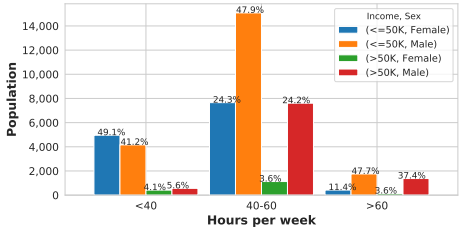


Figure 1: Adult: Bayesian network (class label: *income*, protected attributes: *sex*, *race*, *age*)

As demonstrated in Figure 1, there is a direct dependency between *income* and *education* as well as between *sex* and *education*. Therefore, we explore in more detail the distribution of the population w.r.t. *education*, *income* and *sex* in Figure 2a. As expected, highly educated people have a high income. However, in the high education segment and for the high-income class, the number of males is at least 5 times higher than that of females showing an under-representation of high education women in the high-income class. Based on the dependence of *hours per week* attribute on *sex*, we plot the weekly working hours w.r.t *income* and *sex* (Figure 2b). The number of males who work more than 40 hours per week is approximately 7 times higher than the number of females.



(a) Distribution of *education* and *income* w.r.t *sex*



(b) Distribution of *weekly working hours* and *income* w.r.t *sex*

Figure 2: Adult: relationships of *education*, *weekly working hours*, *education* and *income* attributes

Interestingly, there are many outgoing edges from the *relationship* and *age* attributes in the Bayesian network. We show the distribution of *sex* in each class based on the *age* (x-axis) and the *relationship* status (y-axis) in Figure 3. A first observation is that a great amount of young (less than 25 years old) or old (more than 60 years old) people do not receive more than 50K. *"Unmarried"* people have an income higher than 50K when they are older than 45 years, while people in the *"Own-child"* group can have a high income when they are young. In general, there are more males than females for almost all relationship statuses for the high-income group.

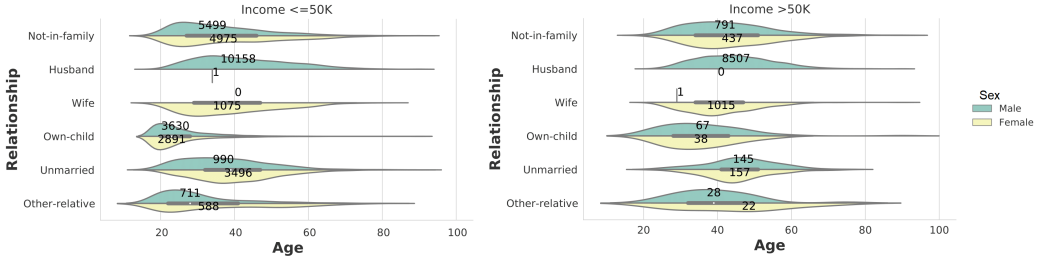


Figure 3: Adult: distribution of *age*, *relationship* and *income* w.r.t *sex*

Another interesting observation is that there is a direct edge from protected attribute *sex* to *race*. This suggests that choosing *sex* as the protected attribute would make the fairness-aware classifier attain fairness w.r.t *race*. Evidence of such outcome is seen in the work of (Friedler et al., 2019).

3.1.2 KDD Census-Income dataset

The KDD Census-Income[8] dataset (Dheeru & Karra Taniskidou, 2017) was collected from Current Population Surveys implemented by the U.S. Census Bureau from 1994 to 1995. The dataset has been considered in numerous related works (Appendix A). The prediction task is to decide if a person receives more than 50,000 US dollars annually or not. The prediction task is the same as the *Adult* dataset. However, the differences between the two datasets described by the dataset's authors (Dheeru & Karra Taniskidou, 2017) are: "the goal field was drawn from the *total person income* field rather than the *adjusted gross income* and may, therefore, behave differently than the original adult goal field".

Dataset characteristics: The dataset contains 299,285 instances with 41 attributes, 32 of which are categorical, 7 are numerical and 2 are binary attributes. An overview of the dataset characteristics[9] is shown in Table 3 and Table 16 (Appendix B). Attribute *weight* is omitted as proposed by the authors of the dataset (Dheeru & Karra Taniskidou, 2017).

Missing values exist in 157,741 (52.71%) instances. Because related studies only focus on a subset of data and features, we clean the dataset by eliminating all missing values. In particular, we remove four features *migration-code-change-in-msa, migration-code-change-in-reg, migration-code-move-within-reg, migration-prev-res-in-sunbelt* due to their high proportion in the missing values, as illustrated in Table 3. The result is a cleaned dataset with 284,556 instances.

Protected attributes: Previous researches consider *sex* as a protected attribute (Iosifidis & Ntoutsi, 2019; Ristanoski, Liu, & Bailey, 2013; Iosifidis & Ntoutsi, 2020). Attribute *race* =

[8] https://archive.ics.uci.edu/ml/datasets/Census-Income+(KDD)
[9] Table 3 describes attributes used in the Bayesian network

Table 3: KDD Census-Income: attributes characteristics

Attributes	Type	Values	#Missing values	Description
age	Numerical	[0 - 90]	0	The age of an individual
workclass	Categorical	9	0	Represents class of the worker
industry	Categorical	52	0	The industry code
occupation	Categorical	47	0	The occupation code
education	Categorical	17	0	The highest level of education
wage-per-hour	Numerical	[0 - 9,999]	0	Wage per hour
marital-status	Categorical	7	0	The marital status
race	Categorical	5	0	Race
sex	Binary	{Male, Female}	0	The biological sex of the individual
employment-status	Categorical	8	0	The employment status (full or part time)
capital-gain	Numerical	[0 - 99,999]	0	The capital gains for an individual
capital-loss	Numerical	[0 - 4,608]	0	The capital loss for an individual
dividends-from-stocks	Numerical	[0 - 99,999]	0	The dividends from stocks
tax-filer-stat	Categorical	6	0	The tax filer status (joint under 65, joint 65+, etc.)
detailed-household-and-family-stat	Categorical	38	0	The detailed household and family (child under 18, grandchild etc.)
detailed-household-summary-in-household	Categorical	8	0	The detailed household summary (spouse, non-relative, etc.)
num-persons-worked-for-employer	Numerical	[0 - 6]	0	The number of persons worked for the employer
family-members-under-18	Categorical	5	0	Family members under 18 (both parent, mother only, etc.)
citizenship	Categorical	5	0	The citizenship
own-business	Categorical	3	0	Own business or self employed
veterans-benefits	Categorical	3	0	Veterans benefits
weeks-worked	Numerical	[0 - 52]	0	The number of weeks worked in a year
year	Categorical	2	0	The year in which the interviewee answered
income (class)	Binary	{≤50K, >50K}	0	Whether an individual makes more than $50,000 annually

{*white, black, asian-pac-islander, amer-indian-eskimo, other*} could be also employed as a protected attribute because it has the same role as in the original *Adult* dataset. Similarly to the Adult dataset, the KDD Census-Income dataset is dominated by *white* people; there are 239,081 (84.01%) *white* people, hence, we encode *race* as a binary attribute for our analysis.

- *sex* = {*male, female*}. The dataset is slightly imbalanced towards female instances, the *male:female* ratio is 136,447:148,109 (48%:52%).
- *race* = {*white, non-white*}. The dataset is dominated by white people, the *white:non-white* ratio is 239,081:29,239 (86%:14%).

Class attribute: The class attribute is *income* $\in \{\leq 50K, > 50K\}$ indicating whether an individual makes less or more than 50K. The positive class is "$> 50K$". The dataset is very imbalanced with an IR $1 : 15.30$ (positive:negative).

Bayesian network: To generate the Bayesian network, we encode the following attributes: *age* = {≤25, 26-60, >60}; *wage-per-hour* = {≤500, 501-1000, >1000}; *industry* = {≤30, >30}; *occupation* = {≤10, >10}; *capital-gain* = {≤500, >500}; *capital-loss* = {≤500, >500}; *dividends-from-stocks* = {≤500, 501-2000, >2000}; *num-persons-worked-for-employer* = {0, >0}; *weeks-worked-in-year* = {≤26, 27-51, 52}. The ranges of encoded attributes are chosen to ensure each group has values. To reduce the complexity, we eliminate these attributes: *enroll-in-edu-inst-last-wk, major-industry, major-occupation* since they have a very low correlation with other features. Also, for efficiency purposes, we generate the Bayesian network on a randomly selected 10% sample of the dataset rather than on the complete dataset. The learned Bayesian network is shown in Figure 4; the class label *income* is set as a leaf node.

As shown in Figure 4, *income* is conditionally dependent on *sex*, *occupation* and the number of week worked in year (*weeks-worked*) attributes. Regarding *sex* attribute, females are largely underrepresented in the high income group, consisting of 13,691 males (∼10.03% of the male population) and only 3,711 females (∼2.51% of the female population). Regarding the number of weeks worked per year and *income*, as shown in Figure 5, women tend to do part-time jobs, i.e., the number of weeks worked per year is less than 26. In addition, women earn less money than men even though they all work 52 weeks per year. That is shown by the number of men

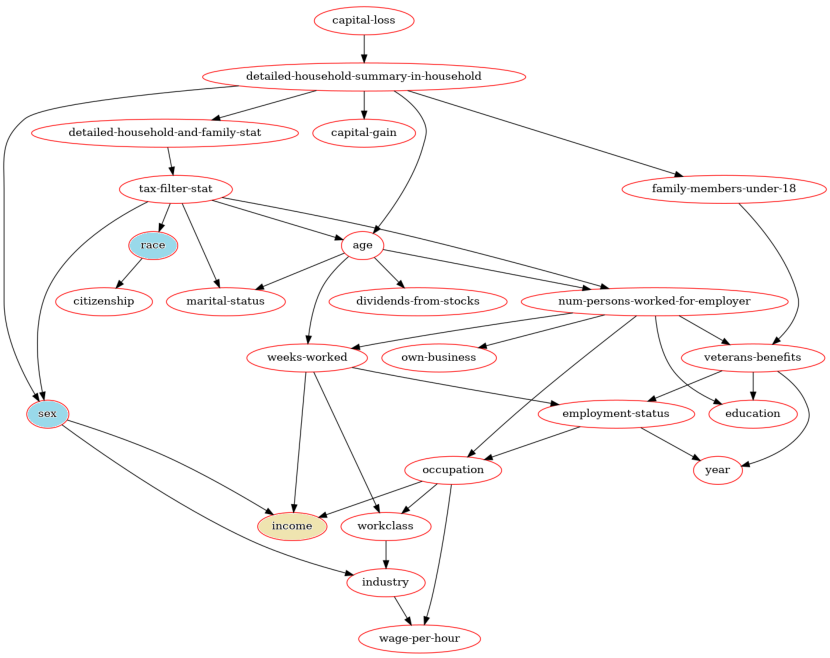


Figure 4: KDD Census-Income: Bayesian network (class label: *income*, protected attributes: *sex, race*)

with high income is approximately five times more than the number of women.

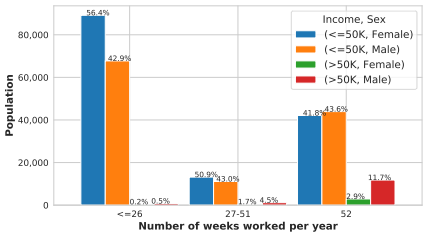


Figure 5: KDD Census-Income: relationship of *the number of weeks worked per year* and *income* w.r.t *sex*

As mentioned, *race* could also be considered as the protected attribute. Based on the data, the income of *non-white* people is significantly different from the income of the *white* group. Only 3.2% of the *non-white* group have an income above 50K, compared to 6.7% for the *white* group. Furthermore, since *age* has a conditional dependence on *marital-status* attribute, we investigate the relationship between these attributes, the protected attribute *sex* and the class label *income* in Figure 6. As shown in this figure, males comprise the majority of the high-income group, especially for certain population segments like the *Married-civilian spouse present* segment where the number of males is 5 times higher than that of females. Interestingly, the number of widows is 1.7 times higher than the number of widowers in terms of high income.

Regarding the *age* effect, most people have a high income when they are over 40 years old. With respect to the protected attributes, there is no edge between *race* and *sex*, which suggests the researchers should perform their fairness-aware models on both these protected attributes.

Figure 6: KDD Census-Income: relationship of *marital status*, *age*, *sex* and *income*

3.1.3 German credit dataset

The German credit[10] dataset (Dheeru & Karra Taniskidou, 2017) consists of samples of bank account holders. The dataset is used for risk assessment prediction, i.e., to determine whether it is risky to grant credit to a person or not. The dataset is frequently employed in fairness-aware learning researches (Appendix A).

Dataset characteristics: The dataset contains only 1,000 instances without any missing values. Each sample is described by 13 categorical, 7 numerical and 1 binary attributes. An overview of all attributes is presented in Table 4. Attribute *personal-status-and-sex* contains information of marital status and the gender of people. We disentangle gender from personal status and create two separate attributes: *marital-status* and *sex*. The original *personal-status-and-sex* attribute is omitted from further analysis.

Table 4: German credit: attributes characteristics

Attributes	Type	Values	#Missing values	Description
checking-account	Categorical	4	0	The status of existing checking account
duration	Numerical	[4 - 72]	0	The duration of the credit (month)
credit-history	Categorical	5	0	The credit history
purpose	Categorical	10	0	Purpose (car, furniture, education, etc.)
credit-amount	Numerical	[250 - 18,424]	0	Credit amount
savings-account	Categorical	5	0	Savings account/bonds
employment-since	Categorical	5	0	Present employment since
installment-rate	Numerical	[1 - 4]	0	The installment rate in percentage of disposable income
personal-status-and-sex	Categorical	4	0	The personal status and sex
other-debtors	Categorical	3	0	Other debtors/guarantors
residence-since	Numerical	[1 - 4]	0	Present residence since
property	Categorical	4	0	Property
age	Numerical	[19 - 75]	0	The age of the individual
other-installment	Categorical	3	0	Other installment plans
housing	Categorical	3	0	Housing (rent, own, for free)
existing-credits	Numerical	[1 - 4]	0	Number of existing credits at this bank
job	Categorical	4	0	Job (unemployed, (un)skilled, management)
number-people-provide-maintenance-for	Numerical	[1 - 2]	0	Number of people being liable to provide maintenance for
telephone	Binary	{Yes, None}	0	Telephone number
foreign-worker	Binary	{Yes, No}	0	Is the individual a foreign worker?
class-label	Binary	{Good, Bad}	0	Class

Protected attributes: In all studies, *sex* is considered as the protected attribute. *Age* can also be considered as the protected attribute after binarization into {*young*, *old*} by *age* thresholding

[10] https://archive.ics.uci.edu/ml/datasets/statlog+(german+credit+data)

at 25 (Kamiran & Calders, 2009; Friedler et al., 2019).

- *sex* $=$ {*male, female*}. The dataset is dominated by male instances, the ratio of *male:female* is 690:310 (69%:31%). The percentage of women identified as *bad* customers is 35.2% while that of men is only 27.7%.
- *age* $=$ {$\leq$*25, >25*}: The dataset is dominated by people older than 25 years, the ratio is 810:190 (81%:19%). We discover that there is a discrimination on the age of customers. There are 42.1% of *young* people are recognized as *bad* customers while this proportion in *old* people is 27.2%.

Class attribute: The class attribute is *class-label* $\in \{good, bad\}$ revealing the customer's level of risk. The positive class is "good". The dataset is imbalanced with an IR $2.33 : 1$ (positive:negative).

Bayesian network: We transform the numerical attributes into categorical as follows: *duration* $=$ {$\leq$6, 7-12, >12} (short, medium and long-term); *credit-amount* $=$ {$\leq$2000, 2000-5000, >5000} (low, medium and high income); *age* $=$ {$\leq$25, >25}. The extracted Bayesian network is shown in Figure 7; *class-label* is set as a leaf node.

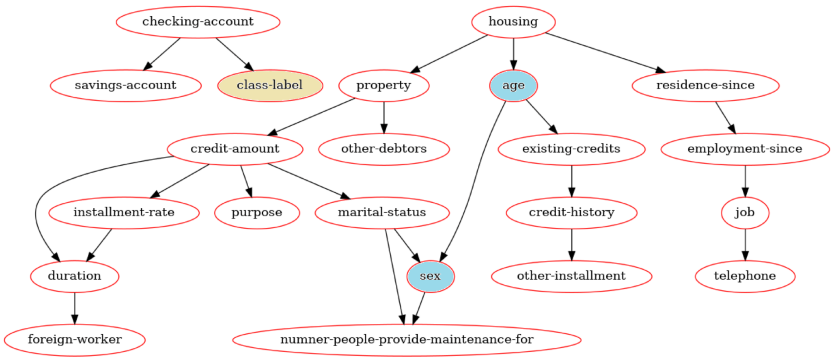
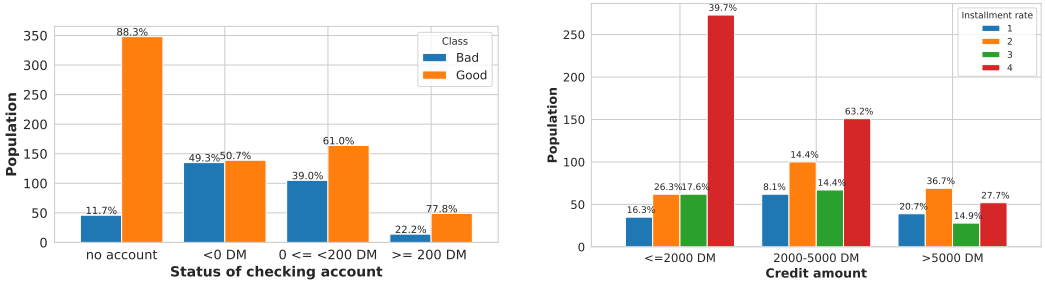


Figure 7: German credit: Bayesian network (class label: *class-label*, protected attributes: *sex, age*)

The Bayesian network consists of two disconnected components. First, *class-label* is conditionally dependent on the *checking-account* attribute. We investigate in more detail this relationship in Figure 8a. As we can see, a very high proportion of people, i.e., 88.3%, having no checking account is identified as the *good* customers while half of the customers having a balance less than 0 DM (*Deutsche Mark*) in their checking account are classified as the *bad* customers.

Second, interestingly, *credit-amount* has a direct effect on many attributes such as *installment-rate, duration*. We discover that people who borrow a great amount of money tend to borrow for a long period. For example, 93.6% of interviewees make a loan of more than 5000 DM with a loan duration of more than 12 months. As illustrated in Figure 8b, the number of customers who have to pay the highest installment rate (visualized as the "red" columns) is inversely proportional to the *credit-amount*. Regarding the protected attributes, a direct edge between *sex* and *age* is observed. This is the starting point of the research question "Does the fairness-aware model obtain fairness w.r.t. *sex* if *age* is chosen as the protected attributes?"



(a) Distribution of *class label* on *status of checking account*

(b) Relationship between *credit amount* and *installment rate*

Figure 8: German credit: relationships of class label and attributes

3.1.4 Dutch census dataset

The Dutch census dataset (Van der Laan, 2000) represented aggregated groups of people in the Netherlands for the year 2001. Researchers (Appendix A) have used Dutch dataset to formulate a binary classification task to predict a person's *occupation* which can be categorized as *high-level* (prestigious) or *low-level* profession.

Dataset characteristics: The dataset includes 60,420 samples[11] where each sample is described by 12 attributes. An overview of attributes is presented in Table 5.

Table 5: Dutch census: attributes characteristics

Attributes	Type	Values	#Missing values	Description
sex	Binary	{Male, Female}	0	The biological sex of the person
age	Categorical	12	0	The age group of the person (0-4, 5-9, etc.)
household_position	Categorical	8	0	The relationship to household head (spouse, child, etc.)
household_size	Categorical	6	0	The size of the household the person belongs to
prev_residence_place	Binary	{Netherlands, non-Netherlands}	0	The place of the person's residence prior to the Census
citizenship	Categorical	3	0	The person's citizenship status
country_birth	Categorical	3	0	Whether the person was born in the Netherlands or elsewhere
edu_level	Categorical	6	0	The person's level of educational attainment
economic_status	Categorical	3	0	The person's economic status (class of worker)
cur_eco_activity	Categorical	12	0	The current economic activity
marital_status	Categorical	4	0	The person's current marital status according to law or custom
occupation	Binary	{0, 1}	0	The person's occupation (0: low-level, 1: high-level)

Protected attributes: In the related work, they consider attribute *sex* $= \{$*male, female*$\}$ as the protected attribute, *male:female* ratio is 30,147:30,273 (49.9%:50.1%).

Class attribute: The class attribute is *occupation* $\in \{0, 1\}$ demonstrating if an individual has a prestigious profession or not. The positive class is *1 (high-level)*. This is a fairly balanced dataset in our survey with an IR $1 : 1.10$ (positive:negative).

Bayesian network: We use all attributes in the dataset to generate the Bayesian network. As illustrated in Figure 9, the leaf node *occupation* is conditionally dependent on *economic status, education level* and *sex* attributes. In fact, 62.6% of males (18,860 out of 30,147) have a high-level occupation, while this proportion on females group is only 32.7%. In addition, people with high education are doing prestigious jobs and vice versa, as depicted in Figure 10. For example, 89.5% of people having *tertiary* level are working in high-level jobs while around 80% of people

[11] https://github.com/tailequy/fairness_dataset/tree/main/Dutch_census

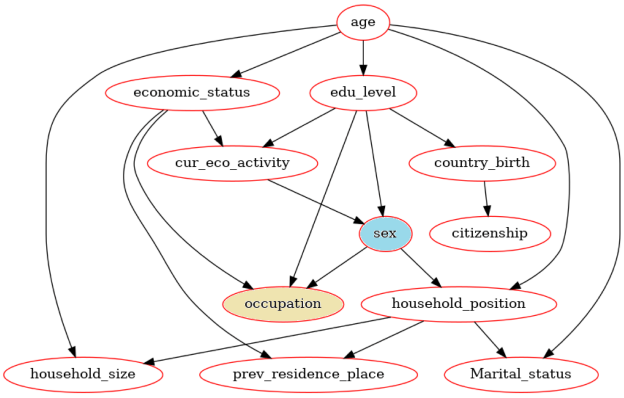


Figure 9: Dutch census: Bayesian network (class label: *occupation*, protected attribute: *sex*)

with *lower secondary* degrees are doing low-level work. Interestingly, *age* has a direct effect on many attributes.

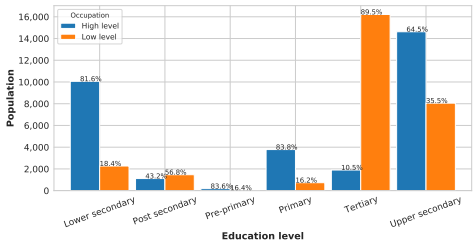


Figure 10: Dutch census: relationship between *education level* and *occupation*

3.1.5 Bank marketing dataset

The bank marketing[12] dataset (Moro, Cortez, & Rita, 2014) is related to the direct marketing campaigns of a Portuguese banking institution from 2008 to 2013. There is a variety of researchers investigating this dataset in their studies (Appendix A). The classification goal is to predict whether a client will make a deposit subscription or not.

Dataset characteristics: The dataset comprises 45,211 samples, each with 6 categorical, 4 binary and 7 numerical attributes, as summarized in Table 6.

Protected attributes: In the literature, *marital-status* can be considered as the protected attribute (Backurs et al., 2019; Hu et al., 2020; Chierichetti, Kumar, Lattanzi, & Vassilvitskii, 2017; Ziko, Yuan, Granger, & Ayed, 2021; Bera, Chakrabarty, Flores, & Negahbani, 2019). Besides, in several studies (Krasanakis et al., 2018; Zafar, Valera, Rogriguez, & Gummadi, 2017; Fish, Kun, & Lelkes, 2016), they consider *age* as the protected attribute which is binary separated into people who are between the age of 25 to 60 years old and less than 25 or more than 60 years old.

- *age* = {*25-60*, *<25 or >60*}: the dataset is dominated by people from 25 to 60 years old, the ratio of *"25-60":"<25 or >60"* is 43,214:1,997 (95.6%:4.4%).

[12] https://archive.ics.uci.edu/ml/datasets/Bank+Marketing

Table 6: Bank marketing: attributes characteristics

Attributes	Type	Values	#Missing values	#Missing values
age	Numerical	[18 - 95]	0	The age of the client
job	Categorical	12	0	The type of job (admin, self-employed, technician, ect.)
marital	Categorical	3	0	The marital status
education	Categorical	4	0	The education level
default	Binary	{Yes, No}	0	Has the credit in default?
balance	Numerical	[-8,019 - 102,127]	0	The balance of this client's account
housing	Binary	{Yes, No}	0	Has a housing loan?
loan	Binary	{Yes, No}	0	Has a personal loan?
contact	Categorical	3	0	The contact communication type
day	Numerical	[1 - 31]	0	The last contact day of the week
month	Categorical	12	0	The last contact month of the year
duration	Numerical	[0 - 4,918]	0	The last contact duration, in seconds
campaign	Numerical	[1 - 63]	0	The number of contacts performed during this campaign and for this client
pdays	Numerical	[-1 - 871]	0	The number of days that passed by after the client was last contacted
previous	Numerical	[0 - 275]	0	The number of contacts performed before this campaign and for this client
poutcome	Categorical	4	0	The outcome of the previous marketing campaign
y (class)	Binary	{Yes, No}	0	Has the client subscribed a term deposit?

- *marital* $=$ {*married, non-married*}: *married* group is the majority with the ratio of *married:non-married* is 27,214:17,997 (60.2%: 39.8%).

Class attribute: The class attribute is $y \in \{Yes, No\}$ presenting whether a customer will subscribe a term deposit or not. The positive class is "Yes". The dataset is imbalanced with an IR $1 : 7.55$ (positive:negative).

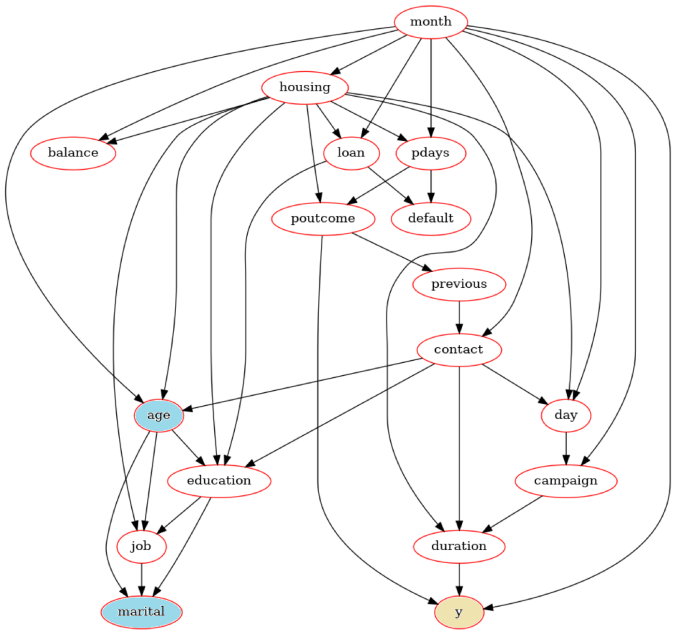


Figure 11: Bank marketing: Bayesian network (class label: *y*, protected attributes: *age, marital*)

Bayesian network: We perform a pre-processing step to transfer the numerical attributes into categorical: *job* $=$ {*blue-collar, management-service, other*}; *balance* $=$ {0, >0}; *day* $=$ {$\leq$15, >15}; *duration* $=$ {$\leq$120, 121-600, >600}; *campaign* $=$ {$\leq$1, 2-5, >5}; *pdays* $=$ {$\leq$30, 31-

180, >180}; *previous* = {0, 1-5, >5}. Figure 11 visualizes the Bayesian network of the Bank marketing dataset. The class label *y*, as illustrated in Figure 11, is conditionally dependent on *poutcome, month* and *duration* attributes. An insight about the relationship between the last contact *duration* and class label *y* is described in Figure 12. The ratio of clients who will make a deposit subscription is proportional to the duration of the last contact. When the talk is taken place in less than 2 minutes, 98.5% of people will not make the deposit subscription. However, if a marketing staff can maintain the talk with customers over 10 minutes, 48.4% of customers will say "*Yes*". Interestingly, in the Bayesian network, both protected attributes *age* and *marital* have no effect on the class label *y*. However, the protected attributes are connected together by an in-direct edge, which could be a reason for a similar accuracy of fairness-aware models of the related work (Hu et al., 2020) and (Krasanakis et al., 2018).

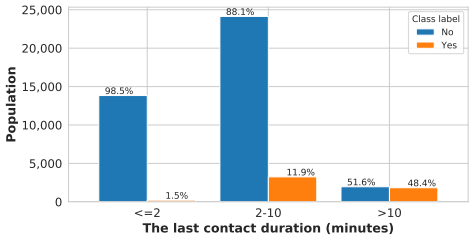


Figure 12: Bank marketing: Relationship between the last contact duration and class label

3.1.6 Credit card clients dataset

The credit card clients[13] dataset (Yeh & Lien, 2009) investigated the customers' default payments in Taiwan in October 2005. The goal is to predict whether a customer will face the default situation in the next month or not. The data have been used for default payment prediction in several studies (Appendix A).

Dataset characteristics: The dataset includes 30,000 customers described by 8 categorical, 14 numerical and 2 binary attributes, as depicted in Table 7. There is no missing value in the dataset.

Protected attributes: In the literature, *sex* (Deepak & Abraham, 2020; Bechavod & Ligett, 2017; Berk et al., 2017), *education, marriage* (Deepak & Abraham, 2020; Bera et al., 2019) are considered as the protected attributes.

- *sex* = {*male, female*}: the dataset is dominated by females, the ratio of *male:female* is 11,888:18,112 (39.6%:60.4%).
- *marriage* = {*married, single, others*}: *single* group is the majority with the ratio of *married:single:others* is 13,659:15,964:377 (45.5%:53.2%:1.3%).
- *education* = {*graduate school, university, high school, others*}: *university* is the biggest group with 14,030 (46.8%) clients.

Class attribute: The class attribute is *default payment* $\in \{0, 1\}$ indicating whether a customer will suffer the default payment situation in the next month (1) or not (0). The positive class is *1*. This is an imbalanced dataset with an IR $1 : 3.52$ (positive:negative).

[13] https://archive.ics.uci.edu/ml/datasets/default+of+credit +card+clients

Table 7: Credit card clients: attributes characteristics

Attributes	Type	Values	#Missing values	Description
limit_bal	Numerical	[10,000 - 1,000,000]	0	The amount of the given credit (New Taiwan dollar)
sex	Binary	{Male, Female}	0	The biological sex of the client
education	Categorical	7	0	The education level
marriage	Categorical	3	0	The marital status
age	Numerical	[21 - 79]	0	The age of the client (year)
pay_0	Categorical	11	0	The repayment status in September 2005 (pay duly, delay 1 month, etc.)
pay_2	Categorical	11	0	The repayment status in August 2005
pay_3	Categorical	11	0	The repayment status in July 2005
pay_4	Categorical	11	0	The repayment status in June 2005
pay_5	Categorical	10	0	The repayment status in May 2005
pay_6	Categorical	10	0	The repayment status in April 2005
bill_amt1	Numerical	[-165,580 - 964,511]	0	The amount of bill statement in September 2005
bill_amt2	Numerical	[-69,777 - 983,931]	0	The amount of bill statement in August 2005
bill_amt3	Numerical	[-157,264 - 1,664,089]	0	The amount of bill statement in July 2005
bill_amt4	Numerical	[-170,000 - 891,586]	0	The amount of bill statement in June 2005
bill_amt5	Numerical	[-81,334 - 927,171]	0	The amount of bill statement in May 2005
bill_amt6	Numerical	[-339,603 - 961,664]	0	The amount of bill statement in April 2005
pay_amt1	Numerical	[0 - 873,552]	0	The amount paid in September 2005
pay_amt2	Numerical	[0 - 1,684,259]	0	The amount paid in August 2005
pay_amt3	Numerical	[0 - 896,040]	0	The amount paid in July 2005
pay_amt4	Numerical	[0 - 621,000]	0	The amount paid in June 2005
pay_amt5	Numerical	[0 - 426,529]	0	The amount paid in May 2005
pay_amt6	Numerical	[0 - 528,666]	0	The amount paid in April 2005
default payment	Binary	{0, 1}	0	Whether or not the client face the default situation

Bayesian network: To generate the Bayesian network, we convert the numerical attributes: *age* = {≤35, 36-60, >60}; the amount of the given credit (*limit_bal*), the amount of the bill statements *(bill_amt_1,..., bill_amt_6)*, and the amount of the previous payments *(pay_amt_1, bill_1,..., pay_amt_6)* = {≤50,000, 50,001-200,000, >200,000} (corresponding to the *low, medium, high* levels); history of the past payments *pay_0,..., pay_6* = {*pay duly, 1-3 months, >3 months*}. The Bayesian network is presented in Figure 13.

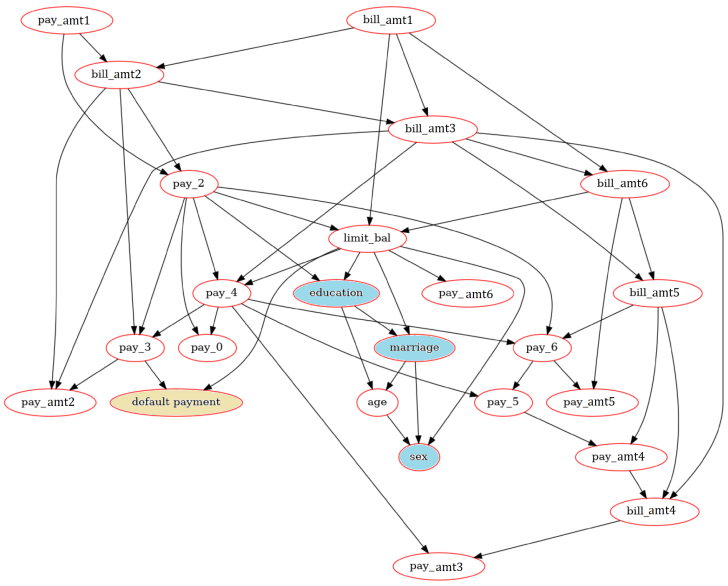


Figure 13: Credit card clients: Bayesian network (class label: *default payment*, protected attributes: *sex, marriage, education*)

The class label *default payment* is directly conditionally dependent on the repayment status in

July 2005 (attribute *pay_3*), and the given credit (attribute *limit_bal*) and indirectly dependent on the amount of bill statements (the attributes with a prefix *bill_amt*). As demonstrated in Figure 14, the ratio of the default payment phenomenon is inversely proportional to the credit limit balance. Moreover, we discover that the percentage of males having the default payment in the next month is higher than that of females. In particular, the ratio of males with the default payment is 24.2% while that of females is only 20.8%. Interestingly, the protected attributes (*sex, education, marriage*) are conditionally dependent on each other.

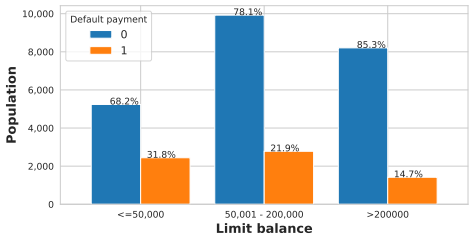


Figure 14: Credit card clients: Relationship between *the credit limit balance* and *default payment*

Summary of the financial datasets: In general, the financial datasets are very diverse as they were collected from several diverse locations (from US, to Taiwan) and at very different time points (from 1994 to 2013). With respect to the collection time, the datasets are pretty old, esp. Adult and KDD census datasets. These datasets have been heavily investigated in the related work and under different protected attributes. The most prevalent protected attribute is *sex* followed by *race, age, marriage* and *education*. An interesting observation is that the protected attributes are often related to each other (a strong or weak relationship), e.g., race with education. Due to these dependencies, ensuring fairness for one protected attribute may positively affect fairness for other protected attributes. Moreover, most of the datasets in this category are imbalanced, with the only exception of the Dutch census dataset which is almost balanced (see Table 1). In terms of class imbalance, datasets demonstrate different imbalance ratios.

3.2 Criminological datasets

3.2.1 COMPAS dataset

The COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) (Angwin et al., 2016) is a recent dataset, compared to the rest of the datasets in our survey, which was released by ProPublica[14] in 2016 based on the Broward County data (collected from Jan 2013 to Dec 2014). Defendant's answers to the COMPAS screening survey are used to generate the recidivism risk scores. The data have been used for crime recidivism risk prediction by a plethora of works (Appendix A). *Risk of recidivism* (denoted as *COMPAS recid.*) and *Risk of violent recidivism* (denoted as *COMPAS viol. recid*) subsets are most widely used in the literature. The former has a classification task to predict if an individual is rearrested within two years after the first arrest. The latter predicts if an individual is rearrested for a violent crime within two years.

Dataset characteristics: *COMPAS recid.* and *COMPAS viol. recid.* datasets contain 7,214 and 4,743 samples, respectively. Each defendant is described by 52 attributes[15] (31 categorical, 6 binary, 14 numerical and a null attribute), as shown in Table 8 and Table 17 (Appendix B).

[14] https://www.propublica.org/datastore/dataset/compas-recidivism-risk-score-data-and-analysis
[15] Table 8 describes attributes used in the Bayesian network and data analysis

Missing data is a common phenomenon in both subsets. There are 6,395 rows (88.6%) containing missing values in the COMPAS recid. subset while this number in the COMPAS viol. revid. subset is 3,748 (79%). Based on (Angwin et al., 2016), we clean the dataset by removing the missing data, such as *violent_recid* = *NULL* or the change date of a crime (attribute *days_b_screening_arrest*) was not within 30 days when he or she was arrested. The cleaned datasets used in our analysis contain 6,172 (COMPAS recid.) and 4,020 (COMPAS viol. recid.) records.

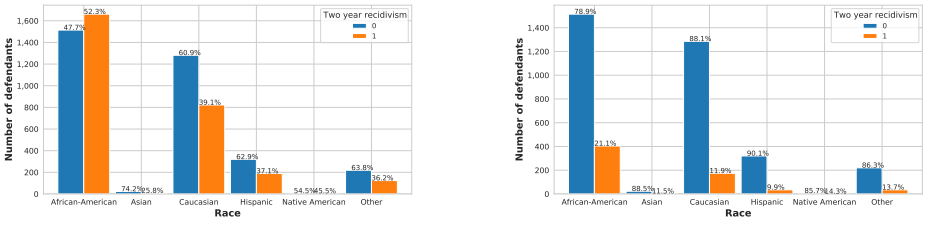
Protected attributes: Typically, *race* is employed as the protected attribute. In both subsets, *black* and *white* are the main races. In the COMPAS recid. subset, the *black:white* ratio is 3,175:2,103 (51.4%:34%) (computed on the total number of defendants), while this ratio in the COMPAS viol. recid. subset is 1,918:1,459 (47.7%:36.3%). Figure 15 describes the distribution of defendants w.r.t. *race* attribute. The recidivism rate in the black defendants is higher than that of the white defendants in both subsets.

Sex has been also considered as the protected attribute (Diana, Gill, Kearns, Kenthapadi, & Roth, 2021; van Berkel, Goncalves, Russo, Hosio, & Skov, 2021; Chakraborty, Majumder, Yu, & Menzies, 2020). The ratio *male:female* is 4,997:1,175 (81%:19%) in the COMPAS recid. subset, while this ratio in the COMPAS viol. recid. subset is 3,179:841 (79.1%:20.9%).

Class attribute: The class attribute is *two_year_recid* $\in \{0, 1\}$ indicating whether an individual will be rearrested within two years (1) or not (0). The positive class is *1*. The COMPAS recid. subset is fairly balanced with an IR $1 : 1.20$ (positive:negative) while the COMPAS viol. recid. subset is imbalanced with an IR $1 : 5.17$.

Table 8: COMPAS recid: attributes characteristics

Attributes	Type	Values	#Missing values	Description
sex	Binary	{Male, Female}	0	Sex
age	Numerical	[18 - 96]	0	Age in years
age_cat	Categorical	3	0	Age category
race	Categorical	6	0	Race
juv_fel_count	Numerical	[0 - 20]	0	The juvenile felony count
juv_misd_count	Numerical	[0 - 13]	0	The juvenile misdemeanor count
juv_other_count	Numerical	[0 - 17]	0	The juvenile other offenses count
priors_count	Numerical	[0 - 38]	0	The prior offenses count
c_charge_degree	Binary	{F, M}	0	Charge degree of original crime
score_text	Categorical	3	0	ProPublica-defined category of decile score
v_score_text	Categorical	3	0	ProPublica-defined category of v_decile_score
two_year_recid	Binary	{0, 1}	0	Whether the defendant is rearrested within two years



(a) COMPAS recid. subset

(b) COMPAS viol. recid. subset

Figure 15: COMPAS: distribution of *two year recidivism* w.r.t. *race*

Bayesian network: To generate the Bayesian network, we remove the temporal attributes such

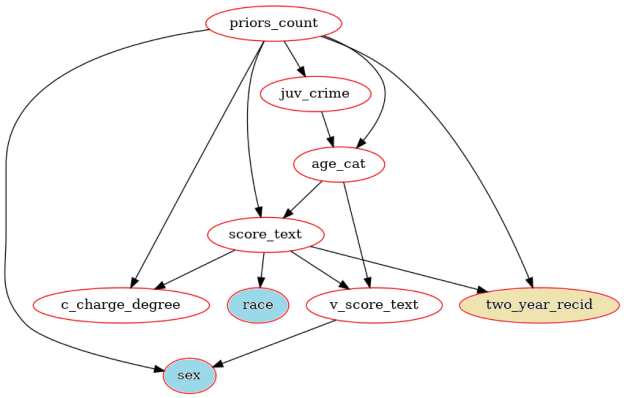


Figure 16: COMPAS recid.: Bayesian network (class label: *two_year_recid*, protected attributes: *race*, *sex*)

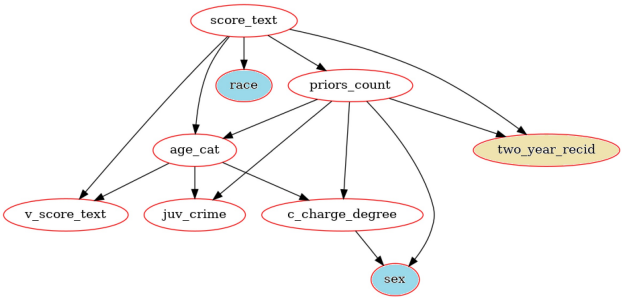
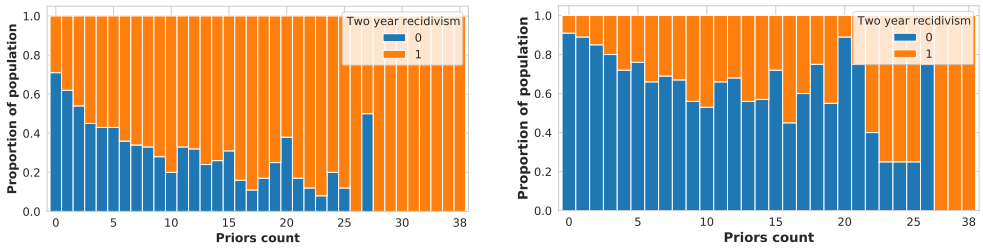


Figure 17: COMPAS viol. recid.: Bayesian network (class label: *two_year_recid*, protected attributes: *race*, *sex*)

as *screening_date* (the date on which the risk of recidivism score was given), *in_custody* (the date on which individual was brought into custody), and several ID-related attributes. A new attribute *juv_crime* is computed by the sum of the juvenile felony count (*juv_fel_count*) and the juvenile misdemeanor count (*juv_misd_count*) and the juvenile other offenses count (*juv_other_count*). We transform the numerical attributes into the categorical type: prior offenses count *priors_count* = {0, 1-5, >5}; the juvenile felony count *juv_crime* = {0, >0}. Figure 16 and Figure 17 are the Bayesian networks of the COMPAS dataset. The class label *two_year_recid* = {0, 1} is assigned as a leaf node. It shows the dependency of many attributes such as *sex*, age category (*age_cat*) on prior offenses count (*priors_count*) feature. For instance, the number of convictions directly affects the frequency of recidivism, as shown in Figure 18. If a defendant has a long history of convictions, his probability of recidivism is higher, especially when the number of convictions is more than 27 times, the recidivism probability is almost 100%.

Interestingly, *score_text* attribute (defines the category of the recidivism score) has many ingoing and outgoing edges as depicted in Figure 17. To clarify this phenomenon, we investigate the distribution of age, recidivism score (*score_text*) w.r.t. *race*, in Figure 19. The majority of recidivists are under the age of 30. In the recidivist group, the number of black criminals is four times and two times more than that of white criminals with a high recidivism score and



(a) COMPAS recid. subset (b) COMPAS viol. recid. subset

Figure 18: COMPAS: Relationship between recidivism and priors count

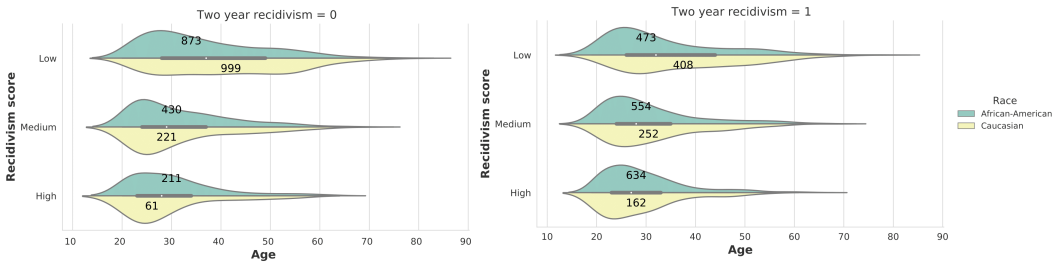


Figure 19: COMPAS recid. : distribution of recidivism score, age w.r.t. race

medium recidivism score, respectively. In the group of defendants with a low recidivism score, the distribution of the *race* is balanced.

3.2.2 Communities and Crime dataset

The Communities and Crime[16] dataset (Dheeru & Karra Taniskidou, 2017) was a small dataset containing the socio-economic data from 46 states of the United States in 1990 (the US Census). The law enforcement data come from the 1990 US LEMAS survey, and crime data come from the 1995 FBI Uniform Crime Reporting (UCR) program. The goal is to predict the total number of violent crimes per 100 thousand population. Many researchers are investigating the dataset in their experiments (Appendix A).

Dataset characteristics: The dataset contains only 1,994 samples; each instance is described by 127 attributes (4 categorical and 123 numerical attributes). A description of attributes[17] is illustrated in Table 9, Table 18 and Table 19 (Appendix B).

There is a very high proportion (84%) of missing values in 25 attributes, as demonstrated in Table 19. Based on the suggestions from the literature (Heidari, Ferrari, Gummadi, & Krause, 2018; Calders, Karim, Kamiran, Ali, & Zhang, 2013) , we remove all columns containing missing values. We create a new binary class label namely *class* based on *ViolentCrimesPerPop* attribute (the total number of violent crimes per 100,000 population). As illustrated in the related work (Kearns, Neel, Roth, & Wu, 2018), a label "high-crime" is set if the crime rate of the communities (possitive class) is greater than 0.7, otherwise, "low-crime" is given. The ratio of *high-crime:low-crime* is: 122:1,872 (6.1%:93.9%). Therefore, the dataset is very imbalanced with an IR $1:$

[16] http://archive.ics.uci.edu/ml/datasets/communities+and+crime

[17] Table 9 contains attributes used in the Bayesian network

Table 9: Communities and Crime: attributes characteristics

Attributes	Type	Values	#Missing values	Description
racepctblack	Numerical	[0.0 - 1.0]	0	The percentage of population that is African American
pctWInvInc	Numerical	[0.0 - 1.0]	0	The percentage of households with investment/rent income in 1989
pctWPubAsst	Numerical	[0.0 - 1.0]	0	The percentage of households with public assistance income in 1989
NumUnderPov	Numerical	[0.0 - 1.0]	0	The number of people under the poverty level
PctPopUnderPov	Numerical	[0.0 - 1.0]	0	The percentage of people under the poverty level
PctUnemployed	Numerical	[0.0 - 1.0]	0	The percentage of people 16 and over, in the labor force, and unemployed
MalePctDivorce	Numerical	[0.0 - 1.0]	0	The percentage of males who are divorced
FemalePctDiv	Numerical	[0.0 - 1.0]	0	The percentage of females who are divorced
TotalPctDiv	Numerical	[0.0 - 1.0]	0	The percentage of population who are divorced
PersPerFam	Numerical	[0.0 - 1.0]	0	The mean number of people per family
PctKids2Par	Numerical	[0.0 - 1.0]	0	The percentage of kids in family housing with two parents
PctYoungKids2Par	Numerical	[0.0 - 1.0]	0	The percentage of kids 4 and under in two parent households
PctTeen2Par	Numerical	[0.0 - 1.0]	0	The percentage of kids age 12-17 in two parent households
NumIlleg	Numerical	[0.0 - 1.0]	0	The number of kids born to never married
PctIlleg	Numerical	[0.0 - 1.0]	0	The percentage of kids born to never married
PctPersOwnOccup	Numerical	[0.0 - 1.0]	0	The percentage of people in owner occupied households
HousVacant	Numerical	[0.0 - 1.0]	0	The number of vacant households
PctHousOwnOcc	Numerical	[0.0 - 1.0]	0	The percentage of households owner occupied
PctVacantBoarded	Numerical	[0.0 - 1.0]	0	The percentage of vacant housing that is boarded up
NumInShelters	Numerical	[0.0 - 1.0]	0	The number of people in homeless shelters
NumStreet	Numerical	[0.0 - 1.0]	0	The number of homeless people counted in the street
ViolentCrimesPerPop	Numerical	[0.0 - 1.0]	0	The total number of violent crimes per 100,000 popuation

15.34.

Protected attributes: In the literature (Kamishima, Akaho, Asoh, & Sakuma, 2012; Kamiran, Žliobaitė, & Calders, 2013), typically, researchers derive a new attribute, namely *Black*, which is considered as the protected attribute, in order to divide the communities according to race by thresholding the attribute *racepctblack* (the percentage of the population that is African American) at 0.06. The ratio of *black:non-black* is 1,038:956 (52.1%:47.9%). The interesting point in the data is that 94.3% (115/122) of the class "high-crime" are communities dominated by *blacks*.

Bayesian network: The dataset contains 122 numerical attributes normalized in the range of $(0, 1)$, which is not competent to the Bayesian network. Hence, we use the median value 0.5 as a threshold to transform these attributes into categorical with two values $\{\leq 0.5, > 0.5\}$. Besides, to ensure the visibility of the chart and the computation time, we use 21 attributes that have a high correlation (at a threshold of 0.25) with the class label. The Bayesian network is visualized in Figure 20. In which, the percentage of *kids born to never married* (*PctIlleg*) and the percentage of *kids in family housing with two parents* (*PctKids2Par*) have a direct impact on the class label and the race. Looking into the dataset, we discover that 92.4% of the communities are dominated by *black* people, where the percentage of *kids in family housing with two parents* less than 50%), while only 55.6% of the communities are dominated by *black* people, where the percentage of *kids in family housing with two parents* greater than 50%.

Summary of the criminological datasets: In summary, the criminological datasets were only surveyed in the US.

Race and *sex* are considered as protected attributes, with *race* being the most prevalent protected attribute. Historical bias w.r.t *race* has been detected in the data, but comprises a challenge for ML models. Furthermore, the datasets consists of many attributes (the richer description among all datasets, see Table 1); hence, a careful selection of attributes for fairness-aware learning is required.

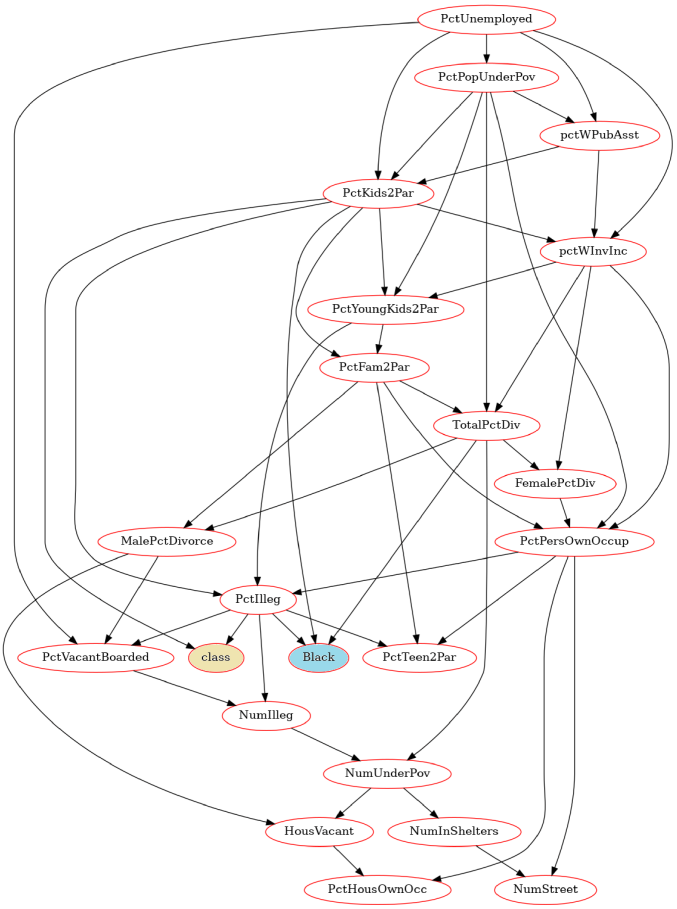


Figure 20: Communities and Crime: Bayesian network (class label: *class*, protected attribute: *black*)

3.3 Healthcare and social datasets

3.3.1 Diabetes dataset

The diabetes[18] dataset (Strack et al., 2014) describes the clinical care at 130 US hospitals and integrated delivery networks from 1999 to 2008. The classification task is to predict whether a patient will readmit within 30 days. The dataset is investigated in several studies (Appendix A).

Dataset characteristics: The dataset contains 101,766 patients described by 50 attributes (10 numerical, 7 binary and 33 categorical). Characteristics of all attributes[19] are summarized in Table 10 and Table 20 (in Appendix B). The attributes *encounter_id* and *patient_nbr* should not be considered in the learning tasks since they are the ID of the patients. Typically, *weight, payer_code, medical_specialty* attributes are removed because they contains at least 40% of missing values. Furthermore, we eliminate the missing values in *race, diag_1, diag_2, diag_3* columns. The class label *readmitted* contains 54,864 rows with "no record of readmission", hence, these

[18] https://archive.ics.uci.edu/ml/datasets/diabetes+130-us+hospitals+for+years+1999-2008

[19] Table 10 describes attributes used in the Bayesian network

rows should be clean. The clean version of the dataset contains 45,715 records.

Table 10: Diabetes: attributes characteristics

Attributes	Type	Values	#Missing values	Description
race	Categorical	6	2,273	Race (Caucasian, Asian, African American, Hispanic, and other)
gender	Categorical	3	0	Gender (male, female, and unknown/invalid)
age	Categorical	10	0	Grouped in 10-year intervals
time_in_hospital	Numerical	[1 - 14]	0	The number of days between admission and discharge
num_procedures	Numerical	[0 - 6]	0	The number of procedures (other than lab tests) performed during the encounter
num_medications	Numerical	[1 - 81]	0	The number of distinct generic names administered during the encounter
number_outpatient	Numerical	[0 - 42]	0	The number of outpatient visits of the patient in the year preceding the encounter
number_emergency	Numerical	[0 - 76]	0	The number of emergency visits of the patient in the year preceding the encounter
number_inpatient	Numerical	[0 - 21]	0	The number of inpatient visits of the patient in the year preceding the encounter
A1Cresult	Categorical	4	0	The range of the results or if the test was not taken
metformin	Categorical	4	0	Whether the drug was prescribed or there was a change in the dosage
chlorpropamide	Categorical	4	0	Whether the drug was prescribed or there was a change in the dosage
glipizide	Categorical	4	0	Whether the drug was prescribed or there was a change in the dosage
rosiglitazone	Categorical	4	0	Whether the drug was prescribed or there was a change in the dosage
acarbose	Categorical	4	0	Whether the drug was prescribed or there was a change in the dosage
miglitol	Categorical	4	0	Whether the drug was prescribed or there was a change in the dosage
diabetesMed	Binary	{Yes, No}	0	Was there any diabetic medication prescribed?
readmitted	Categorical	3	0	The number of days to inpatient readmission (No, < 30, > 30)

Protected attributes: Typically *Gender* $= \{$*male, female*$\}$ is chosen as the protected attribute. The ratio of *male:female* is 20,653:25,062 (45.2%:54.8%). The ratio of males or females who have to readmit hospital in less than 30 days is approximately 24%.

Class attribute: The class attribute is *readmitted* $\in \{< 30, > 30\}$ indicating whether a patient will readmit within 30 days. The positive class is "< 30". The dataset is imbalanced with an IR $1 : 3.13$ (positive:negative).

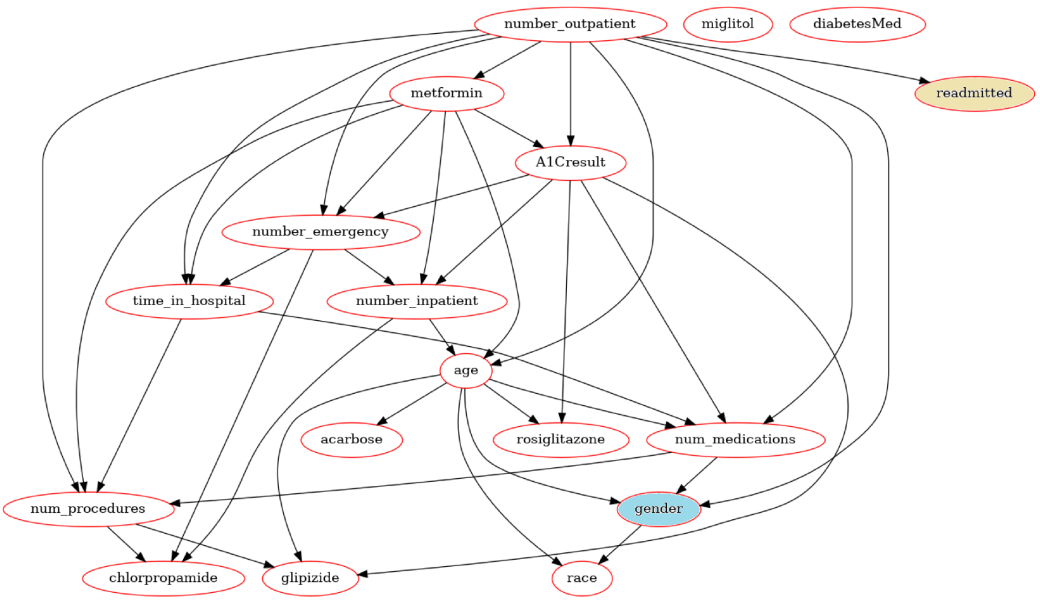


Figure 21: Diabetes: Bayesian network (class label: *readmitted*, protected attribute: *gender*)

Bayesian network: To prepare the dataset for Bayesian network generating process, we encode the attributes: *age* $= \{$<40, 40-59, 60-79, 80-99$\}$; *time_in_hospital* $= \{\leq 5, >5\}$; *num_lab_-*

procedures = {≤50, 50}; *num_procedures* = {≤1, >1}; *number_outpatient* = {0, >0}; *num_medications* = {≤15, >15}; *number_emergency* = {0, >0}; *number_inpatient* = {0, >0}; *number_diagnoses* = {0, >0}. To reduce the computation time, we use 17 attributes that have an absolute correlation coefficient higher than 0.005 with *gender* and *readmitted* attributes to generate the Bayesian network in Figure 21.

The class label *readmitted* is directly conditionally dependent on the number of outpatient visits of the patient in the year preceding the encounter (*number_outpatient*). The attribute *number_outpatient* also has an impact on 8 other features. Interestingly, there is no connection between the protected attribute *gender* and the class label.

3.3.2 Ricci dataset

The Ricci[20] dataset was generated by the Ricci v.DeStefano case (Supreme Court of the United States, 2009), in which they investigated the results of a promotion exam within a fire department in Nov 2003 and Dec 2003. Although it is a relatively small dataset, it has been employed for fairness-aware classification tasks in many studies (Appendix A). The classification task is to predict whether an individual obtains a promotion based on the exam results.

Dataset characteristics: The dataset consists of 118 samples, where each sample is characterized by 6 attributes (3 numerical and 3 binary attributes), as presented in Table 11.

Table 11: Ricci: attributes characteristics

Attributes	Type	Values	#Missing values	Description
Position	Binary	{Lieutenant, Captain}	0	The desired promotion
Oral	Numerical	[40.83 - 92.08]	0	The oral exam score
Written	Numerical	[46 - 95]	0	The written exam score
Race	Binary	{White, Non-White}	0	Race
Combine	Numerical	[45.93 - 92.80]	0	The combined score (the written exam gets 60% weight)
Promoted	Binary	{True, False}	0	Whether an individual obtains a promotion or not

Protected attributes: In this dataset, only attribute *race* can be used as a protected attribute. *Race* contains three values (*black, white*, and *hispanic*). As described in the literature, "black" and "hispanic" are grouped as "non-white" community. The ratio of *white:non-white* is 68:50 (57.6%:42.4%).

Class attribute: The class attribute is *promoted* $\in \{True, False\}$ revealing whether an individual achieves a promotion or not. The positive class is "True". The dataset is almost balanced with an IR $1 : 1.11$ (positive:negative).

Bayesian network: We encode 3 numerical attributes *oral, written* and *combine* as following: *oral* = {<70, ≥70}, *written* = {<70, ≥70}, *combine* = {<70, ≥70}. The Bayesian network of the Ricci dataset is demonstrated in Figure 22.

It is easy to observe that the combined grade (attribute *combine*) has a direct effect on the class label (*promoted*). Figure 23 illustrates the relationship between the combined grade and the promotion status. 100% of people whose combined oral and written exams are equal to or above 70 are promoted. Besides, as depicted in Figure 24, the number of promotions are granted for *white* people is higher than that for *non-white* people. The opposite trend is true in the group of candidates with no promotion.

Summary of the healthcare and social datasets: In summary, the datasets in healthcare and society domains were only surveyed in the US. *Race* and *gender* are considered as protected

[20] https://www.key2stats.com/data-set/view/690

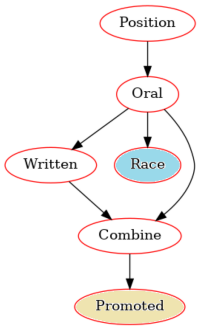


Figure 22: Ricci: Bayesian network (class label: *promoted*, protected attribute: *race*)

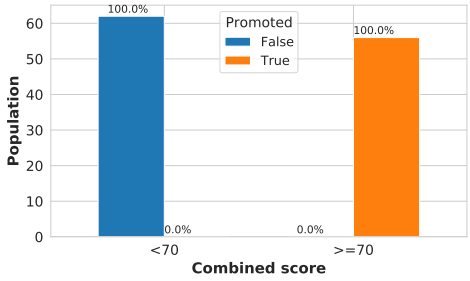


Figure 23: Ricci: Relationship between *combined score* and *promotion status*

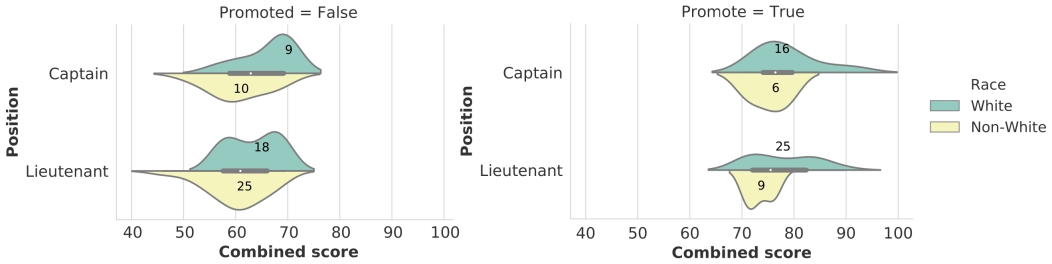


Figure 24: Ricci: Distribution of *combined score*, *position* and *promotion decision* w.r.t. *race*

attributes. In terms of class imbalance, these datasets are less imbalanced than datasets in other domains, although the Diabetes dataset is still imbalanced. Interestingly, there is no connection between the protected attribute and the class label in both two datasets, which implies fairness can be observed in the results of fairness-aware ML models.

3.4 Educational datasets

3.4.1 Student performance dataset

The student performance dataset (Cortez & Silva, 2008) described students' achievement in the secondary education of two Portuguese schools in 2005 - 2006 with two distinct subjects:

Mathematics and Portuguese.[21]. The regression task is to predict the final year grade of the students. It is investigated in several researches (Appendix A) with fairness-aware regression and clustering approaches.

Dataset characteristics: The dataset contains information of 395 (Mathematics subject) and 649 (Portuguese subject) students described by 33 attributes (4 categorical, 13 binary and 16 numerical attributes). Characteristics of all attributes is described in Table 12. To simply the classification problem, we create a class label based on attribute *G3*, *class* = {*Low, High*}, corresponding to *G3* = {$<$10, $\geq$10}. The positive class is "High". The dataset is imbalanced with imbalance ratios 1:2.04 (Mathematics subject) and 1:5.09 (Portuguese subject).

Table 12: Student performance: attributes characteristics

Attributes	Type	Values	#Missing values	Description
school	Binary	{GP, MS}	0	The student's school ('GP': Gabriel Pereira, 'MS': Mousinho da Silveira)
sex	Binary	{Male, Female}	0	Sex
age	Numerical	[15 - 22]	0	Age (in years)
address	Binary	{U, R}	0	The address type ('U': urban, 'R':rural)
famsize	Binary	{LE3, GT3}	0	The family size ('LE3': less or equal to 3, 'GT3': greater than 3)
Pstatus	Binary	{T, A}	0	The parent's cohabitation status ('T': living together, 'A': apart)
Medu	Numerical	[0 - 4]	0	Mother's education
Fedu	Numerical	[0 - 4]	0	Father's education
Mjob	Categorical	5	0	Mother's job
Fjob	Categorical	5	0	Father's job
reason	Categorical	4	0	The reason to choose this school
guardian	Categorical	3	0	The student's guardian (mother, father, other)
traveltime	Numerical	[1 - 4]	0	The travel time from home to school
studytime	Numerical	[1 - 4]	0	The weekly study time
failures	Numerical	[0 - 3]	0	The number of past class failures
schoolsup	Binary	{Yes, No}	0	Is there an extra educational support?
famsup	Binary	{Yes, No}	0	Is there any family educational support?
paid	Binary	{Yes, No}	0	Is there an extra paid classes within the course subject (Math or Portuguese)
activities	Binary	{Yes, No}	0	Are there extra-curricular activities?
nursery	Binary	{Yes, No}	0	Did the student attend a nursery school?
higher	Binary	{Yes, No}	0	Does the student want to take a higher education?
internet	Binary	{Yes, No}	0	Does the student have an Internet access at home?
romantic	Binary	{Yes, No}	0	Does the student have a romantic relationship with anyone?
famrel	Numerical	[1 - 5]	0	The quality of family relationships (from 1: very bad to 5: excellent)
freetime	Numerical	[1 - 5]	0	Free time after school (from 1: very low to 5: very high)
goout	Numerical	[1 - 5]	0	How often does the student go out with friends? (from 1: very low to 5: very high)
Dalc	Numerical	[1 - 5]	0	The workday alcohol consumption (from 1: very low to 5: very high)
Walc	Numerical	[1 - 5]	0	The weekend alcohol consumption (from 1: very low to 5: very high)
health	Numerical	[1 - 5]	0	The current health status (from 1: very bad to 5:very good)
absences	Numerical	[0 - 32]	0	The number of school absences
G1	Numerical	[0 - 19]	0	The first period grade
G2	Numerical	[0 - 19]	0	The second period grade
G3	Numerical	[0 - 19]	0	The final grade

Protected attributes: Typically, in the literature, *sex* is considered as the protected attribute. In the work of (Kearns, Neel, Roth, & Wu, 2019; Deepak & Abraham, 2020), they also select *age* as the protected attribute. Especially, in the research (Kearns et al., 2019), they consider atttributes *romatic* (relationship) and *dalc, walc* (alcohol consumption) as the protected attributes. However, because of the unpopularity of these attributes, we did not consider those within the scope of this paper.

- *sex* = {*male, female*}: the dataset is dominated by female students. The ratios of *male:female* are 208:187 (52.7%:47.3%) and 383:266 (59%:41%) for the Mathematics subject and Portuguese subject, respectively.
- *age* = {$<$*18,* $\geq$ *18*}: young students (less than 18 years old) are the majority with the ratios of "< 18":"≥ 18" are 284:111 (71.9%: 28.1%) and 468:181 (72.1%:27.9%) for the

[21] https://archive.ics.uci.edu/ml/datasets/student+performance

Mathematics subject and Portuguese subject, respectively.

Bayesian network: We perform a transformation of numerical variables: the number of school absences, *absences* = {0-5, 6-20, >20}; grade *G1* = {<10, ≥10}; *G2* = {<10, ≥10}. Due to the computation of the Bayesian network generator and the correlation coefficient with the class label (with a threshold of 0.02), we select 26 variables for the network. The Bayesian networks of the dataset on Portuguese and Mathematics subjects are visualized in Figure 25 and Figure 26, respectively.

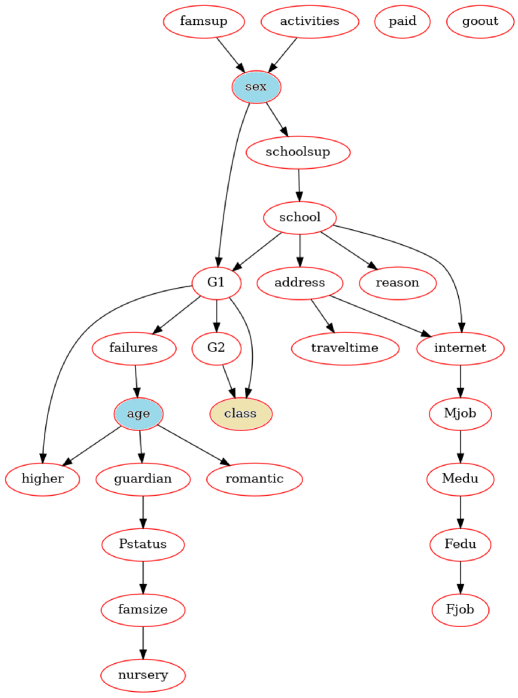


Figure 25: Student performance - Portuguese subject: Bayesian network (class label: *class*, protected attributes: *age*, *sex*)

The *class label* attribute is conditionally dependent on the grade *G2* in both subsets (Mathematics and Portuguese subjects). This is explained by a very high correlation coefficient (above 90%) between *G2* and *G3* variables. In addition, we investigate the distribution of the final grade *G3* on *sex* because the attribute *sex* has an indirect relationship with the *class label*. Figure 27 reveals that the male students tend to receive high scores in the Portuguese subject, while the scores of Math are relatively evenly distributed across both sexes.

3.4.2 OULAD dataset

The Open University Learning Analytics (OULAD) dataset[22] was collected from the OU analysis project (Kuzilek, Hlosta, & Zdrahal, 2017) of The Open University (England) in 2013 - 2014. The dataset contains information of students and their activities in the virtual learning environment (VLE) for 7 courses. The dataset is investigated in several papers (Appendix A), on fairness-aware problems. The goal is to predict the success of students.

[22] https://analyse.kmi.open.ac.uk/open_dataset

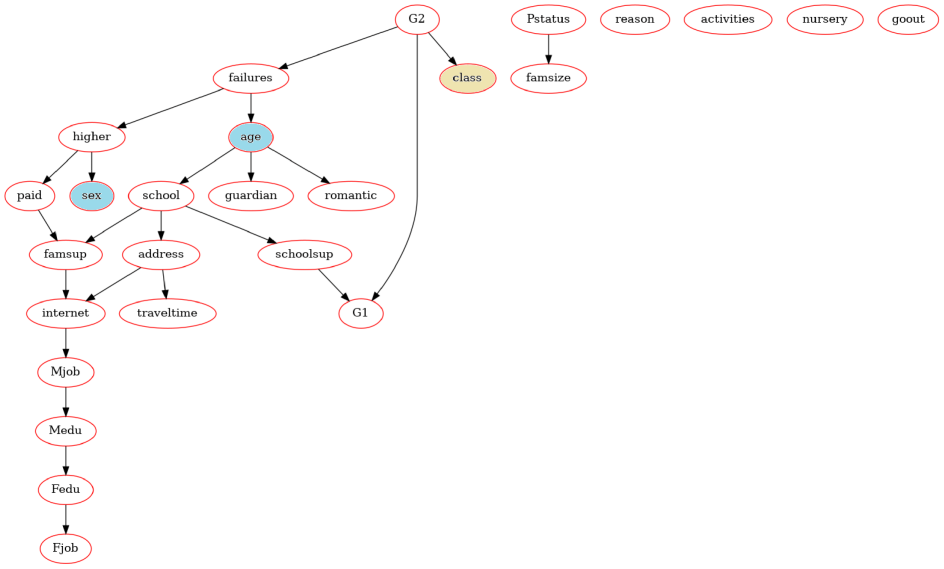
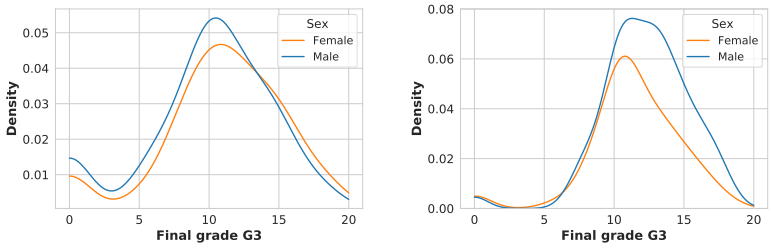


Figure 26: Student performance - Mathematics subject: Bayesian network (class label: *class*, protected attributes: *age, sex*)



(a) Mathematics subject (b) Portuguese subject

Figure 27: Student performance: Distribution of the final grade G3 w.r.t. *sex*

Dataset characteristics: The dataset contains information of 32,593 students characterized by 12 attributes (7 categorical, 2 binary and 3 numerical attributes). An overview of all attributes is illustrated in Table 13. Attribute *id_student* should be ignored in the analysis. Typically, in the related work, they consider the prediction task on the class label *final_result* $= \{$*pass, fail*$\}$. Therefore, we investigate the cleaned dataset with 21,562 instances after removing the missing values and rows with *final_result* $=$ *"withdrawn"*. "Pass" is the positive class. The ratio of *pass:fail* is 14,655:6,907 (68%:32%). In other words, the dataset is imbalanced with the IR is 2.12:1 (positive:negative).

Protected attributes: *gender* $= \{$*male, female*$\}$ is considered as the protected attribute, in the literature. Male is the majority group with the ratio *male:female* is 11,568:9994 (56.6%:46.4%).

Bayesian network: The numerical attributes are encoded for generating the Bayesian network: *num_of_prev_attempts* $= \{0, >0\}$, *studied_credits* $= \{\leq 100, >100\}$. The network is depicted in Figure 28. The final result attribute is directly conditionally dependent on the highest education

Table 13: OULAD: attributes characteristics

Attributes	Type	Values	#Missing values	Description
code_module	Categorical	7	0	The identification code of the module on which the student is registered
code_presentation	Categorical	4	0	The identification code of the presentation on which the student is registered
id_student	Numerical	[3,733 - 2,716,795]	0	A unique identification number for the student
gender	Binary	{Male, Female}	0	Gender
region	Categorical	13	0	The geographic region
highest_education	Categorical	5	0	The highest student education level
imd_band	Categorical	10	1111	The index of multiple deprivation (IMD) band of the place where the student lived
age_band	Categorical	3	0	The category of the student's age
num_of_prev_attempts	Numerical	[0 - 6]	0	The number times the student has attempted this module
studied_credits	Numerical	[30 - 655]	0	The total number of credits for the modules the student is currently studying
disability	Binary	{Yes, No}	0	Whether the student has declared a disability
final_result	Categorical	4	0	The student's final result (in the module-presentation)

level (*highest_education*) and the number times the student has attempted the module (*num_of_prev_attempts*) attributes, while *gender* has a more negligible effect on the outcome.

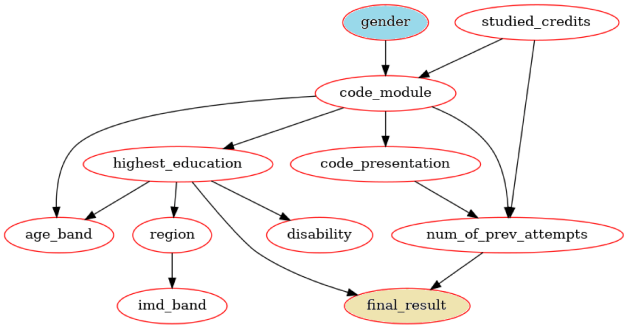


Figure 28: OULAD: Bayesian network (class label: *final_result*, protected attributes: *gender*)

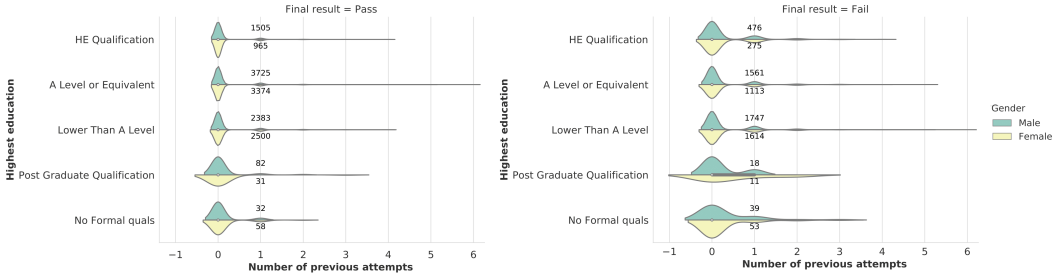


Figure 29: OULAD: Distribution of *the number of previous attempts*, *the highest education* and *the final result* w.r.t. *gender*

We perform the analysis on the relationship of the highest education, number of previous attempts and the final result for each gender. As demonstrated in Figure 29, students have a higher probability of failing when they tried to attempt the exam many times in the past. The ratio of male students having the *highest education* is "A-level or equivalent" or "higher education (HE) qualification" is around 1.5 times higher than that of female students.

3.4.3 Law school dataset

The Law school[23] dataset (Wightman, 1998) was conducted by a Law School Admission Council (LSAC) survey across 163 law schools in the United States in 1991. The dataset contains the law school admission records. The prediction task is to predict whether a candidate would pass the bar exam or predict a student's first-year average grade (FYA). The dataset is investigated in a variety of studies (Appendix A).

Dataset characteristics: The dataset contains information of 20,798 students characterized by 12 attributes (3 categorical, 3 binary and 6 numerical attributes). An overview of all attributes is depicted in Table 14.

Table 14: Law schoool: attributes characteristics

Attributes	Type	Values	#Missing values	Description
decile1b	Numerical	[1.0 - 10.0]	0	The student's decile in the school given his grades in Year 1
decile3	Numerical	[1.0 - 10.0]	0	The student's decile in the school given his grades in Year 3
lsat	Numerical	[11.0 - 48.0]	0	The student's LSAT score
ugpa	Numerical	[1.5 - 4.0]	0	The student's undergraduate GPA
zfygpa	Numerical	[-3.35 - 3.48]	0	The first year law school GPA
zgpa	Numerical	[-6.44 - 4.01]	0	The cumulative law school GPA
fulltime	Binary	{1, 2}	0	Whether the student will work full-time or part-time
fam_inc	Categorical	5	0	The student's family income bracket
male	Binary	{0, 1}	0	Whether the student is a male or female
tier	Categorical	6	0	Tier
race	Categorical	6	0	Race
pass_bar	Binary	{0, 1}	0	Whether the student passed the bar exam on the first try

Protected attributes: In the literature, *race* (Bechavod & Ligett, 2017; Lahoti et al., 2020; Russell, Kusner, Loftus, & Silva, 2017; Kusner, Loftus, Russell, & Silva, 2017; Chzhen, Denis, Hebiri, Oneto, & Pontil, 2020; Kearns et al., 2019; Ruoss, Balunovic, Fischer, & Vechev, 2020; Yang, Cisse, & Koyejo, 2020) and *male* (Berk et al., 2017; Lahoti et al., 2020; Kusner et al., 2017; Kearns et al., 2019; Yang et al., 2020) are considered as the protected attributes.

- *male* $= \{1, 0\}$. *Male* is the majority group. The ratio of *male* (1):*female* (0) is 11,675:9,123 (56.1%:43.9%).
- *race* $= \{$*white, black, Hispanic, Asian, other*$\}$. As introduced in the related work, we encode *race* $= \{$*white, non-white*$\}$ based on the original attribute. *Non-white* is the minority group with the ratio *white:non-white* is 17,491:3,307 (84%:16%).

Class attribute: The class label *pass_bar* $= \{0, 1\}$ is used for the classification task. The positive class is *1 - pass*. The dataset is imbalanced with an imbalance ratio 8.07:1 (positive:negative).

Bayesian network: To generate the Bayesian network, we encode the numerical attributes as follows: *decile1b* $= \{\leq 5, >5\}$, *decile3* $= \{\leq 5, >5\}$, *lsat* $= \{37, >37\}$, *ugpa* $= \{<3.3, \geq 3.3\}$, *zgpa* $= \{\leq 0, >0\}$, *zfygpa* $= \{\leq 0, >0\}$. The Bayesian network is visualized in Figure 30.

It is easy to observe that the bar exam's result is conditionally dependent on the law school admission test (LSAT) score, undergraduate grade point average (UGPA) and *Race*. We discover that 92.1% of *white* students (16,114/17,491) pass the bar exam, while this ratio in *non-white* students is only 72.3%. In general, the percentage of students who passed the bar exam is increased in proportion to the LSAT and UGPA scores, which is depicted in Figure 31.

[23] https://github.com/tailequy/fairness_dataset/tree/main/Law_school

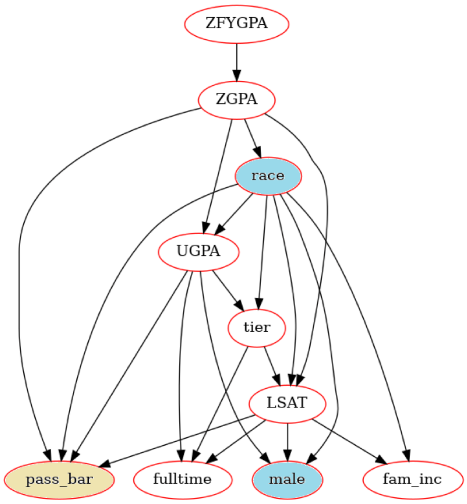
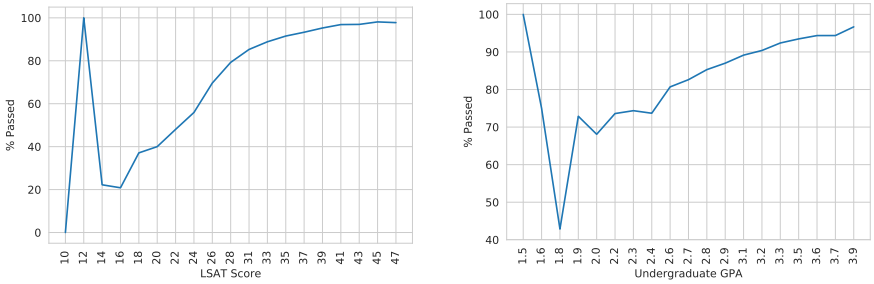


Figure 30: Law school: Bayesian network (class label: *pass_bar*, protected attributes: *male, race*)



(a) LSAT score

(b) Undergraduate GPA

Figure 31: Law school: The percentage of students that passed the bar exam by LSAT and UGPA scores

Summary of the educational datasets: The educational datasets were collected in many countries around the world. *Gender* is the most popular protected attribute, followed by *age* and *race*. The typical learning task is to predict students' outcome or grades. Therefore, many machine learning tasks are applied to the datasets, such as classification, regression, or clustering. All datasets are imbalanced with very different imbalance ratios in terms of class imbalance. The bias is observed in the datasets w.r.t protected attributes, i.e., *race, sex*; hence, fairness-aware algorithms need to take into account these attributes to achieve fairness in education

4 Experimental evaluation

The goal of our survey is to summarize the different datasets on fairness-aware learning in terms of their application domain, fairness-aware and learning-related challenges. An experimental evaluation of the different fairness-aware learning methods (pre-,in-and post-processing) is be-

yond the scope of this survey. However, in order to characterize the different datasets in terms of the difficulty of the fairness-aware learning task, in this section, we present a short fairness-vs-predictive performance evaluation[24] using a popular classification method (namely, logistic regression).

4.1 Evaluation setup

Predictive model. As our classification model, we use *logistic regression* (Cox, 1958), a statistical model using a logistic function to model a binary dependent variable. To simplify the task, we apply the logistic regression model to the binary classification problem.

Metrics. Based on the confusion matrix in Figure 32 (in which, *prot* and *non-prot* stand for *protected, non-protected*, respectively), we report the performance of the predictive model on the following measures.

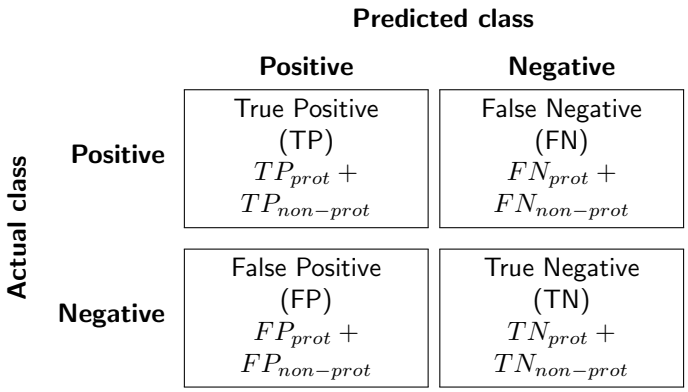


Figure 32: The confusion matrix, including protected/ non-protected groups.

- Accuracy

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{8}$$

- Balanced accuracy

$$Balanced\,accuracy = \frac{1}{2} \times \left(\frac{TP}{TP + FN} + \frac{TN}{TN + FP} \right) \tag{9}$$

- True positive rate (TPR) on protected group

$$TPR_{prot} = \frac{TP_{prot}}{TP_{prot} + FN_{prot}} \tag{10}$$

- TPR on non-protected group

$$TPR_{non-prot} = \frac{TP_{non-prot}}{TP_{non-prot} + FN_{non-prot}} \tag{11}$$

[24]The source code is available at: https://github.com/tailequy/fairness_dataset

- True negative rate (TNR) on protected group

$$TNR_{prot} = \frac{TN_{prot}}{TN_{prot} + FP_{prot}} \tag{12}$$

- TNR on non-protected group

$$TNR_{non-prot} = \frac{TN_{non-prot}}{TN_{non-prot} + FP_{non-prot}} \tag{13}$$

- Statistical parity (Eq. 4)
- Equalized odds (Eq. 6)
- ABROCA (Eq. 7)

Training/test set spliting. The ratio of training set and test set in our experiment is 70%:30% (single split) applied for each dataset.

4.2 Experimental results

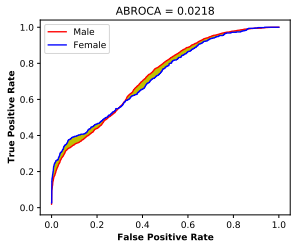
Table 15 describes the performance of the logistic regression model on all datasets. We believe that our experimental results can be considered as the baseline for the researchers' future studies.

Table 15: Predictive- and fairness-relatedperformance of logistic regression model

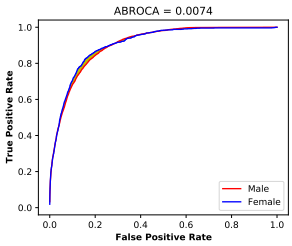
Dataset	Protected attribute	Group distribution (%) $[s_+, s_-, \bar{s}_+, \bar{s}_-]$	Accuracy	Balanced accuracy	Statistical Parity	Equalized odds	ABROCA	TPR prot.	TPR non-prot.	TNR prot.	TNR non-prot.
Adult	Sex	[3.7, 28.8, 21.1, 46.4]	0.7864	0.6249	0.0555	0.0281	0.0218	0.3194	0.3007	0.9521	0.9426
KDD Census-Income	Sex	[1.3, 50.7, 4.8, 43.2]	0.9474	0.6031	0.0198	0.0403	0.0074	0.1825	0.2195	0.9961	0.9928
German credit	Sex	[20.1, 10.9, 49.9, 19.1]	0.6967	0.5713	-0.0770	0.1634	0.1228	0.9831	0.8533	0.2759	0.2419
Dutch census	Sex	[16.4, 33.7, 31.2, 18.7]	0.8149	0.8138	0.3568	0.3746	0.0202	0.6984	0.8382	0.9219	0.6871
Bank marketing	Marital	[5.6, 34.2, 6.1, 54.1]	0.8855	0.5720	0.0153	0.0261	0.025	0.1527	0.1726	0.9849	0.9787
Credit card clients	Sex	[12.5, 47.8, 9.7, 30.0]	0.7822	0.5	**0.0**	**0.0**	0.0220	0.0	0.0	**1.0**	**1.0**
COMPAS recid.	Race	[31.5, 28.7, 15.5, 24.3]	0.6414	0.6299	-0.3398	0.6452	0.0675	0.5996	0.2058	0.6793	0.9307
COMPAS viol. recid.	Race	[12.0, 44.8, 5.2, 38.0]	0.8432	0.5541	-0.0659	0.2195	0.0584	0.1826	0.0	0.9606	0.9975
Communites & Crime	Black	[5.8, 46.3, 0.3, 47.6]	0.9683	0.7011	0.0396	0.4507	0.031	0.0	0.44	1.0	0.9892
Diabetes	Gender	[11.1, 34.1, 13.1, 41.7]	0.7584	0.5	**0.0**	**0.0**	0.0189	0.0	0.0	**1.0**	**1.0**
Ricci	Race	[12.7, 29.7, 34.7, 22.9]	**1.0**	**1.0**	0.1714	**0.0**	**0.0**	**1.0**	**1.0**	**1.0**	**1.0**
Student - Mathematics	Sex	[33.7, 19.0, 33.4, 13.9]	0.9412	0.9360	0.2041	0.1616	0.0177	0.9354	0.9762	0.9630	0.8421
Student - Portuguese	Sex	[51.3, 7.7, 33.3, 7.7]	0.9282	0.8447	-0.0682	0.0490	0.0273	0.9633	0.95	0.75	0.7143
OULAD	Gender	[32.1, 14.2, 35.9, 17.8]	0.6751	0.5	**0.0**	**0.0**	0.0088	**1.0**	**1.0**	0.0	0.0
Law School	Race	[11.5, 4.4, 77.5, 6.6]	0.9072	0.6260	0.1937	0.5043	0.0325	0.9100	0.9955	0.5251	0.1063

In general, a significant difference in terms of predictive performance and fairness measures is observed between the datasets. In particular, the *Ricci* dataset is an exception where the performance of the predictive model reaches the peak regarding both accuracy and fairness measures. Apart from that, the logistic regression model shows the best performance on the *Communities & Crime* dataset in terms of accuracy. The worst accuracy is seen in the result of the model on the *OULAD* dataset. Regarding balanced accuracy, the *Student - Mathematics* is the dataset showing the best result of the predictive model, followed by the *Student - Portuguese* and the *Dutch census* datasets. Logistic regression model shows the worst balanced accuracy on the *Credit card clients, Diabetes* and *OULAD* datasets.

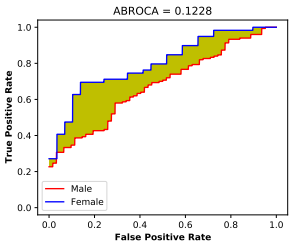
Regarding the statistical parity measure, in general, 10/15 datasets have an absolute value of statistical parity less than 0.1. The *Diabetes, Credit card clients* and *OULAD* datasets have the best value (0.0) of statistical parity while the *Bank marketing* dataset has the worst value.



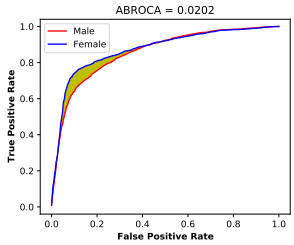
(a) Adult



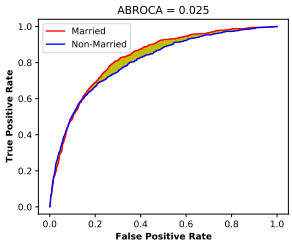
(b) KDD Census-Income



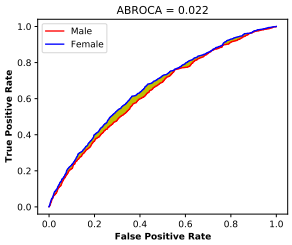
(c) German credit



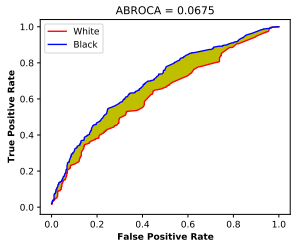
(d) Dutch census



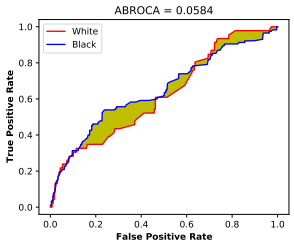
(e) Bank marketing



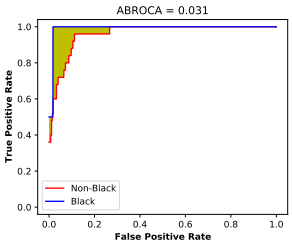
(f) Credit card clients



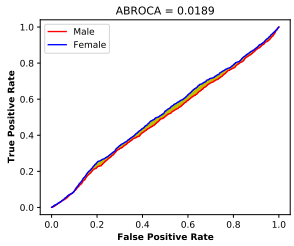
(g) COMPAS recid.



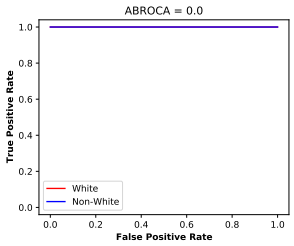
(h) COMPAS viol. recid.



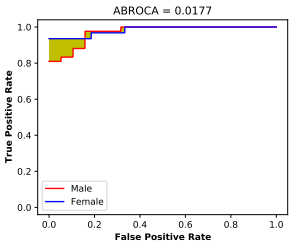
(i) Communities & Crime



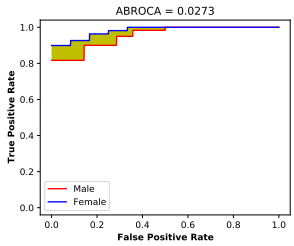
(j) Diabetes



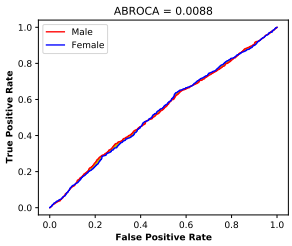
(k) Ricci



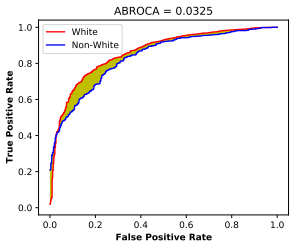
(l) Student-Mathematics



(m) Student-Portuguese



(n) OULAD



(o) Law school

Figure 33: ABROCA slice plot on datasets

Interestingly, in terms of the equalized odds measure, the best value (0.0) is observed in four datasets (*Credit card clients, Diabetes, OULAD* and *Ricci*). The predictive model results in the worst performance on the *COMPAS recid.* dataset with a high value of equalized odds, followed by the *Law school* and the *Communities & Crime* datasets.

In addition, we plot the ABROCA slicing of all datasets in Figure 33. In the Figure, the *red* ROC curve represents the non-protected group (e.g., Male) while the *blue* ROC is the curve of the protected group (e.g., Female). The best value of the ABROCA is seen in the *Ricci* dataset, followed by the *OULAD* and the *KDD Census-Income* datasets. The worst cases are the *German credit* and the *COMPAS* datasets.

5 Open issues on datasets for fairness-aware ML

In the previous sections, we have summarized the most popular datasets for fairness-aware learning. In this section, we extend the discussion to also include recently proposed (and therefore, not adequately exploited) real datasets (Section 5.1), synthetic datasets (Section 5.2 and datasets for sequential decision making (Section 5.3). We advocate that the community should focus more on new datasets representing diverse fairness scenarios, in parallel to new methods and algorithms for fairness-aware learning.

5.1 Adult reconstruction and ACS PUMS datasets

The Adult reconstruction dataset[25](Ding, Hardt, Miller, & Schmidt, 2021) is a new dataset reconstructed from the Current Population Survey (CPS) data (Sarah Flood & Warren, 2020) from 1994. The dataset consists of 49,531 instances with 14 attributes, in which 13 of 15 attributes of the Adult dataset (see Section 3.1.1) are matched. Differently from the vanilla Adult dataset, the class attribute *income* is now represented as a continuous variable. A possible prediction task of the Adult reconstruction dataset is to decide whether an individual earns annually more than 50,000 US dollars. Apart from the Adult reconstruction dataset, (Ding et al., 2021) introduce further datasets based on the American Community Survey (ACS) Public Use Microdata Sample (PUMS)[26] with five new prediction tasks w.r.t. *income, public health insurance, residential address, employment* and *commuting time to workplace*.

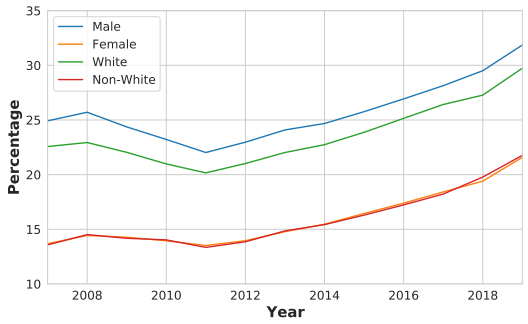


Figure 34: ACS PUMS dataset (California, 2007-2019): Proportion of people with an income above 50K$ in each group over the years

[25]https://github.com/zykls/folktables/
[26]https://www2.census.gov/programs-surveys/acs/data/pums/

For our study, we focus on a particular state (California) for a specific period (2007-2019), using the provided tool (Ding et al., 2021) and *income* as the prediction task. We consider two protected attributes: $Sex = \{male, female\}$ and $race = \{white, non\text{-}white\}$, with "female" and "non-white" being the corresponding protected values.

In Figure 34, we depict the proportion of people in each income class over time split per gender and race. It is easy to observe a lower representation of the protected groups (female, non-white) over the years. In relation to the population size, which is shown in Figure 35, the number of people with an income above 50K$ gradually increases in both sexes. However, the growth rate in the male group is slightly higher than that in the female group.

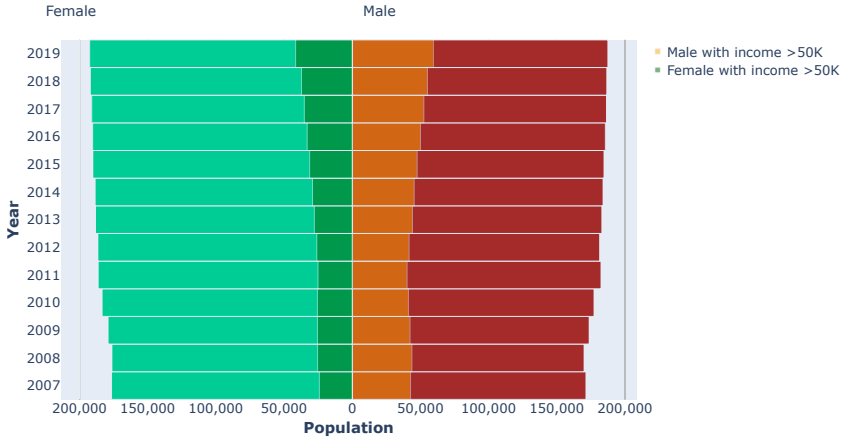


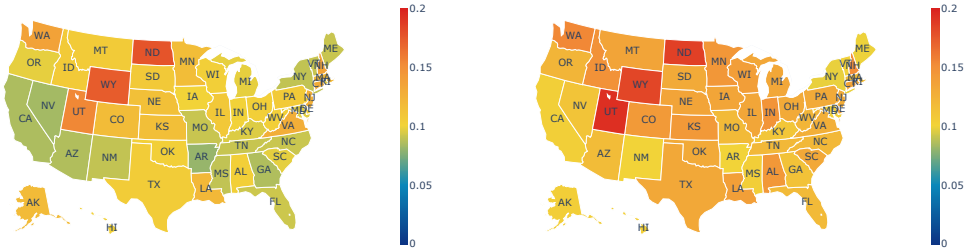
Figure 35: ACS PUMS dataset (California, 2007-2019): The distribution of people w.r.t *sex* and *income* over the years

The ACS PUMS datasets were only recently proposed (Ding et al., 2021). We believe they comprise a very interesting collection since they also contain spatial and temporal information, albeit only for the US, and can therefore be used to analyze the dynamics of discrimination across space and time. As a preliminary investigation, in Figure 36, we illustrate the gender percentage differences in the positive class, i.e., income over 50K$, for different US states in 2011 and 2019. Many states have low *gender differences* (depicted in green) in 2011. However, the *gender differences* increase over the years, as seen in 2019. A further investigation of the potential effect of spatial and temporal parameters is of course required.

5.2 Synthetic datasets

Apart from using real-world datasets, it is typical for machine learning evaluation (Ntoutsi, Schubert, Zimek, & Zimmermann, 2019) to also employ synthetic data which allow for evaluation under different learning complexity scenarios. Synthetic datasets have been also used for the evaluation of fairness-aware learning methods (Zafar, Valera, Gomez Rodriguez, & Gummadi, 2017; Loh, Cao, & Zhou, 2019; D'Amour et al., 2020; Tu et al., 2020; Reddy et al., 2021) to produce desired testing scenarios, which may not yet be captured by the existing real-world datasets, but are essential for the development and evaluation of theoretically sound fair algorithms.

For example, the works (D'Amour et al., 2020; Tu et al., 2020) study the long-term effects of a currently fair decision-making system and therefore require data that capture the decision of



(a) 2011 (b) 2019

Figure 36: ACS PUMS dataset: Gender differences (%males-%females) in the positive class ($> 50K$ income) for different US states

the classifier continuously through time and change the underlying population accordingly. To this end, they simulate dataset changes over time.

In a different direction, (Iosifidis & Ntoutsi, 2018) use synthetic data augmentation to increase the representation of the underrepresented protected groups in the overall population. The synthetic instances are generated via SMOTE (Chawla, Bowyer, Hall, & Kegelmeyer, 2002) by interpolating between original instances.

5.3 Sequential datasets

While algorithmic fairness in decision-making has been mostly studied in static/batch settings, increasing attention has been gained in sequential decision-making environments (Liu, Dean, Rolf, Simchowitz, & Hardt, 2018; Heidari & Krause, 2018; Wen, Bastani, & Topcu, 2021), where a sequence of instances possibly infinitely arrives continuously over time which widely exists in many real-world applications such as when making decisions about lending and employment. In contrast to the batch based static environments, sequential decision-making requires the operating model processes each new individual at each time step while making an irrevocable decision based on observations made so far (W. Zhang & Bifet, 2020; W. Zhang & Ntoutsi, 2019). Often times, the processing also needs to be on the fly and without the need for storage and reprocessing (W. Zhang, Bifet, Zhang, Weiss, & Nejdl, 2021).

The aforementioned unique characteristics require the datasets being used for fair sequential decision-making studies fulfill these additional demanding requirements. Among the previously discussed datasets, the *Adult* (Kohavi, 1996) and *Census* (Asuncion & Newman, 2007) are rendered as discriminated data streams to fit for this purpose by processing the individuals in sequence (W. Zhang & Bifet, 2020). In addition, the datasets are ordered based on the sensitive attribute of their particular task at hand before sequential processing to further simulating the potential concept and fairness drifts in the online settings (W. Zhang et al., 2021). Revevantly, the *Crime and Communities* dataset (Asuncion & Newman, 2007) is also sequentially processed for sequential fairness-aware studies (Heidari & Krause, 2018). However, sequential-friendly datasets, due to their magnified requirements, are still in scarce, albeit their significance for the development of fair sequential models which are widely applicable in many real-world applications (W. Zhang, Tang, & Wang, 2019). A continuation on fair sequential datasets efforts is therefore required for a unified fairness-aware research. The new Adult dataset(s) (see Section 5.1) might be suitable for sequential learning as they contain temporal information (year of

data collection).

More recently, the uncertainty due to censorship in fair sequential decision-making has also been researched (W. Zhang & Weiss, 2021, 2022). Distinct from existing fairness studies assuming certainty on the class label by designed, this line of works addresses fairness in the presence of uncertainty on the class label due to censorship. Take the motivating clinical prediction therein as the example (e.g., SUPPORT dataset (Knaus et al., 1995)), whether the patient relapses/discharges (event of interest) could be unknown for various reasons (e.g., loss to followup) leading to uncertainty on the class label, i.e., censorship (W. Zhang, Tang, & Wang, 2016). This problem extends beyond the medical domain with examples in marketing analytics (e.g., KKBox dataset (Kvamme, Borgan, & Scheel, 2019)) and recidivism prediction instruments (e.g., ROSSI (Fox & Carvalho, 2012) and COMPAS dataset (Angwin et al., 2016)). The censorship information, including survival time and an event indicator, in addition to the observed features, is thus also included, which is normally excluded in fairness studies that do not consider censorship. As the exclusion of censorship information could lead to important information loss and introduce substantial bias (Wang, Zhang, Jadhav, & Weiss, 2021), more attention on the censorship of fairness datasets is warranted.

Related to the topic of fairness, is the topic of explainability. Explainability tools can help debugging ML models and uncover biased decision making. For sequential decision making, the notion of sequential counterfactuals (Naumann & Ntoutsi, 2021) seems prominent as it takes into account longer-term consequences of feature-value changes. The experiments were conducted on the Adult dataset, however the fairness of the decisions were not investigated. Further research in this direction is required.

6 Conclusion and outlook

There are several approaches and discussions that can be implemented in studies on fairness-aware ML. First, in this survey, we investigate the tabular data as the most prevalent data representation. However, in practice, other data types such as text (Zhao, Wang, Yatskar, Ordonez, & Chang, 2018) and images (Buolamwini & Gebru, 2018) are also used in fairness-aware machine learning problems. Obviously, these data types are closely related to the domain, and the method of handling data sets is also very different and specialized. This requires the fairness-aware algorithms to be tweaked to apply to different datasets.

Second, by generating the Bayesian network, we discover the relationship between attributes showing their conditional dependence. The results from data analysis and experiments show that the bias may appear in the data itself and/or in the outcome of predictive models. It is understandable that if a dataset contains bias and discrimination, it would be difficult for fairness-aware algorithms to find the trade-off between fairness requirement and performance. Furthermore, based on our experimental results, a significant variation in outcomes between the datasets suggests that the fairness-aware models need to be performed on the diverse datasets.

Third, bias and discrimination are the common problems of almost all domains in reality. In this paper, we study the well-known datasets describing the important aspects of social life such as finance, education, healthcare and criminology. The definition of fairness, of course, is different across domains. It isn't easy to evaluate the efficiency of fairness-aware algorithms because they must be based on such fairness notions. Therefore, it is crucial and necessary to select or define the appropriate fairness notions for each problem in each domain because there is no universal fairness notion for every problem. This remains a major challenge for researchers.

Fourth, the selection of the protected attributes is also a matter of consideration. In the datasets surveyed in this paper, *gender (sex), race, age* and *marriage* are the prevalent protected attributes. The selection of one or more protected attributes for the experiment depends on many factors such as domain, problem and the purpose of the experiment. In our experiments, for each dataset, we only demonstrate the performance of the predictive model w.r.t one of the most popular protected attributes. In addition, the identification and handling of "proxy" attributes is also an issue that requires more research.

Fifth, collecting new datasets is always a requirement of data scientists. The surveyed datasets were all collected quite a long time in the past with an average *age* of about 20 years. The oldest dataset was obtained 48 years ago, while the newest dataset was identified from 7 years ago. Of course, the newer the data, the more up-to-date with the trends of the modern society, so the analysis and application of fairness-aware algorithms on the new datasets will reflect the manifestations of the social behaviors more realistic. On the other hand, the old datasets are of reference value in comparing and contrasting the movement and variation of fairness in the same or different domains. The datasets are collected in the US and European countries where the data protection laws are in place. However, the general policies on data quality or collection still need to be studied and proposed (Ntoutsi et al., 2020).

To conclude, fairness-aware ML has attracted many recently in various domains from criminology, healthcare, finance to education. This paper reviews the most popular datasets used in fairness-aware ML researches. We explore the relationship of the variables as well as analyze their correlation concerning protected attributes and the class label. We believe our analysis will be the basis for developing frameworks or simulation environments to evaluate fairness-aware algorithms. In another aspect, an excellent understanding of well-known datasets can also inspire researchers to develop synthetic data generators because finding a suitable real-world dataset is never a simple task.

Funding Information

Ministry of Science and Education of LowerSaxony, German, project ID: 51410078

Acknowledgements

The work of the first author is supported by the Ministry of Science and Education of Lower Saxony, Germany, within the PhD programme "LernMINT: Data-assisted teaching in the MINT subjects". The work of the second author is supported by the Volkswagen Foundation under the call "Artificial Intelligence and the Society of the Future" (the BIAS project).

References

Abbasi, M., Bhaskara, A., & Venkatasubramanian, S. (2021). Fair clustering via equitable group representations. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 504–514). doi: https://doi.org/10.1145/3442188.3445913

Abraham, S. S., Sundaram, S. S., & Deepak, P. (2020). Fairness in clustering with multiple sensitive attributes. In *23rd International Conference on Extending Database Technology (EDBT)* (pp. 287–298).

Agarwal, A., Beygelzimer, A., Dudík, M., Langford, J., & Wallach, H. (2018). A reductions approach to fair classification. In *International Conference on Machine Learning* (pp. 60–69).

Ahn, Y., & Lin, Y.-R. (2019). Fairsight: Visual analytics for fairness in decision making. *IEEE transactions on visualization and computer graphics*, *26*(1), 1086–1095. doi: https://doi.org/10.1109/TVCG.2019.2934262

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). Machine bias. *ProPublica, May*, *23*.

Asuncion, A., & Newman, D. (2007). *Uci machine learning repository.* Irvine, CA, USA.

Backurs, A., Indyk, P., Onak, K., Schieber, B., Vakilian, A., & Wagner, T. (2019). Scalable fair clustering. In *International Conference on Machine Learning* (pp. 405–413).

Bechavod, Y., & Ligett, K. (2017). Learning fair classifiers: A regularization approach. *4th Workshop on Fairness, Accountability, and Transparency in Machine Learning (FAT/ML 2017)*.

Bera, S. K., Chakrabarty, D., Flores, N. J., & Negahbani, M. (2019). Fair algorithms for clustering. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems* (pp. 4954–4965).

Berk, R., Heidari, H., Jabbari, S., Joseph, M., Kearns, M., Morgenstern, J., ... Roth, A. (2017). A convex framework for fair regression. *4th Workshop on Fairness, Accountability, and Transparency in Machine Learning (FAT/ML 2017)*.

Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency* (pp. 77–91).

Calders, T., & Kamiran, F. (2010). Classification with no discrimination by preferential sampling. In *Proceeding 19th Machine Learning conference Belgium and the Netherlands*.

Calders, T., Kamiran, F., & Pechenizkiy, M. (2009). Building classifiers with independency constraints. In *IEEE international conference on data mining workshops, 2009. ICDMW'09.* (pp. 13–18). doi: https://doi.org/10.1109/ICDMW.2009.83

Calders, T., Karim, A., Kamiran, F., Ali, W., & Zhang, X. (2013). Controlling attribute effect in linear regression. In *2013 IEEE 13th International Conference on Data Mining (ICDM)* (pp. 71–80). doi: https://doi.org/10.1109/ICDM.2013.114

Calders, T., & Verwer, S. (2010). Three naive bayes approaches for discrimination-free classification. *Data Mining and Knowledge Discovery*, *21*(2), 277–292. doi: https://doi.org/10.1007/s10618-010-0190-x

Calmon, F. P., Wei, D., Vinzamuri, B., Ramamurthy, K. N., & Varshney, K. R. (2017). Optimized pre-processing for discrimination prevention. In *Proceedings of the 31st International Conference on Neural Information Processing Systems* (pp. 3995–4004).

Chakraborty, J., Majumder, S., Yu, Z., & Menzies, T. (2020). Fairway: A way to build fair ml software. In *Proceedings of the 28th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering* (pp. 654–665). doi: https://doi.org/10.1145/3368089.3409697

Chakraborty, J., Peng, K., & Menzies, T. (2020). Making fair ML software using trustworthy explanation. In *2020 35th IEEE/ACM International Conference on Automated Software Engineering (ASE)* (pp. 1229–1233).

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, *16*, 321–357. doi: https://doi.org/10.1613/jair.953

Chen, Y.-C., Wheeler, T. A., & Kochenderfer, M. J. (2017). Learning discrete bayesian networks from continuous data. *Journal of Artificial Intelligence Research*, *59*, 103–132. doi:

https://doi.org/10.1613/jair.5371
Chhabra, A., Masalkovaitė, K., & Mohapatra, P. (2021). An overview of fairness in clustering. *IEEE Access*. doi: https://doi.org/10.1109/ACCESS.2021.3114099
Chierichetti, F., Kumar, R., Lattanzi, S., & Vassilvitskii, S. (2017). Fair clustering through fairlets. In *Proceedings of the 31st International Conference on Neural Information Processing Systems* (pp. 5036–5044).
Choi, Y., Farnadi, G., Babaki, B., & Van den Broeck, G. (2020). Learning fair naive bayes classifiers by discovering and eliminating discrimination patterns. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 34, pp. 10077–10084). doi: https://doi.org/10.1609/aaai.v34i06.6565
Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, *5*(2), 153–163. doi: https://doi.org/10.1089/big.2016.0047
Chzhen, E., Denis, C., Hebiri, M., Oneto, L., & Pontil, M. (2020). Fair regression via plug-in estimator and recalibration with statistical guarantees. *Advances in Neural Information Processing Systems*, *33*.
Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., & Huq, A. (2017). Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 797–806). doi: https://doi.org/10.1145/3097983.3098095
Cortez, P., & Silva, A. M. G. (2008). Using data mining to predict secondary school student performance. *Proceedings of 5th FUture BUsiness TEChnology Conference (FUBUTEC 2008)*, 5-12.
Cox, D. R. (1958). The regression analysis of binary sequences. *Journal of the Royal Statistical Society: Series B (Methodological)*, *20*(2), 215–232. doi: https://doi.org/10.1111/j.2517-6161.1958.tb00292.x
D'Amour, A., Srinivasan, H., Atwood, J., Baljekar, P., Sculley, D., & Halpern, Y. (2020). Fairness is not static: deeper understanding of long term fairness via simulation studies. In *Proceedings of the 2020 conference on fairness, accountability, and transparency* (pp. 525–534).
Daniel, K. (2017). *Thinking, fast and slow.* Farrar Straus Giroux.
Datta, A., Fredrikson, M., Ko, G., Mardziel, P., & Sen, S. (2017). Use privacy in data-driven systems: Theory and experiments with machine learnt programs. In *Proceedings of the 2017 ACM SIGSAC conference on Computer and Communications Security* (pp. 1193–1210). doi: https://doi.org/10.1145/3133956.3134097
Datta, A., Tschantz, M. C., & Datta, A. (2015). Automated experiments on ad privacy settings. *Proceedings on Privacy Enhancing Technologies*, *2015*(1), 92–112.
Deepak, P., & Abraham, S. S. (2020). Fair outlier detection. In *International Conference on Web Information Systems Engineering* (pp. 447–462). doi: https://doi.org/10.1007/978-3-030-62008-0_31
Dheeru, D., & Karra Taniskidou, E. (2017). *UCI machine learning repository.* Retrieved from http://archive.ics.uci.edu/ml
Diana, E., Gill, W., Kearns, M., Kenthapadi, K., & Roth, A. (2021). Minimax group fairness: Algorithms and experiments. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 66–76). doi: https://dl.acm.org/doi/10.1145/3461702.3462523
Ding, F., Hardt, M., Miller, J., & Schmidt, L. (2021). Retiring adult: New datasets for fair machine learning. In *Thirty-fifth conference on neural information processing systems.*

Du, M., Yang, F., Zou, N., & Hu, X. (2020). Fairness in deep learning: A computational perspective. *IEEE Intelligent Systems*. doi: https://doi.org/10.1109/MIS.2020.3000681

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012). Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference* (pp. 214–226). doi: https://doi.org/10.1145/2090236.2090255

Esmaeili, S., Brubach, B., Tsepenekas, L., & Dickerson, J. (2020). Probabilistic fair clustering. *Advances in Neural Information Processing Systems*, *33*.

Faliagka, E., Ramantas, K., Tsakalidis, A., & Tzimas, G. (2012). Application of machine learning algorithms to an online recruitment system. In *Proceeding of the International Conference on Internet and Web Applications and Services*.

Feldman, M. (2015). Computational fairness: Preventing machine-learned discrimination.(2015). *URL https://scholarship.tricolib.brynmawr.edu/handle/10066/17628*.

Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., & Venkatasubramanian, S. (2015). Certifying and removing disparate impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 259–268). doi: https://doi.org/10.1145/2783258.2783311

Fish, B., Kun, J., & Lelkes, A. D. (2015). Fair boosting: a case study. In *Workshop on Fairness, Accountability, and Transparency in Machine Learning*.

Fish, B., Kun, J., & Lelkes, Á. D. (2016). A confidence-based approach for balancing fairness and accuracy. In *Proceedings of the 2016 SIAM International Conference on Data Mining* (pp. 144–152). doi: https://doi.org/10.1137/1.9781611974348.17

Foster, I., Ghani, R., Jarmin, R. S., Kreuter, F., & Lane, J. (2016). *Big data and social science: A practical guide to methods and tools*. crc Press.

Fox, J., & Carvalho, M. S. (2012). The rcmdrplugin. survival package: Extending the r commander interface to survival analysis. *Journal of Statistical Software*, *49*(1), 1–32. doi: https://doi.org/10.18637/jss.v049.i07

Friedler, S. A., Scheidegger, C., Venkatasubramanian, S., Choudhary, S., Hamilton, E. P., & Roth, D. (2019). A comparative study of fairness-enhancing interventions in machine learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (pp. 329–338). doi: https://doi.org/10.1145/3287560.3287589

Galhotra, S., Saisubramanian, S., & Zilberstein, S. (2021). Learning to generate fair clusters from demonstrations. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society (AIES '21)*. doi: https://doi.org/10.1145/3461702.3462558

Gardner, J., Brooks, C., & Baker, R. (2019). Evaluating the fairness of predictive student models through slicing analysis. In *Proceedings of the LAK19 conference* (pp. 225–234). doi: https://doi.org/10.1145/3303772.3303791

Grari, V., Ruf, B., Lamprier, S., & Detyniecki, M. (2019). Fair adversarial gradient tree boosting. In *2019 IEEE International Conference on Data Mining (ICDM)* (pp. 1060–1065). doi: https://doi.org/10.1109/ICDM.2019.00124

Grgić-Hlača, N., Zafar, M. B., Gummadi, K. P., & Weller, A. (2018). Beyond distributive fairness in algorithmic decision making: Feature selection for procedurally fair learning. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 32).

Haeri, M. A., & Zweig, K. A. (2020). The crucial role of sensitive attributes in fair classification. In *2020 IEEE Symposium Series on Computational Intelligence (SSCI)* (pp. 2993–3002). doi: https://doi.org/10.1109/SSCI47803.2020.9308585

Hajian, S., & Domingo-Ferrer, J. (2013). A methodology for direct and indirect discrimination prevention in data mining. *IEEE Transactions on Knowledge and Data Engineering*, *25*(7), 1445–1459. doi: https://doi.org/10.1109/TKDE.2012.72

Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. In *Proceedings of the 30th International Conference on Neural Information Processing Systems* (pp. 3323–3331).

Hébert-Johnson, U., Kim, M., Reingold, O., & Rothblum, G. (2018). Multicalibration: Calibration for the (computationally-identifiable) masses. In *International Conference on Machine Learning* (pp. 1939–1948).

Heidari, H., Ferrari, C., Gummadi, K. P., & Krause, A. (2018). Fairness behind a veil of ignorance: a welfare analysis for automated decision making. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems* (pp. 1273–1283).

Heidari, H., & Krause, A. (2018). Preventing disparate treatment in sequential decision making. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence* (pp. 2248–2254).

Holmes, D. E., & Jain, L. C. (2008). Introduction to Bayesian networks. In *Innovations in Bayesian Networks* (pp. 1–5). doi: https://doi.org/10.1007/978-3-540-85066-3_1

Hu, T., Iosifidis, V., Liao, W., Zhang, H., Yang, M. Y., Ntoutsi, E., & Rosenhahn, B. (2020). FairNN-conjoint learning of fair representations for fair decisions. In *International Conference on Discovery Science* (pp. 581–595). doi: https://doi.org/10.1007/978-3-030-61527-7_38

Huang, L., Jiang, S. H.-C., & Vishnoi, N. K. (2019). Coresets for clustering with fairness constraints. In *Proceedings of the 33rd international conference on neural information processing systems* (pp. 7589–7600).

Husmeier, D., Dybowski, R., & Roberts, S. (2006). *Probabilistic modeling in bioinformatics and medical informatics*. Springer Science & Business Media.

Ignatiev, A., Cooper, M. C., Siala, M., Hebrard, E., & Marques-Silva, J. (2020). Towards formal fairness in machine learning. In *International Conference on Principles and Practice of Constraint Programming* (pp. 846–867). doi: https://doi.org/10.1007/978-3-030-58475-7_49

Iosifidis, V., & Ntoutsi, E. (2018). Dealing with bias via data augmentation in supervised learning scenarios. *Jo Bates Paul D. Clough Robert Jäschke*, *24*.

Iosifidis, V., & Ntoutsi, E. (2019). AdaFair: Cumulative fairness adaptive boosting. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management* (pp. 781–790). doi: https://doi.org/10.1145/3357384.3357974

Iosifidis, V., & Ntoutsi, E. (2020). FABBOO - online fairness-aware learning under class imbalance. In *International Conference on Discovery Science* (pp. 159–174). doi: https://doi.org/10.1007/978-3-030-61527-7_11

Kamiran, F., & Calders, T. (2009). Classifying without discriminating. In *2009 2nd international conference on computer, control and communication* (p. 1-6). doi: https://doi.org/10.1109/IC4.2009.4909197

Kamiran, F., & Calders, T. (2012). Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems*, *33*(1), 1–33. doi: https://doi.org/10.1007/s10115-011-0463-8

Kamiran, F., Calders, T., & Pechenizkiy, M. (2010). Discrimination aware decision tree learning. In *2010 IEEE 10th International Conference on Data Mining (ICDM)* (pp. 869–874). doi: https://doi.org/10.1109/ICDM.2010.50

Kamiran, F., Žliobaitė, I., & Calders, T. (2013). Quantifying explainable discrimination and removing illegal discrimination in automated decision making. *Knowledge and Information Systems*, *35*(3), 613–644. doi: https://doi.org/10.1007/s10115-012-0584-8

Kamishima, T., Akaho, S., Asoh, H., & Sakuma, J. (2012). Fairness-aware classifier with

prejudice remover regularizer. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases* (pp. 35–50). doi: https://doi.org/10.1007/978-3-642-33486-3_3

Kearns, M., Neel, S., Roth, A., & Wu, Z. S. (2018). Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In *International Conference on Machine Learning* (pp. 2564–2572).

Kearns, M., Neel, S., Roth, A., & Wu, Z. S. (2019). An empirical study of rich subgroup fairness for machine learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (pp. 100–109). doi: https://doi.org/10.1145/3287560.3287592

Knaus, W. A., Harrell, F. E., Lynn, J., Goldman, L., Phillips, R. S., Connors, A. F., ... others (1995). The support prognostic model: Objective estimates of survival for seriously ill hospitalized adults. *Annals of internal medicine*, *122*(3), 191–203. doi: https://doi.org/10.7326/0003-4819-122-3-199502010-00007

Kohavi, R. (1996). Scaling up the accuracy of Naive-Bayes classifiers: a decision-tree hybrid. In *Proceedings of the 2nd International Conference on Knowledge Discovery and Data mining, Portland, 1996* (pp. 202–207).

Krasanakis, E., Spyromitros-Xioufis, E., Papadopoulos, S., & Kompatsiaris, Y. (2018). Adaptive sensitive reweighting to mitigate bias in fairness-aware classification. In *Proceedings of the 2018 World Wide Web conference on World Wide Web* (pp. 853–862).

Krop, P. S. (1981). Age discrimination and the disparate impact doctrine. *Stan. L. Rev.*, *34*, 837.

Kusner, M., Loftus, J., Russell, C., & Silva, R. (2017). Counterfactual fairness. In *Proceedings of the 31st International Conference on Neural Information Processing Systems* (pp. 4069–4079).

Kuzilek, J., Hlosta, M., & Zdrahal, Z. (2017). Open university learning analytics dataset. *Scientific data*, *4*, 170171. doi: https://doi.org/10.1038/sdata.2017.171

Kvamme, H., Borgan, Ø., & Scheel, I. (2019). Time-to-event prediction with neural networks and cox regression. *Journal of Machine Learning Research*, *20*(129), 1–30.

Lahoti, P., Beutel, A., Chen, J., Lee, K., Prost, F., Thain, N., ... Chi, E. (2020). Fairness without demographics through adversarially reweighted learning. In *34th Conference on Neural Information Processing Systems*.

Lahoti, P., Gummadi, K. P., & Weikum, G. (2019). Operationalizing individual fairness with pairwise fair representations. *Proceedings of the VLDB Endowment*, *13*(4), 506–518. doi: https://doi.org/10.14778/3372716.3372723

L. Cardoso, R., Meira Jr, W., Almeida, V., & J. Zaki, M. (2019). A framework for benchmarking discrimination-aware models in machine learning. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 437–444). doi: https://doi.org/10.1145/3306618.3314262

Le Quy, T., Roy, A., Friege, G., & Ntoutsi, E. (2021). Fair-capacitated clustering. In *Proceedings of the 14th International Conference on Educational Data Mining (EDM21).* (pp. 407–414).

Liu, L. T., Dean, S., Rolf, E., Simchowitz, M., & Hardt, M. (2018). Delayed impact of fair machine learning. In *International Conference on Machine Learning* (pp. 3150–3158).

Loh, W.-Y., Cao, L., & Zhou, P. (2019). Subgroup identification for precision medicine: A comparative review of 13 methods. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *9*(5), e1326.

Luong, B. T., Ruggieri, S., & Turini, F. (2011). k-nn as an implementation of situation testing for discrimination discovery and prevention. In *Proceedings of the 17th ACM SIGKDD*

International Conference on Knowledge discovery and Data mining (pp. 502–510). doi: https://doi.org/10.1145/2020408.2020488

Madras, D., Creager, E., Pitassi, T., & Zemel, R. (2019). Fairness through causal awareness: Learning causal latent-variable models for biased data. In *Proceedings of the conference on fairness, accountability, and transparency* (pp. 349–358). doi: https://doi.org/10.1145/3287560.3287564

Mahabadi, S., & Vakilian, A. (2020). Individual fairness for k-clustering. In *International Conference on Machine Learning* (pp. 6586–6596).

Mancuhan, K., & Clifton, C. (2014). Combating discrimination using bayesian networks. *Artificial Intelligence and Law*, *22*(2), 211–238. doi: https://doi.org/10.1007/s10506-014-9156-4

Martinez, N., Bertran, M., & Sapiro, G. (2020). Minimax pareto fairness: A multi objective perspective. In *International Conference on Machine Learning* (pp. 6755–6764).

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, *54*(6), 1–35. doi: https://doi.org/10.1145/3457607

Moore, J. S. (1998). An expert system approach to graduate school admission decisions and academic performance prediction. *Omega*, *26*(5), 659–670. doi: https://doi.org/10.1016/S0305-0483(98)00008-5

Moro, S., Cortez, P., & Rita, P. (2014). A data-driven approach to predict the success of bank telemarketing. *Decision Support Systems*, *62*, 22–31. doi: https://doi.org/10.1016/j.dss.2014.03.001

Mukerjee, A., Biswas, R., Deb, K., & Mathur, A. P. (2002). Multi–objective evolutionary algorithms for the risk–return trade–off in bank loan management. *International Transactions in Operational research*, *9*(5), 583–597. doi: https://doi.org/10.1111/1475-3995.00375

Narasimhan, H., Cotter, A., Gupta, M., & Wang, S. (2020). Pairwise fairness for ranking and regression. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 34, pp. 5248–5255). doi: https://doi.org/10.1609/aaai.v34i04.5970

Naumann, P., & Ntoutsi, E. (2021). Consequence-aware sequential counterfactual generation. In N. Oliver, F. Pérez-Cruz, S. Kramer, J. Read, & J. A. Lozano (Eds.), *European Conference on Machine Learning and Data Mining - ECML-PKDD 2021* (pp. 682–698). Cham: Springer International Publishing. doi: https://doi.org/10.1007/978-3-030-86520-7_42

Ntoutsi, E., Fafalios, P., Gadiraju, U., Iosifidis, V., Nejdl, W., Vidal, M.-E., ... others (2020). Bias in data-driven artificial intelligence systems—an introductory survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *10*(3), e1356. doi: https://doi.org/10.1002/widm.1356

Ntoutsi, E., Schubert, E., Zimek, A., & Zimmermann, A. (2019). Evaluation and experimental design in data mining and machine learning: Motivation and summary of edml 2019. In *CEUR Workshop Proceedings 2436 (2019)* (Vol. 2436, pp. 4–4).

Oneto, L., Doninini, M., Elders, A., & Pontil, M. (2019). Taking advantage of multitask learning for fair classification. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 227–237). doi: https://doi.org/10.1145/3306618.3314255

Paullada, A., Raji, I. D., Bender, E. M., Denton, E., & Hanna, A. (2021). Data and its (dis) contents: A survey of dataset development and use in machine learning research. *Patterns*, *2*(11), 100336.

Pedreschi, D., Ruggieri, S., & Turini, F. (2009). Measuring discrimination in socially-sensitive decision records. In *Proceedings of the 2009 SIAM International Conference on Data Mining* (pp. 581–592). doi: https://doi.org/10.1137/1.9781611972795.50

Pedreshi, D., Ruggieri, S., & Turini, F. (2008). Discrimination-aware data mining. In *Proceedings*

of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 560–568). doi: https://doi.org/10.1145/1401890.1401959

Pitoura, E., Stefanidis, K., & Koutrika, G. (2021). Fairness in rankings and recommendations: An overview. *The VLDB Journal*. doi: https://doi.org/10.1007/s00778-021-00697-y

Quadrianto, N., & Sharmanska, V. (2017). Recycling privileged learning and distribution matching for fairness.

Reddy, C., Sharma, D., Mehri, S., Romero, A., Shabanian, S., & Honari, S. (2021). Benchmarking bias mitigation algorithms in representation learning through fairness metrics.

Riazy, S., & Simbeck, K. (2019). Predictive algorithms in learning analytics and their fairness. *DELFI 2019*. doi: https://dx.doi.org/10.18420/delfi2019_305

Riazy, S., Simbeck, K., & Schreck, V. (2020). Fairness in learning analytics: Student at-risk prediction in virtual learning environments. In *CSEDU (1)* (pp. 15–25). doi: https://doi.org/10.5220/0009324100150025

Ristanoski, G., Liu, W., & Bailey, J. (2013). Discrimination aware classification for imbalanced datasets. In *Proceedings of the 22nd ACM International Conference on Information & Knowledge Management* (pp. 1529–1532). doi: https://doi.org/10.1145/2505515.2507836

Ruggieri, S., Pedreschi, D., & Turini, F. (2010). Data mining for discrimination discovery. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, *4*(2), 9. doi: https://doi.org/10.1145/1754428.1754432

Ruoss, A., Balunovic, M., Fischer, M., & Vechev, M. T. (2020). Learning certified individually fair representations. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems (NeurIPS 2020)*.

Russell, C., Kusner, M. J., Loftus, J. R., & Silva, R. (2017). When worlds collide: integrating different counterfactual assumptions in fairness. *Advances in Neural Information Processing Systems 30. Pre-proceedings*, *30*.

Sarah Flood, R. R. S. R., Miriam King, & Warren, J. R. (2020). Integrated public use microdata series, current population survey: Version 8.0 [dataset]. *Minneapolis, MN: IPUMS*. doi: https://doi.org/10.18128/D030.V8.0

Schelter, S., He, Y., Khilnani, J., & Stoyanovich, J. (2020). Fairprep: Promoting data to a first-class citizen in studies on fairness-enhancing interventions. In *Edbt*. doi: https://doi.org/10.5441/002/edbt.2020.41

Sharifi-Malvajerdi, S., Kearns, M., & Roth, A. (2019). Average individual fairness: Algorithms, generalization and experiments. *Advances in Neural Information Processing Systems*, *32*, 8242–8251.

Simoiu, C., Corbett-Davies, S., & Goel, S. (2017). The problem of infra-marginality in outcome tests for discrimination. *The Annals of Applied Statistics*, *11*(3), 1193–1216.

Simonite, T. (2015). Probing the dark side of google's ad-targeting system. *MIT Technology Review*.

Slack, D., Friedler, S. A., & Givental, E. (2020). Fairness warnings and fair-maml: learning fairly with minimal data. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (pp. 200–209). doi: https://doi.org/10.1145/3351095.3372839

Strack, B., DeShazo, J. P., Gennings, C., Olmo, J. L., Ventura, S., Cios, K. J., & Clore, J. N. (2014). Impact of HbA1c measurement on hospital readmission rates: analysis of 70,000 clinical database patient records. *BioMed research international*, *2014*. doi: https://doi.org/10.1155/2014/781670

Supreme Court of the United States. (2009). Ricci v. destefano. In *557 u.s. 557, 174.*

Tu, R., Zhang, X., Liu, Y., Kjellström, H., Liu, M., Zhang, K., & Zhang, C. (2020). How do fair decisions fare in long-term qualification? In *Thirty-fourth Conference on Neural Information Processing Systems.*

Valdivia, A., Sánchez-Monedero, J., & Casillas, J. (2021). How fair can we go in machine learning? Assessing the boundaries of accuracy and fairness. *International Journal of Intelligent Systems*, *36*(4), 1619–1643. doi: https://doi.org/10.1002/int.22354

van Berkel, N., Goncalves, J., Russo, D., Hosio, S., & Skov, M. B. (2021). Effect of information presentation on fairness perceptions of machine learning predictors. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (pp. 1–13). doi: https://doi.org/10.1145/3411764.3445365

Van der Laan, P. (2000). The 2001 census in the netherlands. In *Conference the Census of Population.*

Verma, S., & Rubin, J. (2018). Fairness definitions explained. In *2018 IEEE/ACM International Workshop on Software Fairness (FairWare)* (pp. 1–7). doi: https://doi.org/10.23919/FAIRWARE.2018.8452913

Wang, X., Zhang, W., Jadhav, A., & Weiss, J. (2021). Harmonic-mean cox models: A ruler for equal attention to risk. In *Survival Prediction-Algorithms, Challenges and Applications* (pp. 171–183).

Warner, R., & Sloan, R. H. (2021). Making artificial intelligence transparent: Fairness and the problem of proxy variables. *Criminal Justice Ethics*, *40*(1), 23–39. doi: https://doi.org/10.1080/0731129X.2021.1893932

Wen, M., Bastani, O., & Topcu, U. (2021). Algorithms for fairness in sequential decision making. In *International Conference on Artificial Intelligence and Statistics* (pp. 1144–1152).

Wightman, L. F. (1998). LSAC national longitudinal bar passage study. LSAC research report series.

Xivuri, K., & Twinomurinzi, H. (2021). A systematic review of fairness in artificial intelligence algorithms. In *Conference on e-Business, e-Services and e-Society* (pp. 271–284). doi: https://doi.org/10.1007/978-3-030-85447-8_24

Xu, R., Cui, P., Kuang, K., Li, B., Zhou, L., Shen, Z., & Cui, W. (2020). Algorithmic decision making with conditional fairness. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (pp. 2125–2135). doi: https://doi.org/10.1145/3394486.3403263

Yang, F., Cisse, M., & Koyejo, S. (2020). Fairness with overlapping groups: a probabilistic perspective. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, & H. Lin (Eds.), *Advances in Neural Information Processing Systems* (Vol. 33, pp. 4067–4078). Curran Associates, Inc.

Yeh, I.-C., & Lien, C.-h. (2009). The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. *Expert Systems with Applications*, *36*(2), 2473–2480. doi: https://doi.org/10.1016/j.eswa.2007.12.020

Yeom, S., Datta, A., & Fredrikson, M. (2018). Hunting for discriminatory proxies in linear regression models. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems* (pp. 4573–4583).

Zafar, M. B., Valera, I., Gomez Rodriguez, M., & Gummadi, K. P. (2017). Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th International Conference on World Wide Web* (pp. 1171–1180). doi: https://doi.org/10.1145/3038912.3052660

Zafar, M. B., Valera, I., Gomez-Rodriguez, M., & Gummadi, K. P. (2019). Fairness constraints: A flexible approach for fair classification. *J. Mach. Learn. Res.*, *20*(75), 1–42.

Zafar, M. B., Valera, I., Rogriguez, M. G., & Gummadi, K. P. (2017, 20–22 Apr). Fairness Constraints: Mechanisms for Fair Classification. In A. Singh & J. Zhu (Eds.), *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics* (Vol. 54, pp. 962–970).

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

Zhang, B. H., Lemoine, B., & Mitchell, M. (2018). Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 335–340). doi: https://doi.org/10.1145/3278721.3278779

Zhang, W., & Bifet, A. (2020). Feat: A fairness-enhancing and concept-adapting decision tree classifier. In *International conference on discovery science* (pp. 175–189). doi: https://doi.org/10.1007/978-3-030-61527-7_12

Zhang, W., Bifet, A., Zhang, X., Weiss, J. C., & Nejdl, W. (2021). Farf: A fair and adaptive random forests classifier. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining* (pp. 245–256). doi: https://doi.org/10.1007/978-3-030-75765-6_20

Zhang, W., & Ntoutsi, E. (2019). FAHT: an adaptive fairness-aware decision tree classifier. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence (IJCAI-19).*

Zhang, W., Tang, J., & Wang, N. (2016). Using the machine learning approach to predict patient survival from high-dimensional survival data. In *IEEE International Conference on Bioinformatics and Biomedicine (BIBM).* doi: https://doi.org/10.1109/BIBM.2016.7822695

Zhang, W., Tang, X., & Wang, J. (2019). On fairness-aware learning for non-discriminative decision-making. In *2019 International Conference on Data Mining Workshops (ICDMW)* (pp. 1072–1079). doi: https://doi.org/10.1109/ICDMW.2019.00157

Zhang, W., & Weiss, J. (2021). Fair decision-making under uncertainty. In *2021 IEEE International Conference on Data Mining (ICDM).*

Zhang, W., & Weiss, J. (2022). Longitudinal fairness with censorship. In *Proceedings of the AAAI Conference on Artificial Intelligence.*

Zhao, J., Wang, T., Yatskar, M., Ordonez, V., & Chang, K.-W. (2018). Gender bias in coreference resolution: Evaluation and debiasing methods. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (Vol. 2). doi: https://doi.org/10.18653/v1/N18-2003

Ziko, I. M., Yuan, J., Granger, E., & Ayed, I. B. (2021). Variational fair clustering. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 35, pp. 11202–11209).

Žliobaitė, I. (2015). On the relation between accuracy and fairness in binary classification. In *The 2nd Workshop on Fairness, Accountability, and Transparency in Machine Learning (FAT/ML 2015) workshop at ICML'15.*

Žliobaitė, I. (2017). Measuring discrimination in algorithmic decision making. *Data Mining and Knowledge Discovery*, *31*(4), 1060–1089. doi: https://doi.org/10.1007/s10618-017-0506-1

Žliobaitė, I., Kamiran, F., & Calders, T. (2011). Handling conditional discrimination. In *2011 IEEE 11th International Conference on Data Mining (ICDM)* (pp. 992–1001). doi: https://doi.org/10.1109/ICDM.2011.72

A Citations

1. **Adult dataset**

   (Krasanakis et al., 2018; Kamiran & Calders, 2012; Kamiran et al., 2013; Calders et al., 2009; Žliobaitė, Kamiran, & Calders, 2011; Calders & Verwer, 2010; Luong et al., 2011; Kamiran, Calders, & Pechenizkiy, 2010; Iosifidis & Ntoutsi, 2018; Calmon, Wei, Vinzamuri, Ramamurthy, & Varshney, 2017; Feldman, Friedler, Moeller, Scheidegger, & Venkatasubramanian, 2015; Hajian & Domingo-Ferrer, 2013; Zafar, Valera, Rogriguez, & Gummadi, 2017; Žliobaitė, 2015; Calders & Kamiran, 2010; Feldman, 2015; Fish, Kun, & Lelkes, 2015; Fish et al., 2016; Friedler et al., 2019; Ristanoski et al., 2013; Chakraborty, Peng, & Menzies, 2020; Quadrianto & Sharmanska, 2017; Xu et al., 2020; Zafar, Valera, Gomez-Rodriguez, & Gummadi, 2019; Choi et al., 2020; Oneto, Doninini, Elders, & Pontil, 2019; Grari, Ruf, Lamprier, & Detyniecki, 2019; L. Cardoso, Meira Jr, Almeida, & J. Zaki, 2019; Agarwal, Beygelzimer, Dudík, Langford, & Wallach, 2018; Backurs et al., 2019; Hu et al., 2020; Chierichetti et al., 2017; Ziko et al., 2021; Haeri & Zweig, 2020; Berk et al., 2017; Esmaeili, Brubach, Tsepenekas, & Dickerson, 2020; Deepak & Abraham, 2020; Mahabadi & Vakilian, 2020; Huang, Jiang, & Vishnoi, 2019; Kearns et al., 2019; Bechavod & Ligett, 2017; Ruoss et al., 2020; B. H. Zhang et al., 2018; Iosifidis & Ntoutsi, 2019; Du, Yang, Zou, & Hu, 2020; W. Zhang & Ntoutsi, 2019; Galhotra, Saisubramanian, & Zilberstein, 2021; Abbasi, Bhaskara, & Venkatasubramanian, 2021; Diana et al., 2021; Galhotra et al., 2021; Chakraborty, Majumder, et al., 2020).

2. **KDD Census-Income dataset**

   (Iosifidis & Ntoutsi, 2019; Ristanoski et al., 2013; Iosifidis & Ntoutsi, 2020; W. Zhang & Ntoutsi, 2019).

3. **German credit dataset**

   (Calders et al., 2009; Luong et al., 2011; Ruggieri, Pedreschi, & Turini, 2010; Pedreshi, Ruggieri, & Turini, 2008; Pedreschi, Ruggieri, & Turini, 2009; Iosifidis & Ntoutsi, 2018; Feldman et al., 2015; Hajian & Domingo-Ferrer, 2013; Kamiran & Calders, 2009; Feldman, 2015; Fish et al., 2016; Zemel, Wu, Swersky, Pitassi, & Dwork, 2013; Friedler et al., 2019; Mancuhan & Clifton, 2014; Ristanoski et al., 2013; Choi et al., 2020; Ruoss et al., 2020; Ahn & Lin, 2019; Chakraborty, Majumder, et al., 2020).

4. **Dutch census dataset**

   (Kamiran et al., 2010; Kamiran & Calders, 2012; Kamiran et al., 2013; Žliobaitė et al., 2011; Kamiran et al., 2010; Xu et al., 2020; L. Cardoso et al., 2019; Agarwal et al., 2018).

5. **Bank marketing dataset**

   (Grari et al., 2019; Zafar et al., 2019; Krasanakis et al., 2018; Zafar, Valera, Rogriguez, & Gummadi, 2017; Fish et al., 2016; Backurs et al., 2019; Hu et al., 2020; Chierichetti et al., 2017; Ziko et al., 2021; Haeri & Zweig, 2020; Bera et al., 2019; Mahabadi & Vakilian, 2020; Huang et al., 2019; Galhotra et al., 2021; Abbasi et al., 2021; Galhotra et al., 2021).

6. **Credit card clients dataset**

   (Yeh & Lien, 2009; Berk et al., 2017; Esmaeili et al., 2020; Bera et al., 2019; Deepak & Abraham, 2020; Bechavod & Ligett, 2017; Chakraborty, Majumder, et al., 2020).

7. **COMPAS dataset**

   (Krasanakis et al., 2018; Calmon et al., 2017; Chouldechova, 2017; Zafar, Valera, Gomez Rodriguez, & Gummadi, 2017; Corbett-Davies, Pierson, Feller, Goel, & Huq, 2017; Friedler et al., 2019; Tu et al., 2020; Quadrianto & Sharmanska, 2017; Xu et al., 2020; Zafar et al., 2019; Slack, Friedler, & Givental, 2020; Choi et al., 2020; Grgić-Hlača, Zafar, Gummadi, & Weller, 2018; Oneto et al., 2019; L. Cardoso et al., 2019; Agarwal et al., 2018; Haeri & Zweig, 2020; Lahoti, Gummadi, & Weikum, 2019; Heidari et al., 2018; Russell et al., 2017; Berk et al., 2017; Ruoss et al., 2020; Du et al., 2020; Diana et al., 2021; van Berkel et al., 2021; Chakraborty, Majumder, et al., 2020).

8. **Communites & Crime dataset**

   (Kamiran & Calders, 2012; Kamiran et al., 2013, 2010; Lahoti et al., 2019; Kearns et al., 2018; Narasimhan, Cotter, Gupta, & Wang, 2020; Slack et al., 2020; Sharifi-Malvajerdi, Kearns, & Roth, 2019; Heidari et al., 2018; Calders et al., 2013; Chzhen et al., 2020; Kearns et al., 2019; Berk et al., 2017; Ruoss et al., 2020; Galhotra et al., 2021; Diana et al., 2021; Galhotra et al., 2021).

9. **Diabetes dataset**

   (Backurs et al., 2019; Chierichetti et al., 2017; Mahabadi & Vakilian, 2020; Huang et al., 2019).

10. **Ricci dataset**

    (Feldman et al., 2015; Feldman, 2015; Friedler et al., 2019; Ignatiev, Cooper, Siala, Hebrard, & Marques-Silva, 2020; Schelter, He, Khilnani, & Stoyanovich, 2020; Valdivia, Sánchez-Monedero, & Casillas, 2021).

11. **Student performance dataset**

    (Deepak & Abraham, 2020; Chzhen et al., 2020; Kearns et al., 2019; Le Quy, Roy, Friege, & Ntoutsi, 2021).

12. **OULAD dataset**

    (Riazy & Simbeck, 2019; Le Quy et al., 2021; Riazy, Simbeck, & Schreck, 2020).

13. **Law School dataset**

    (Chzhen et al., 2020; Kearns et al., 2019; Kusner et al., 2017; Russell et al., 2017; Lahoti et al., 2020; Bechavod & Ligett, 2017; Berk et al., 2017; Yang et al., 2020; Ruoss et al., 2020).

B Datasets' characteristics

Table 16: KDD Census-Income: attributes characteristics (continued)

Attributes	Type	Values	#Missing values	Description
enroll-in-edu-inst-last-wk	Categorical	3	0	An individual enrolled in an educational institute last week?
major-industry	Categorical	24	0	The major industry code
major-occupation	Categorical	15	0	The major occupation code
hispanic-origin	Categorical	9	1,279	The Hispanic origin
member-union	Categorical	3	0	Member of a labor union
reason-unemployment	Categorical	6	0	The reason for unemployment
region-previous	Categorical	6	0	The region of previous residence
state-previous	Categorical	50	1038	The state of previous residence
migration-code-change-in-msa	Categorical	10	149,642	Migration code-change in MSA
migration-code-change-in-reg	Categorical	9	149,642	Migration code-change in region
migration-code-move-within-reg	Categorical	10	149,642	Migration code-move within region
live-hour-1-year-ago	Categorical	3	0	Live in this house 1 year ago
migration-prev-res-in-sunbelt	Categorical	4	149,642	Migration from the previous residence in the sunbelt
country-father	Categorical	42	10,142	The country of birth of the father
country-mother	Categorical	42	9,191	The country of birth of the mother
country-birth	Categorical	42	5,157	The country of birth
fill-questionnaire	Categorical	3	0	Fill the questionnaire for veteran's admin

Table 17: COMPAS recid: attributes characteristics (continued)

Attributes	Type	Values	#Missing values	Description
name	Categorical	7,158	0	First and last name of the defendant
first	Categorical	2,800	0	First name
last	Categorical	3,950	0	Last name
compas_screening_date	Categorical	690	0	The date on which the decile score was given
dob	Categorical	5,452	0	Date of birth
decile_score	Numerical	[1 - 10]	0	The COMPAS Risk of Recidivism score
days_b_screening_arrest	Numerical	[-414 - 1,057]	307	The number of days between COMPAS screening and arrest
c_jail_in	Categorical	6,907	307	The jail entry date for original crime
c_jail_out	Categorical	6,880	307	The jail exit date for original crime
c_case_number	Categorical	7,192	22	The case number for original crime
c_offense_date	Categorical	927	1,159	The offense date of original crime
c_arrest_date	Categorical	580	6,077	The arrest date for original crime
c_days_from_compas	Numerical	[0 - 9,485]	22	Between the COMPAS screening and the original crime offense date
c_charge_desc	Categorical	437	29	Description of charge for original crime
is_recid	Binary	{0, 1}	0	The binary indicator of recidivation
r_case_number	Categorical	3,471	3,743	The case number of follow-up crime
r_charge_degree	Categorical	10	3,743	Charge degree of follow-up crime
r_days_from_arrest	Numerical	[-1 - 993]	4,898	Between the follow-up crime and the arrest date (days)
r_offense_date	Categorical	1,075	3,743	The date of follow-up crime
r_charge_desc	Categorical	340	3,801	Description of charge for follow-up crime
r_jail_in	Categorical	972	4,898	The jail entry date for follow-up crime
r_jail_out	Categorical	938	4,898	The jail exit date for follow-up crime
violent_recid		NULL	7,214	Values are all NA. This column is ignored
is_violent_recid	Binary	{0, 1}	0	The binary indicator of violent follow-up crime
vr_case_number	Categorical	819	6,395	The case number for violent follow-up crime
vr_charge_degree	Categorical	9	6,395	Charge degree for violent follow-up crime
vr_offense_date	Categorical	570	6,395	The date of offense for violent follow-up crime
vr_charge_desc	Categorical	83	6,395	Description of charge for violent follow-up crime
type_of_assessment	Categorical	1	0	The type of COMPAS score given for decile score
decile_score.1	Numerical	[1 - 10]	0	Repeat column of decile score
screening_date	Categorical	690	0	Repeat column of compas_screening_date
v_type_of_assessment	Categorical	1	0	The type of COMPAS score given for v_decile_score
v_decile_score	Numerical	[1 - 10]	0	The COMPAS Risk of Violence score from 1 to 10
v_screening_date	Categorical	690	0	The date on which v_decile_score was given
in_custody	Categorical	1,156	236	The date on which individual was brought into custody
out_custody	Categorical	1,169	236	The date on which individual was released from custody
priors_count.1	Numerical	0 - 38	0	Repeat column of priors_count
start	Numerical	[0 - 937]	0	No information
end	Numerical	[0 - 1,186]	0	No information
event	Binary	{0, 1}	0	No information

Table 18: Communities and Crime: attributes characteristics (continued)

Attributes	Type	Values	#Missing values	Description
state	Categorical	46	0	The US state (by number)
county	Categorical	109	1174	The numeric code for county
community	Categorical	800	1,177	The numeric code for community
communityname	Categorical	1,828	0	The community name
fold	Numerical	[1 - 10]	0	The fold number for non-random 10 fold cross validation
population	Numerical	[0.0 - 1.0]	0	The population for community
householdsize	Numerical	[0.0 - 1.0]	0	The mean people per household
racePctWhite	Numerical	[0.0 - 1.0]	0	The percentage of population that is Caucasian
racePctAsian	Numerical	[0.0 - 1.0]	0	The percentage of population that is of Asian heritage
racePctHisp	Numerical	[0.0 - 1.0]	0	The percentage of population that is of Hispanic heritage
agePct12t21	Numerical	[0.0 - 1.0]	0	The percentage of population that is 12-21 in age
agePct12t29	Numerical	[0.0 - 1.0]	0	The percentage of population that is 12-29 in age
agePct16t24	Numerical	[0.0 - 1.0]	0	The percentage of population that is 16-24 in age
agePct65up	Numerical	[0.0 - 1.0]	0	The percentage of population that is 65 and over in age
numbUrban	Numerical	[0.0 - 1.0]	0	The number of people living in areas classified as urban
pctUrban	Numerical	[0.0 - 1.0]	0	The percentage of people living in areas classified as urban
medIncome	Numerical	[0.0 - 1.0]	0	The median household income
pctWWage	Numerical	[0.0 - 1.0]	0	The percentage of households with wage or salary income in 1989
pctWFarmSelf	Numerical	[0.0 - 1.0]	0	The percentage of households with farm or self employment income in 1989
pctWSocSec	Numerical	[0.0 - 1.0]	0	The percentage of households with social security income in 1989
pctWRetire	Numerical	[0.0 - 1.0]	0	The percentage of households with retirement income in 1989
medFamInc	Numerical	[0.0 - 1.0]	0	The median family income
perCapInc	Numerical	[0.0 - 1.0]	0	Per capita income (national income divided by population size)
whitePerCap	Numerical	[0.0 - 1.0]	0	Per capita income for Caucasians
blackPerCap	Numerical	[0.0 - 1.0]	0	Per capita income for African Americans
indianPerCap	Numerical	[0.0 - 1.0]	0	Per capita income for native Americans
AsianPerCap	Numerical	[0.0 - 1.0]	0	Per capita income for people with Asian heritage
OtherPerCap	Numerical	[0.0 - 1.0]	1	Per capita income for people with 'other' heritage
HispPerCap	Numerical	[0.0 - 1.0]	0	Per capita income for people with Hispanic heritage
PctLess9thGrade	Numerical	[0.0 - 1.0]	0	The percentage of people 25 and over with less than a 9th grade education
PctNotHSGrad	Numerical	[0.0 - 1.0]	0	The percentage of people 25 and over that are not high school graduates
PctBSorMore	Numerical	[0.0 - 1.0]	0	The percentage of people 25 and over with a bachelors degree or higher education
PctUnemployed	Numerical	[0.0 - 1.0]	0	The percentage of people 16 and over, in the labor force, and unemployed
PctEmploy	Numerical	[0.0 - 1.0]	0	The percentage of people 16 and over who are employed
PctEmplManu	Numerical	[0.0 - 1.0]	0	The percentage of people 16 and over who are employed in manufacturing
PctEmplProfServ	Numerical	[0.0 - 1.0]	0	The percentage of people 16 and over who are employed in professional services
PctOccupManu	Numerical	[0.0 - 1.0]	0	The percentage of people 16 and over who are employed in manufacturing
PctOccupMgmtProf	Numerical	[0.0 - 1.0]	0	The percentage of people 16 and over who are employed in management
MalePctNevMarr	Numerical	[0.0 - 1.0]	0	The percentage of males who have never married
PersPerFam	Numerical	[0.0 - 1.0]	0	The mean number of people per family
PctWorkMomYoungKids	Numerical	[0.0 - 1.0]	0	The percentage of moms of kids 6 and under in labor force
PctWorkMom	Numerical	[0.0 - 1.0]	0	The percentage of moms of kids under 18 in labor force
NumImmig	Numerical	[0.0 - 1.0]	0	The total number of people known to be foreign born
PctImmigRecent	Numerical	[0.0 - 1.0]	0	The percentage of immigrants who immigated within the last 3 years
PctImmigRec5	Numerical	[0.0 - 1.0]	0	The percentage of immigrants who immigated within the last 5 years
PctImmigRec8	Numerical	[0.0 - 1.0]	0	The percentage of immigrants who immigated within the last 8 years
PctImmigRec10	Numerical	[0.0 - 1.0]	0	The percentage of immigrants who immigated within the last 10 years
PctRecentImmig	Numerical	[0.0 - 1.0]	0	The percentage of the population who have immigrated within the last 3 years
PctRecImmig5	Numerical	[0.0 - 1.0]	0	The percentage of the population who have immigrated within the last 5 years
PctRecImmig8	Numerical	[0.0 - 1.0]	0	The percentage of the population who have immigrated within the last 8 years
PctRecImmig10	Numerical	[0.0 - 1.0]	0	The percentage of the population who have immigrated within the last 10 years
PctSpeakEnglOnly	Numerical	[0.0 - 1.0]	0	The percentage of the population who speak only English
PctNotSpeakEnglWell	Numerical	[0.0 - 1.0]	0	The percentage of population who do not speak English well
PctLargHouseFam	Numerical	[0.0 - 1.0]	0	The percentage of family households that are large (6 or more)
PctLargHouseOccup	Numerical	[0.0 - 1.0]	0	The percentage of all occupied households that are large (6 or more people)
PersPerOccupHous	Numerical	[0.0 - 1.0]	0	The mean persons per household
PersPerOwnOccHous	Numerical	[0.0 - 1.0]	0	The mean persons per owner occupied household
PersPerRentOccHous	Numerical	[0.0 - 1.0]	0	The mean persons per rental household
PctPersDenseHous	Numerical	[0.0 - 1.0]	0	The percentage of persons in dense housing (more than 1 person per room)
PctHousLess3BR	Numerical	[0.0 - 1.0]	0	The percentage of housing units with less than 3 bedrooms
MedNumBR	Numerical	[0.0 - 1.0]	0	The median number of bedrooms
PctHousOccup	Numerical	[0.0 - 1.0]	0	The percentage of housing occupied
PctVacMore6Mos	Numerical	[0.0 - 1.0]	0	The percentage of vacant housing that has been vacant more than 6 months
MedYrHousBuilt	Numerical	[0.0 - 1.0]	0	The median year housing units built
PctHousNoPhone	Numerical	[0.0 - 1.0]	0	The percentage of occupied housing units without phone (in 1990)
PctWOFullPlumb	Numerical	[0.0 - 1.0]	0	The percentage of housing without complete plumbing facilities
OwnOccLowQuart	Numerical	[0.0 - 1.0]	0	Owner-occupied housing - lower quartile value
OwnOccMedVal	Numerical	[0.0 - 1.0]	0	Owner-occupied housing - median value
OwnOccHiQuart	Numerical	[0.0 - 1.0]	0	Owner-occupied housing - upper quartile value
RentLowQ	Numerical	[0.0 - 1.0]	0	Rental housing - lower quartile rent
RentMedian	Numerical	[0.0 - 1.0]	0	Rental housing - median rent

Table 19: Communities and Crime: attributes characteristics (continued)

Attributes	Type	Values	#Missing values	Description
RentHighQ	Numerical	[0.0 - 1.0]	0	Rental housing - upper quartile rent
MedRent	Numerical	[0.0 - 1.0]	0	The median gross rent
MedRentPctHousInc	Numerical	[0.0 - 1.0]	0	The median gross rent as a percentage of household income
MedOwnCostPctInc	Numerical	[0.0 - 1.0]	0	The median owners cost (with a mortgage) as a percentage of household income
MedOwnCostPctIncNoMtg	Numerical	[0.0 - 1.0]	0	The median owners cost (without a mortgage) as a percentage of household income
PctForeignBorn	Numerical	[0.0 - 1.0]	0	The percentage of people foreign born
PctBornSameState	Numerical	[0.0 - 1.0]	0	The percentage of people born in the same state as currently living
PctSameHouse85	Numerical	[0.0 - 1.0]	0	The percentage of people living in the same house as in 1985 (5 years before)
PctSameCity85	Numerical	[0.0 - 1.0]	0	The percentage of people living in the same city as in 1985 (5 years before)
PctSameState85	Numerical	[0.0 - 1.0]	0	The percentage of people living in the same state as in 1985 (5 years before)
LemasSwornFT	Numerical	[0.0 - 1.0]	1,675	The number of sworn full-time police officers
LemasSwFTPerPop	Numerical	[0.0 - 1.0]	1,675	The number of sworn full-time police officers in field operations
LemasSwFTFieldOps	Numerical	[0.0 - 1.0]	1,675	The sworn full-time police officers in field operations per 100,000 population
LemasSwFTFieldPerPop	Numerical	[0.0 - 1.0]	1,675	The number of sworn full time police officers in field operations
LemasTotalReq	Numerical	[0.0 - 1.0]	1,675	The total requests for police
LemasTotReqPerPop	Numerical	[0.0 - 1.0]	1,675	The total requests for police per 100,000 popuation
PolicReqPerOffic	Numerical	[0.0 - 1.0]	1,675	The total requests for police per police officer
PolicPerPop	Numerical	[0.0 - 1.0]	1,675	The number of police officers per 100,000 population
RacialMatchCommPol	Numerical	[0.0 - 1.0]	1,675	A measure of the racial match between the community and the police force
PctPolicWhite	Numerical	[0.0 - 1.0]	1,675	The percentage of police that are Caucasian
PctPolicBlack	Numerical	[0.0 - 1.0]	1,675	The percentage of police that are African American
PctPolicHisp	Numerical	[0.0 - 1.0]	1,675	The percentage of police that are Hispanic
PctPolicAsian	Numerical	[0.0 - 1.0]	1,675	The percentage of police that are Asian
PctPolicMinor	Numerical	[0.0 - 1.0]	1,675	The percentage of police that are minority of any kind
OfficAssgnDrugUnits	Numerical	[0.0 - 1.0]	1,675	The number of officers assigned to special drug units
NumKindsDrugsSeiz	Numerical	[0.0 - 1.0]	1,675	The number of different kinds of drugs seized
PolicAveOTWorked	Numerical	[0.0 - 1.0]	1,675	Police average overtime worked
LandArea	Numerical	[0.0 - 1.0]	0	Land area in square miles
PopDens	Numerical	[0.0 - 1.0]	0	The population density in persons per square mile
PctUsePubTrans	Numerical	[0.0 - 1.0]	0	The percentage of people using public transit for commuting
PolicCars	Numerical	[0.0 - 1.0]	1,675	The number of police cars
PolicOperBudg	Numerical	[0.0 - 1.0]	1,675	Police operating budget
LemasPctPolicOnPatr	Numerical	[0.0 - 1.0]	1,675	The percentage of sworn full-time police officers on patrol
LemasGangUnitDeploy	Numerical	[0.0 - 1.0]	1,675	Gang unit deployed
LemasPctOfficDrugUn	Numerical	[0.0 - 1.0]	0	The percentage of officers assigned to drug units
PolicBudgPerPop	Numerical	[0.0 - 1.0]	1,675	Police operating budget per population

Table 20: Diabetes: attributes characteristics (continued)

Attributes	Type	Values	#Missing values	Description
encounter_ID	Numerical	[12,522 - 443,867,222]	0	Encounter's unique identifier
patient_nbr	Numerical	[135 - 189,502,619]	0	Patient's unique identifier
weight	Categorical	10	98,569	Weight (pounds)
admission_type_id	Categorical	8	0	The admission type (emergency, urgent, etc.
discharge_disposition_id	Categorical	26	0	Discharge disposition (discharged to home, expired, etc.)
admission_source_id	Categorical	17	0	The admission source (physician referral, emergency room, etc.)
payer_code	Categorical	18	40,256	Payer code (Medicare, self-pay, etc.
medical_specialty	Categorical	73	49,949	The specialty of the admitting physician
num_lab_procedures	Numerical	[1 - 132]	0	The number of lab tests performed during the encounter
diag_1	Categorical	717	21	The primary diagnosis
diag_2	Categorical	749	358	Secondary diagnosis
diag_3	Categorical	790	1,423	Additional secondary diagnosis
number_diagnoses	Numerical	[1 - 16]	0	The number of diagnoses entered to the system
max_glu_serum	Categorical	4	0	The range of the results or if the test was not taken
repaglinide	Categorical	4	0	Whether the drug was prescribed or there was a change in the dosage
nateglinide	Categorical	4	0	Whether the drug was prescribed or there was a change in the dosage
glimepiride	Categorical	4	0	Whether the drug was prescribed or there was a change in the dosage
acetohexamide	Categorical	2	0	Whether the drug was prescribed or there was a change in the dosage
glyburide	Categorical	4	0	Whether the drug was prescribed or there was a change in the dosage
tolbutamide	Categorical	2	0	Whether the drug was prescribed or there was a change in the dosage
pioglitazone	Categorical	4	0	Whether the drug was prescribed or there was a change in the dosage
troglitazone	Categorical	2	0	Whether the drug was prescribed or there was a change in the dosage
tolazamide	Categorical	3	0	Whether the drug was prescribed or there was a change in the dosage
examide	Categorical	1	0	Whether the drug was prescribed or there was a change in the dosage
citoglipton	Categorical	1	0	Whether the drug was prescribed or there was a change in the dosage
insulin	Categorical	4	0	Whether the drug was prescribed or there was a change in the dosage
glyburide-metformin	Categorical	4	0	Whether the drug was prescribed or there was a change in the dosage
glipizide-metformin	Categorical	2	0	Whether the drug was prescribed or there was a change in the dosage
glimepiride-pioglitazone	Categorical	2	0	Whether the drug was prescribed or there was a change in the dosage
metformin-rosiglitazone	Categorical	1	0	Whether the drug was prescribed or there was a change in the dosage
metformin-pioglitazone	Categorical	1	0	Whether the drug was prescribed or there was a change in the dosage
change	Binary	{No, Ch}	0	Was there a change in diabetic medications?