

A survey on datasets for fairness-aware machine learning

Tai Le Quy^{1*}, Arjun Roy², Vasileios Iosifidis¹ and Eirini Ntoutsi²

¹L3S Research Center, Leibniz University Hannover, Germany.

²Institute of Computer Science, Free University Berlin, Germany.

*Corresponding author(s). E-mail(s): tai@l3s.de;

Contributing authors: arjun.roy@fu-berlin.de; iosifidis@l3s.de; eirini.ntoutsi@fu-berlin.de;

Abstract

As decision-making increasingly relies on machine learning and (big) data, the issue of fairness in data-driven AI systems is receiving increasing attention from both research and industry. A large variety of fairness-aware machine learning solutions have been proposed which propose 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 among the different attributes, particularly w.r.t. protected attributes and class attributes, using a Bayesian network. For a deeper understanding of bias and fairness in the datasets, we investigate the interesting relationships using exploratory analysis.

Keywords: datasets for fairness, fairness-aware machine learning, bias, discrimination, benchmark datasets

1 Introduction

Artificial Intelligence (AI) and Machine Learning (ML) 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 [1], recruiting [2], school admission [3], risk prediction [4], etc. There are many advantages of using algorithmic decision-making as computers can quickly 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 [5], 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 comparing to female ones [6, 7].

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 [8] 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” [9] or the “credit rating” can be used as a proxy for “safe driving” [10]. Hence, even if the protected attributes like race or gender are not used, the resulting ML models can still be biased [5] due to the causal effects of such attributes. Although methods for detection of proxy attributes exist, e.g., [11] 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 a lot of interest in recent years, and as a result, several surveys exist that provide a broad overview of the area, its technical challenges and solutions [12–15]. However, an overview of the datasets used for fairness-aware machine learning evaluation is still missing. As data is a vital part of ML and benchmark datasets a decisive factor for the success of AI research¹, we believe our survey is serving to fill a gap in the extant research.

In this paper, 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 the 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 provides an illustration of the conditional dependence/independence between the protected attribute and the class label; thus, it reduces the space and complexity of data analysis that needs to be performed to discover and 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 on measures related to predictive and fairness performance.

The rest of the paper is structured as follows: In Section 2, we present 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 presents a quantitative evaluation of a classification model on the different datasets w.r.t. predictive performance and fairness. Finally, the conclusion and outlook are summarized in Section 5.

2 Methodology of the survey process

In this section, we describe our strategy for dataset collection, and we 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 (GS)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 GS 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, etc. 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 repository³, Kaggle⁴, etc., are not taken into consideration.

2.2 Bayesian network

A Bayesian network (BN) [16] is a directed and acyclic probabilistic graphical model (PGM)

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

²<https://scholar.google.com>

³<https://archive.ics.uci.edu>

⁴<https://www.kaggle.com>

which provides a graphical representation to understand the complex relationships between a set of random variables. In the case of a dataset, the random variables correspond 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 consists of a set of nodes of random variables/attributes \mathcal{V} and a set of directed edges \mathcal{E} . Let X_1, X_2, \dots, X_d be the attributes that define 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., given that we know the parents of X_i denoted by X_{pa_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 the given attributes, i.e., the joint probability distribution of the attributes in the form of conditional independence relations. Formally:

$$P(X_1, X_2, \dots, X_d) = \prod_{i=1}^d P(X_i | X_{pa_i}) \quad (1)$$

Learning the structure of a BN from the data \mathcal{D} is an optimization problem [17], namely to learn an optimal BN model \mathcal{M}^* which maximizes the likelihood of generating \mathcal{D} . The set of parameters of any BN model \mathcal{M} , denoted by $|\mathcal{M}|$, is the set of edges \mathcal{E} which represents the conditional independence relationship between the attribute set \mathcal{V} . Moreover, among the possible models M , the less complex one, i.e., the one with the least $|\mathcal{M}|$, should be selected.

Note that in a learned BN \mathcal{M} , the position of the class attribute y can be in any position (root-, kinternal- or leaf-node), since the objective is to maximize $P(\mathcal{D} | \mathcal{M})$. 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:

$$\begin{aligned} \max_{\mathcal{M}^*} & \{P(\mathcal{D} | \mathcal{M}) - \gamma |\mathcal{M}| \} \\ & \text{subjected to } y \in \mathcal{L} \end{aligned} \quad (2)$$

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 $P(X_i | X_{pa_i})$ for each $X_i \in \mathcal{X}$ (Eq. 1 and Eq. 2).

A high conditional probability often refers to a strong correlation [18]. 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 among them will be negligible. As a consequence, if we find any direct/indirect edge from any protected attribute to the class attribute in our learned BN structure \mathcal{M}^* 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 [19]. 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 the next section.

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 [20, 21]. For our quantitative analysis, we report on three prevalent fairness measures: *statistical parity (Parity)*, *equalized odds (Eq.Odds)* and *Absolute Between-ROC Area (ABROCA)*.

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 $S \in \{s, \bar{s}\}$ with s and \bar{s} denoting the protected and non-protected groups, respectively. We use the notation s_+ (s_-), \bar{s}_+ (\bar{s}_-) to denote

the protected and non-protected groups for the positive (negative, respectively) class.

2.3.1 Statistical parity

Statistical parity (shortly *Parity*) [22, 23] reports on the percentage difference between two populations w.r.t. the positive class. It is formally defined as follows:

$$\text{Parity} = \frac{|\{x \in \mathcal{D} \mid S = \bar{s}, y = +\}|}{|\{x \in \mathcal{D} \mid S = \bar{s}\}|} - \frac{|\{x \in \mathcal{D} \mid S = s, y = +\}|}{|\{x \in \mathcal{D} \mid S = s\}|} \quad (3)$$

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

2.3.2 Equalized odds

Equalized odds [24] (shortly *Eq.Odds*) measures the difference in prediction errors between the protected and non-protected groups. Let ΔFPR and ΔFNR denote the differences in *false positive rates* and *false negative rates*, respectively between the protected and non-protected groups, defined as follows:

$$\begin{aligned} \Delta FPR &= P(y \neq \hat{y} \mid S = \bar{s}_-) - P(y \neq \hat{y} \mid S = s_-) \\ \Delta FNR &= P(y \neq \hat{y} \mid S = \bar{s}_+) - P(y \neq \hat{y} \mid S = s_+) \end{aligned} \quad (4)$$

where y is the true class label, \hat{y} is the predicted label.

Equalized odds aims at minimizing both ΔFPR and ΔFNR and is defined as:

$$\text{Eq.Odds} = |\Delta FPR| + |\Delta FNR| \quad (5)$$

The value domain is: $\text{Eq.Odds} \in [0, 2]$, 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 [25]. It is based on the Receiver Operating Characteristics (ROC) curve. ABROCA

measures the divergence between the protected (ROC_s) and non-protected group ($ROC_{\bar{s}}$) 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:

$$\int_0^1 |ROC_s(t) - ROC_{\bar{s}}(t)| dt \quad (6)$$

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 society domain (Section 3.3) and educational datasets (Section 3.4). A summary of the statistics of the different datasets 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 among 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).

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	Gender, race, age	Income	1994	The US
KDD Census-Income	299,285	284,556	32/2/7	Binary	Finance	1:15.30	Sex, race	Income	1994-1995	The US
German credit	1,000	1,000	13/1/7	Binary	Finance	2.33:1	Sex	Credit score	1973-1975	Germany
Dutch census	189,725	60,420	10/2/0	Binary	Finance	1:1.1	Sex, age	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	Default payment	2005	Taiwan
COMPAS recid.	7,214	6,172	31/6/14	Binary	Criminology	1:1.20	Race	Two year recidivism	2013-2014	The US
COMPAS viol. recid.	4,743	4,020	31/6/14	Binary	Criminology	1:5.17	Race	Two year violent recid.	2013-2014	The US
Communities&Crime	1,994	1,994	4/0/123	Multi	Criminology	-	Black	Violent crimes rate	1995	The US
Diabetes	101,766	45,715	33/7/10	Binary	Healthcare	1:3.13	Gender	Readmit in 30 days	1999-2008	The US
Ricci	118	118	0/3/3	Binary	Society	1:1.11	Race	Promotion	2003	The 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	The US

3.1 Financial datasets

3.1.1 Adult dataset

The adult dataset [26] (also known as "Census Income" dataset⁵) is one of the most popular datasets for fairness-aware classification studies [9, 27–75]. The classification task is to decide based on demographic characteristics whether the annual income of a person exceeds 50K dollar.

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 [28, 30, 41, 70]. Missing values exist in 3,620 (7.41%) records. Several studies remove instances with missing values [35, 39, 51, 71] from their experiments; other researches consider the whole dataset or do not clarify how the missing values were handled. To avoid the effect of missing values on the analysis, we remove the missing data and obtain a clean dataset of 45,222 instances.

Protected attributes: Typically the following attributes have been used as bias triggers in the literature⁶:

- *gender* = {*male*, *female*} : the dataset is dominated by male instances. The *male:female* ratio is 32,650:16,192 (66.9%:33.1%).
- *race* = {*white*, *black*, *asian-pac-islander*, *american-indian-eskimo*, *other*} . Typically, *race* is used as a binary attribute in the related work [33, 39, 47]: *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 [39]: *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%).

Bayesian network: Fig. 1 illustrates the Bayesian network learned from the dataset. The class label *income* is the leaf node, i.e., there are

⁵<https://archive.ics.uci.edu/ml/datasets/adult>

⁶Please note that the majority of fairness-aware ML methods can handle only single protected attributes. The problem of multi-fairness has only recently been addressed [76–78]

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	Represents the employment status
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
gender	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

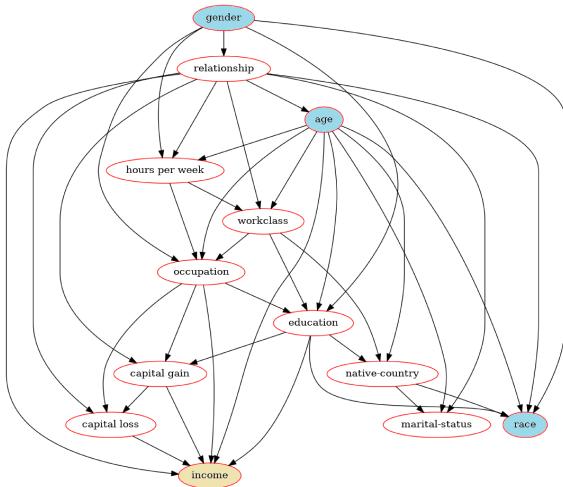


Fig. 1: Adult: Bayesian network (class label: *income*, protected attributes: *gender*, *race*, *age*)

no outgoing edges. To generate the Bayesian network, we discretize the 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 into binary 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}.

As demonstrated in Fig. 1, there is a direct dependency between *income* and *education* as well as between *gender* and *education*. Therefore, we

explore in more detail the distribution of the population w.r.t. *education*, *income* and *gender* in Fig. 2. 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 *gender*, we plot the weekly working hours w.r.t *income* and *gender* (Fig. 3). The number of males who work more than 40 hours per week is more than 7 times higher than that of the females.

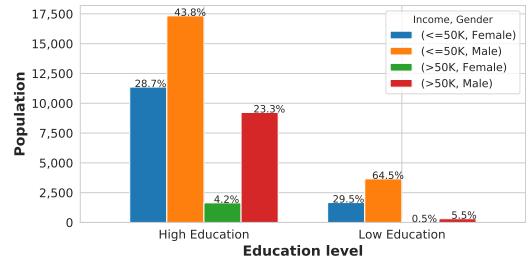


Fig. 2: Adult: distribution of income and gender w.r.t education

Interestingly, there are many outgoing edges from the *relationship* and *age* attributes in the BN. We show the distribution of *gender* in each class based on the *age* (x-axis) and the *relationship* status (y-axis) in Fig. 4. A first observation is that a great amount of young (less than 25 years old) or old (more than 50 years old) people do not receive more than 50K. “Unmarried” people

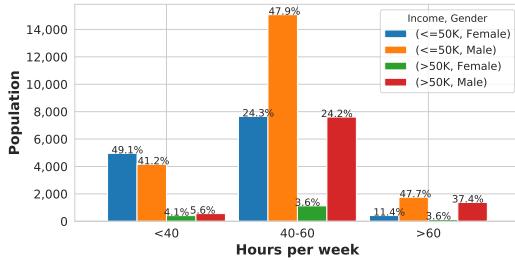


Fig. 3: Adult: distribution of weekly working hours and income w.r.t gender

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.

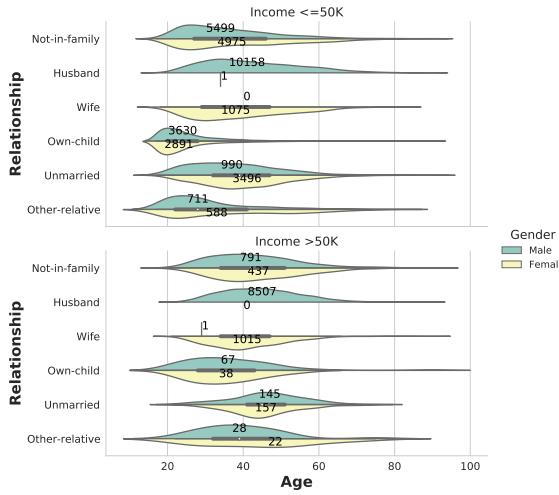


Fig. 4: Adult: distribution of age, relationship and income w.r.t gender

3.1.2 KDD Census-Income dataset

The KDD Census-Income⁷ dataset [79] 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 [46, 71, 73, 80]. The prediction

task is to decide if a person receives more than 50 thousand dollars annually or not. The prediction task is the same as the *Adult* dataset. However, the differences between the two datasets described by the authors [79] 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 is shown in Table 3. The attribute *weight* is omitted as proposed by the authors [79].

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 missing values, as illustrated in Table 3. The result is a cleaned dataset with 284,556 instances.

Protected attributes: Previous researches consider *sex* as the protected attribute [46, 71, 80]. The attribute *race* = {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 *Adult*, the 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%).

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,

⁷[https://archive.ics.uci.edu/ml/datasets/Census-Income+\(KDD\)](https://archive.ics.uci.edu/ml/datasets/Census-Income+(KDD))

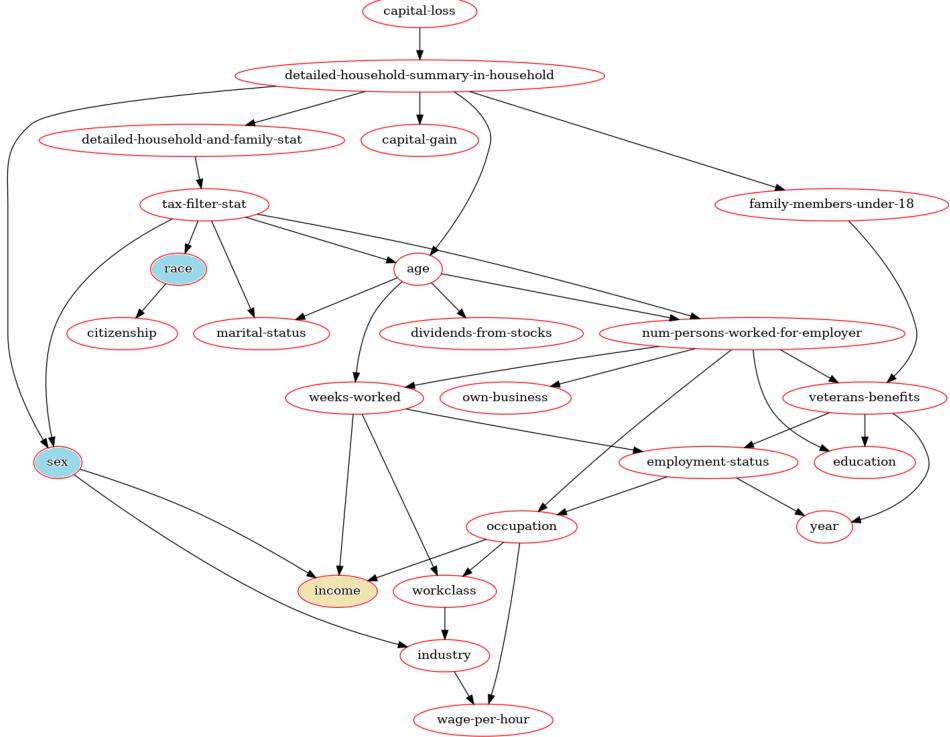


Fig. 5: KDD Census-Income: Bayesian network (class label: *income*, protected attributes: *sex*, *race*)

$>500\}$; $\text{capital-loss} = \{\leq 500, >500\}$; $\text{dividends-from-stocks} = \{\leq 500, 501-2000, >2000\}$; $\text{num-persons-worked-for-employer} = \{0, >0\}$; $\text{weeks-worked-in-year} = \{\leq 26, 27-51, 52\}$. The ranges of encoded attributes are chosen to ensure each group has values. To reduce the complexity, we eliminate the 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 BN on a randomly selected 10% sample of the dataset rather than on the complete dataset. The learned BN is shown in Fig. 5; the class label *income* is set as a leaf node.

As shown in Fig. 5, *income* is conditionally dependent on *sex*, *occupation* and the number of week worked in year (*weeks-worked*) attributes. Regarding sex attribute, females are largely under-represented 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 Fig. 6, 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 than men even though they all work 52 weeks a year. That is shown by the number of men with high income is approximately five times more than the number of women.

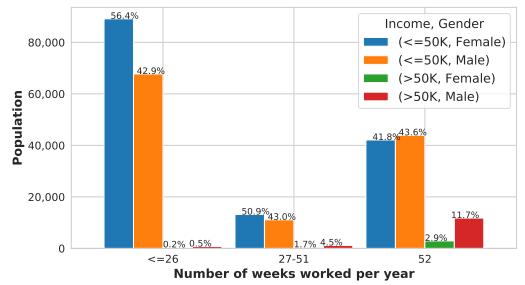


Fig. 6: KDD Census-Income: relationship of the number of weeks worked in a year and income w.r.t gender

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 the employment status
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
enroll-in-edu-inst-last-wk	Categorical	3	0	An individual enrolled in an educational institute last week?
marital-status	Categorical	7	0	The marital status
major-industry	Categorical	24	0	The major industry code
major-occupation	Categorical	15	0	The major occupation code
race	Categorical	5	0	Race
hispanic-origin	Categorical	9	1,279	The Hispanic origin
sex	Binary	{Male, Female}	0	The biological sex of the individual
member-union	Categorical	3	0	Member of a labor union
reason-unemployment	Categorical	6	0	The reason for unemployment
employment-status	Categorical	8	0	The employment status
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
region-previous	Categorical	6	0	The region of previous residence
state-previous	Categorical	50	1038	The state of previous residence
household-family-stat	Categorical	38	0	The detailed household and family
household-summary	Categorical	8	0	The detailed household summary
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
num-persons-worked	Numerical	[0 - 6]	0	The number of persons worked for the employer
family-members-under-18	Categorical	5	0	Family members under 18
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
citizenship	Categorical	5	0	The citizenship
own-business	Categorical	3	0	Own business or self employed
fill-questionnaire	Categorical	3	0	Fill the questionnaire for veteran's admin
veterans-benefits	Categorical	3	0	Veterans benefits
weeks-worked-in-year	Numerical	[0 - 52]	0	The number of weeks worked in a year
year	Categorical	2	0	The year in which the interviewee answered
income	Binary	{≤50K, >50K}	0	Whether an individual makes more than \$50,000 annually

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 among these attributes, the protected attribute *sex* and the class label *income* in Fig. 7. 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.

3.1.3 German credit dataset

The German credit⁸ dataset [79] 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. Customers are classified into two classes: {Good, Bad}. The dataset is frequently employed in fairness-aware learning research [30, 33, 35, 37, 38, 42, 44–46, 51, 69, 81–87].

Dataset characteristics: The dataset contains only 1,000 instances without any missing values. Each sample is described by 13 categorical, 7

⁸[https://archive.ics.uci.edu/ml/datasets/statlog+\(german+credit+data\)](https://archive.ics.uci.edu/ml/datasets/statlog+(german+credit+data))

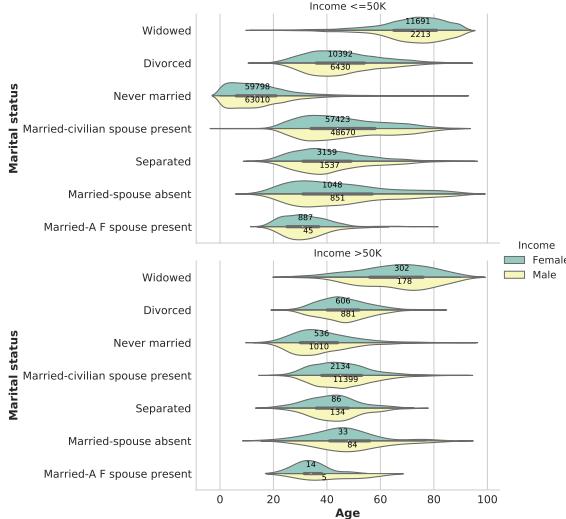


Fig. 7: KDD Census-Income: relationship of marital status, age, sex and income

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.

Protected attributes: Typically, in all studies, *sex* is considered as the protected attribute. *Age* can also be considered as the protected attribute after binarization into $\{\text{young}, \text{old}\}$ by *age* thresholding at 25 [45].

- $sex = \{\text{male}, \text{female}\}$: 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* one is 27.2%.

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 BN is shown in Fig. 8; *class-label* is set as a leaf node.

The BN consists of two disconnected components. In the first component, *class-label* is conditionally dependent on the *checking-account* attribute. We investigate in more detail this relationship in Fig. 9. 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 (stands for *Deutsche Mark*) in their checking account are classified as the *bad* customers.

In the second component, 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 Fig. 10, the number of customers who have to pay the highest installment rate (visualized as the “red” columns) is inversely proportional to the *credit-amount*.

3.1.4 Dutch census dataset

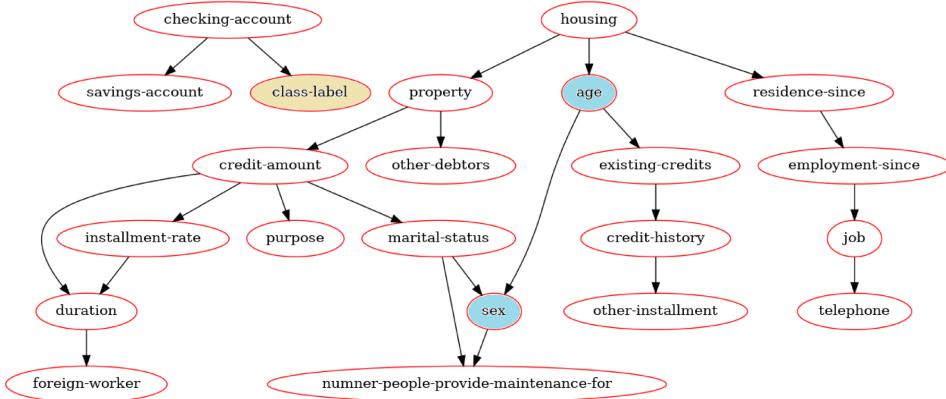
The Dutch census of the year 2001 [88] represented aggregated groups of people in the Netherlands for the year 2001. Researchers [28, 29, 31, 34, 34, 49, 54, 55] have used Dutch dataset to formulate a binary classification task to predict a person’s *occupation* which can be categorized as *high level* (prestigious) or *low level* profession.

Dataset characteristics: The dataset includes 189,725 samples where each sample is described by 12 attributes. An overview of attributes is presented in Table 5. Typically, in the literature, the dataset is pre-processed by dropping samples of *under-aged* people and people whose profession is *unknown* or *middle level* which leads to 129,305 removed samples. The cleaned data contains 60,420 instances.

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

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
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
existing-credits	Numerical	[1 - 4]	0	Number of existing credits at this bank
job	Categorical	4	0	Job
number-people	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

**Fig. 8:** German credit: Bayesian network (class label: *class-label*, protected attributes: *sex, age*)**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 of the person
household_position	Categorical	8	0	The household position
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

Bayesian network: We use all attributes in the dataset to generate the Bayesian network. As illustrated in Fig. 11, 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

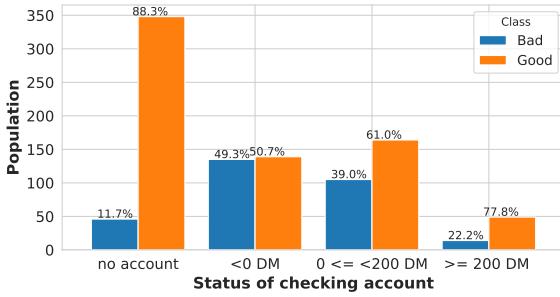


Fig. 9: German credit: distribution of class label on status of checking account

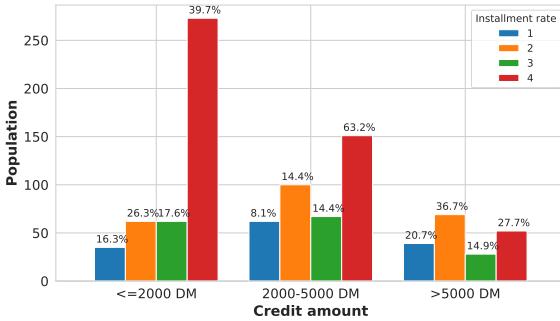


Fig. 10: German credit: relationship between credit amount and installment rate

only 32.7%. In addition, people with high education are doing prestigious jobs and vice versa, as depicted in Fig. 12. For example, 89.5% of people having *tertiary* level are working in high-level jobs while around 80% of people with *lower secondary* degrees are doing low-level work. Interestingly, *age* has a direct effect on many attributes.

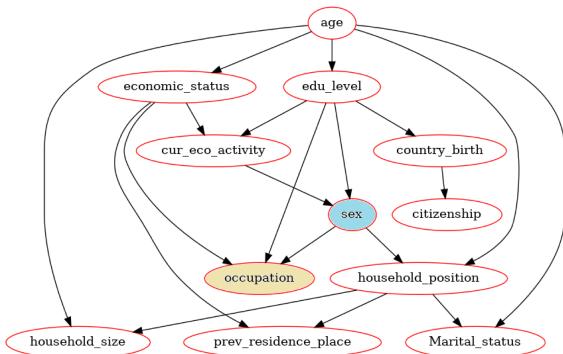


Fig. 11: Dutch census: Bayesian network (class label: *occupation*, protected attribute: *sex*)

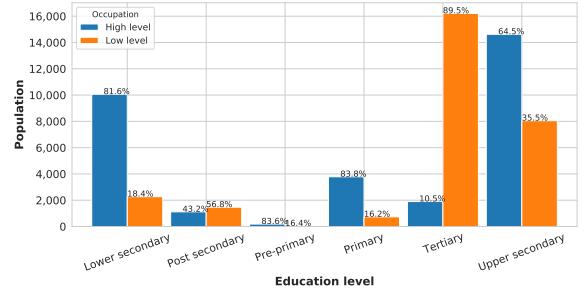


Fig. 12: Dutch census: relationship between education level and occupation

3.1.5 Bank marketing dataset

The bank marketing⁹ dataset [89] 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 [27, 39, 44, 50, 53, 56–60, 64–66, 74, 75, 90]. 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 [56–59, 90]. Besides, in the studies [27, 39, 44], 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%).
- *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%).

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* = {≤15, >15}; *duration* = {≤120, 121-600, >600}; *campaign* =

⁹<https://archive.ics.uci.edu/ml/datasets/Bank+Marketing>

Table 6: Bank marketing: attributes characteristics

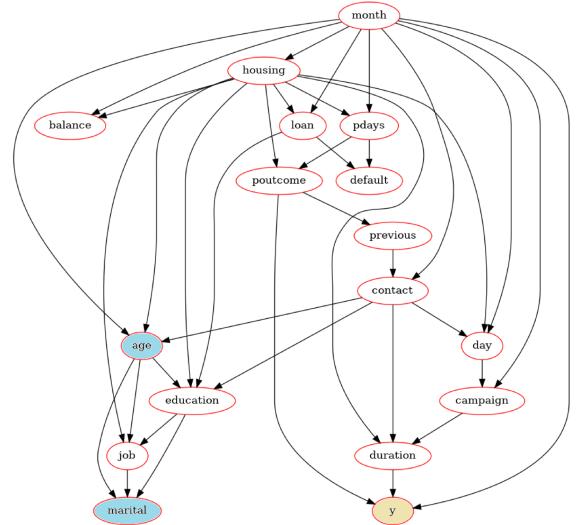
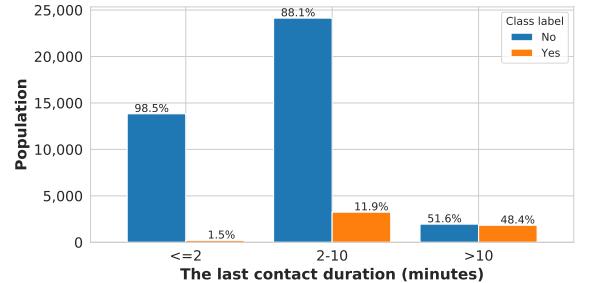
Attributes	Type	Values	#Missing values	#Missing values
age	Numerical	[18 - 95]	0	The age of the client
job	Categorical	12	0	The type of job
marital	Categorical	3	0	The marital status
education	Categorical	4	0	The education level
default	Binary	{Yes, No}	0	Has the credit in default?
balance	Numerical	[-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?

$\{\leq 1, 2-5, >5\}$; $pdays = \{\leq 30, 31-180, >180\}$; $previous = \{0, 1-5, >5\}$. Fig. 13 visualizes the Bayesian network of the Bank marketing dataset. Class label y , as illustrated in Fig. 13, 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 Fig. 14. 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 the bank 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 .

3.1.6 Credit card clients dataset

The credit card clients¹⁰ dataset [4] 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 [4, 61–63, 65, 68, 90].

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.

**Fig. 13:** Bank marketing: Bayesian network (class label: y , protected attributes: age , $marital$)**Fig. 14:** Bank marketing: Relationship between the last contact duration and class label

¹⁰<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	Education
marriage	Categorical	4	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_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

Protected attributes: In the literature, *sex* [61, 63, 68], *education*, *marriage* [63, 90] 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.

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*) = {≤50000, 50001-200000, >200000} (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 Fig. 15. 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 Fig. 16, the ratio of the default payment phenomenon is inversely proportional to the credit limit balance. Interestingly, the protected attributes (*sex*, *education*, *marriage*) are conditionally dependent on each other. 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%.

3.2 Criminological datasets

3.2.1 COMPAS dataset

The COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) [5] is a recent dataset, compared to the rest of the datasets in our work, which was released by ProPublica¹¹ 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 [27, 36, 45, 48–52, 54, 55, 60, 61, 69, 72, 91–101]. *Risk of recidivism* (denoted as *COMPAS recid.*) and *Risk of violent*

¹¹<https://www.propublica.org/datastore/dataset/compas-recidivism-risk-score-data-and-analysis>

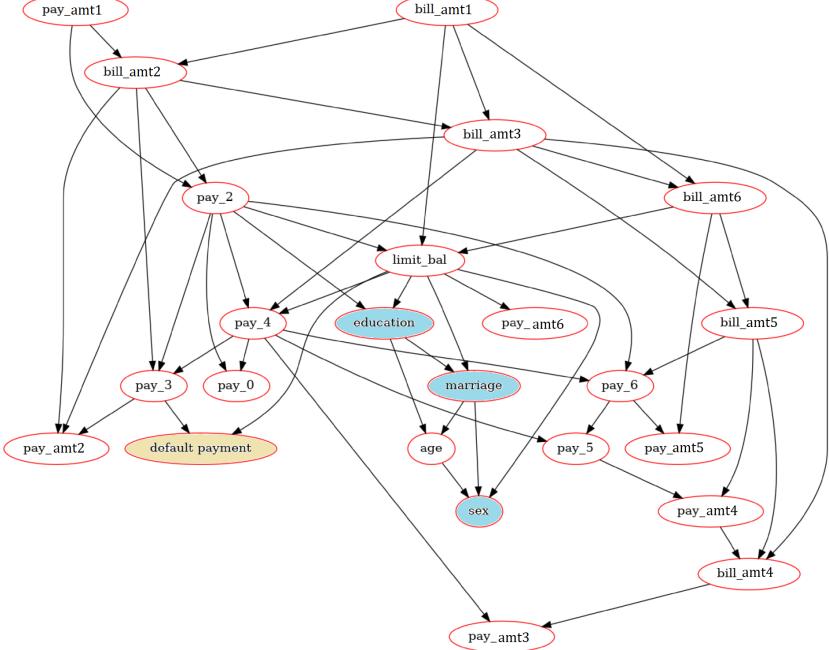


Fig. 15: Credit card clients: Bayesian network (class label: *default payment*, protected attributes: *sex*, *marriage*, *education*)

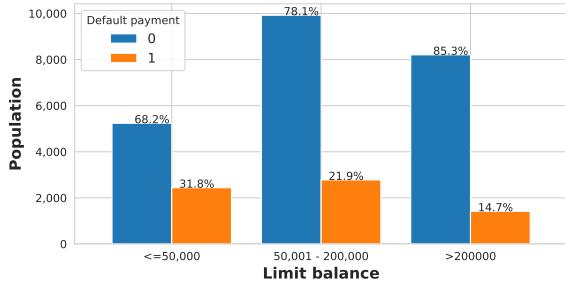


Fig. 16: Credit card clients: Relationship between the credit limit balance and default payment

recidivism (denoted as *COMPAS viol. recid*) subsets are most widely used in the literature. The former data 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 (31 categorical, 6 binary, 14 numerical and a null attribute), as shown in Table 8. As summarized in Table 8,

missing data is a common phenomenon in both subsets. There are 6,395 rows (88.6%) containing missing values in COMPAS recid. subset while this number of COMPAS viol. revid. subset is 3,748 instances (79%). Based on [5], we clean the dataset by removing the missing data, such as “*score_text*” = “*N/A*” 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%) while the ratio in COMPAS viol. recid. subset is 1,918:1,459 (47.7%:36.3%). Fig. 17 and Fig. 18 describe the distribution of defendants w.r.t. *race*. The recidivism rate in the black defendants is higher than that of the white defendants in both subsets.

Bayesian network: To generate the Bayesian network, we remove the temporal attributes such as *screening_date* (the date on which the risk of recidivism score was given), *in_custody* (the date

Table 8: COMPAS recid: attributes characteristics

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
sex	Binary	{Male, Female}	0	Sex
dob	Categorical	5,452	0	Date of birth
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
decile_score	Numerical	[1 - 10]	0	The COMPAS Risk of Recidivism score
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
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 (days)
c_charge_degree	Binary	{F, M}	0	Charge degree of original crime
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	The number of days between the follow-up crime and the arrest date
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
score_text	Categorical	3	0	ProPublica-defined category 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.score_text	Categorical	3	0	ProPublica-defined category of v.decile_score, (High, Medium, Low)
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
two_year_recid	Binary	{0, 1}	0	Whether the defendant is rearrested within two years

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}. Fig. 19 and Fig. 20 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 Fig. 21 and Fig. 22. 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 Fig. 20. To

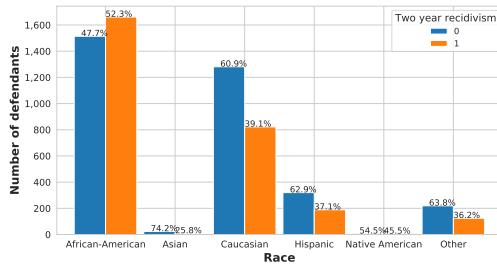


Fig. 17: COMPAS recid.: distribution of *Two year recidivism* w.r.t. *race*

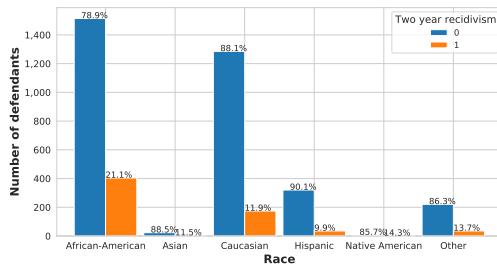


Fig. 18: COMPAS viol. recid.: distribution of *Two year recidivism* w.r.t. *race*

clarify this phenomenon, we investigate the distribution of age, recidivism score (*score_text*) w.r.t. race, in Fig. 23. The majority of recidivists are under the age of 30. In the recidivist group, the number of black criminals is 4 times and 2 times more than that of white criminals, although they have the same high and 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 Crime¹² dataset [79] was a small dataset containing the socio-economic data from 46 states of the United States in 1990 (US Census). The law enforcement data come from the 1990 US LEMAS survey, and crime data come from the 1995 FBI UCR. The goal is to predict the total number of violent crimes per 100 thousand population. Many researchers are investigating the

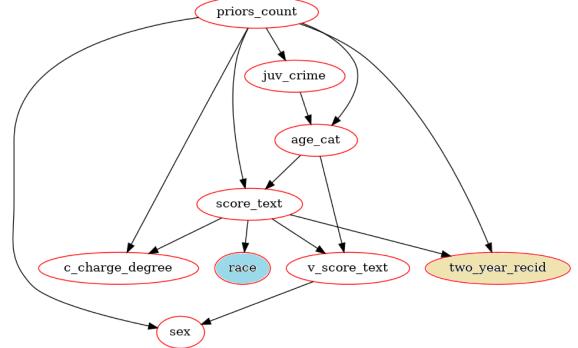


Fig. 19: COMPAS recid.: Bayesian network (class label: *two_year_recid*, protected attribute: *race*)

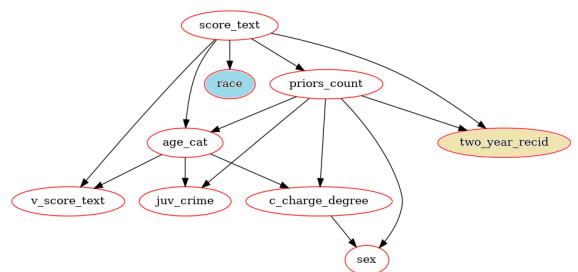


Fig. 20: COMPAS viol. recid.: Bayesian network (class label: *two_year_recid*, protected attribute: *race*)

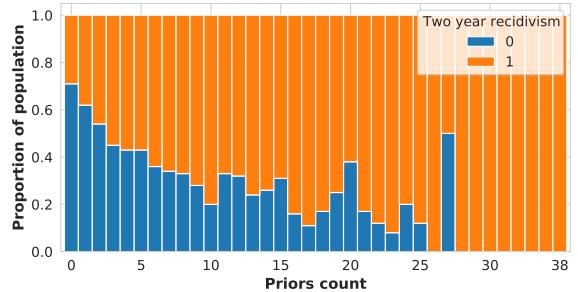


Fig. 21: COMPAS recid.: Relationship between recidivism and priors count

dataset in their experiments [28, 29, 34, 61, 67, 69, 74, 97, 99, 100, 102–106].

Dataset characteristics: The dataset contains only 1,994 samples; each instance is described by 127 attributes (4 categorical and 123 numerical attributes). A summary of attributes is illustrated in Table 9.

¹²<http://archive.ics.uci.edu/ml/datasets/communities+and+crime>

Table 9: Communities and Crime: attributes characteristics

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	1177	The numeric code for community
communityname	Categorical	1828	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
racePctblack	Numerical	[0.0 - 1.0]	0	The percentage of population that is African American
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
pctWInvInc	Numerical	[0.0 - 1.0]	0	The percentage of households with investment/rent income in 1989
pctWSocSec	Numerical	[0.0 - 1.0]	0	The percentage of households with social security income in 1989
pctWPubAsst	Numerical	[0.0 - 1.0]	0	The percentage of households with public assistance 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
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
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
MalePctDivorce	Numerical	[0.0 - 1.0]	0	The percentage of males who are divorced
MalePctNevMarr	Numerical	[0.0 - 1.0]	0	The percentage of males who have never married
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
PctFam2Par	Numerical	[0.0 - 1.0]	0	The percentage of families (with kids) that are headed by two parents
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
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
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
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 immigrated within the last 3 years
PctImmigRec5	Numerical	[0.0 - 1.0]	0	The percentage of immigrants who immigrated within the last 5 years
PctImmigRec8	Numerical	[0.0 - 1.0]	0	The percentage of immigrants who immigrated within the last 8 years
PctImmigRec10	Numerical	[0.0 - 1.0]	0	The percentage of immigrants who immigrated 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)

Table 9: Communities and Crime: attributes characteristics (continued)

Attributes	Type	Values	Missing values	Description
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
PctPersOwnOccup	Numerical	[0.0 - 1.0]	0	The percentage of people in owner occupied households
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
HousVacant	Numerical	[0.0 - 1.0]	0	The number of vacant households
PctHousOccup	Numerical	[0.0 - 1.0]	0	The percentage of housing occupied
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
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
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
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
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 population
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
ViolentCrimesPerPop	Numerical	[0.0 - 1.0]	0	The total number of violent crimes per 100,000 population

There is a very high ratio (84%) of missing values in 25 attributes, as demonstrated in Table 9. Based on the suggestions from the literature, 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, a label “high-crime” is set if the crime rate of the communities 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%).

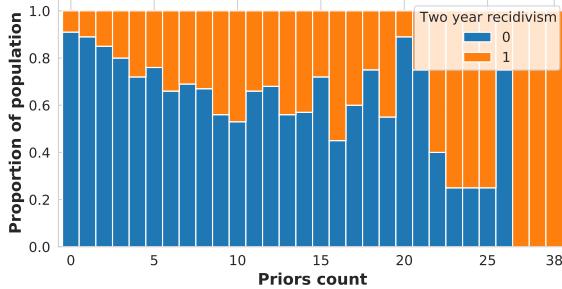


Fig. 22: COMPAS viol. recid.: Relationship between recidivism and priors count

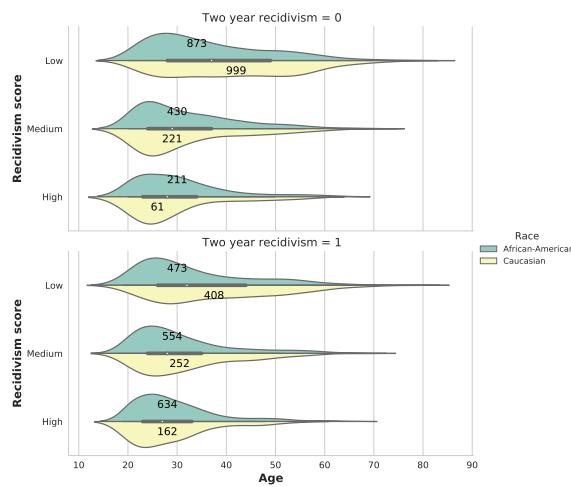


Fig. 23: COMPAS recid. : distribution of recidivism score, age w.r.t. race

Protected attributes: In the literature, typically, researchers derive a new attribute, namely *Black*, which is 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. Therefore, 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 threshold 0.25) with the class label. The Bayesian network is visualized in Fig. 24.

As shown in Fig. 24, 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%.

3.3 Healthcare and social datasets

3.3.1 Diabetes dataset

The diabetes¹³ dataset [107] described the clinical care at 130 US hospitals and integrated delivery networks from 1999 to 2008. The possible classification task is to predict whether the patient will readmit within 30 days or not. The dataset is investigated in several studies [56, 58, 64–66, 108].

Dataset characteristics: The dataset contains 101,766 patients described by 50 attributes (10 numerical, 7 binary and 33 categorical). Characteristics of all attributes are summarized in Table 10. 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 contain 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 rows should be clean. The clean version of the dataset contains 45,715 records.

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%.

¹³<https://archive.ics.uci.edu/ml/datasets/diabetes+130-us+hospitals+for+years+1999-2008>

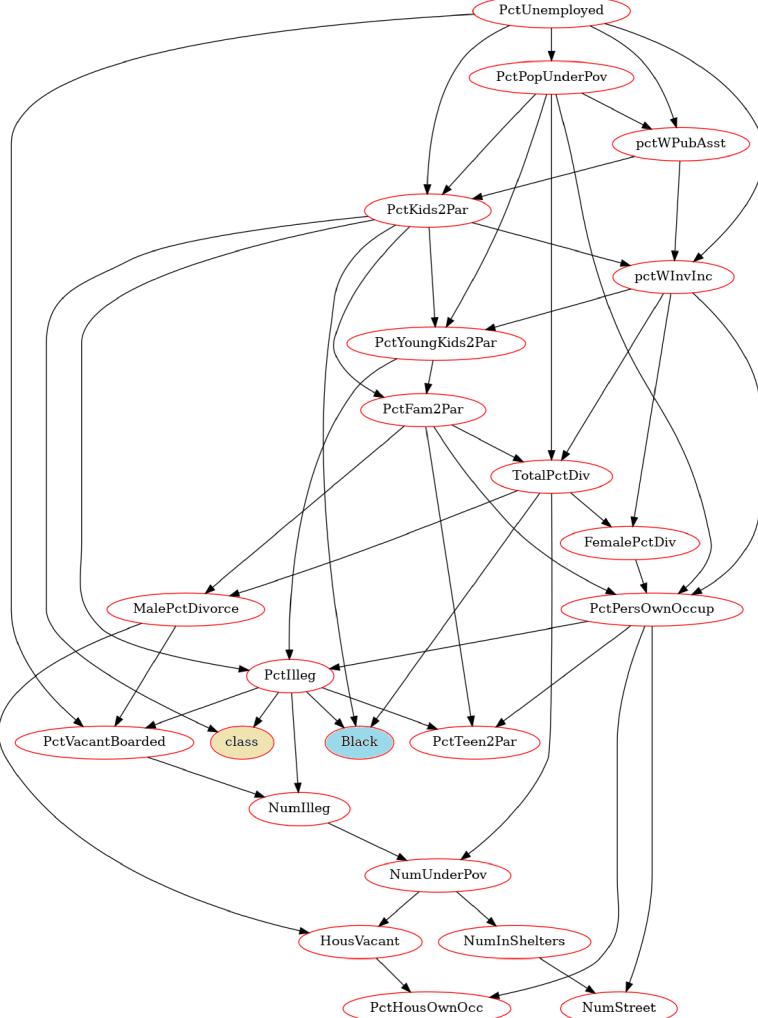


Fig. 24: Communities and Crime: Bayesian network (class label: *class*, protected attribute: *black*)

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* = {≤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 Fig. 25.

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*). Attribute *number_outpatient* also has an impact on 8 other features. Interestingly, there is no connection between protected attributes *gender* and the class label.

3.3.2 Ricci dataset

The Ricci¹⁴ dataset was generated by the Ricci v. DeStefano case [109] in which they investigated the results of a promotion exam within a fire department in November and December of 2003. Although it is a relatively small dataset, it

¹⁴<https://www.key2stats.com/data-set/view/690>

Table 10: Diabetes: attributes characteristics

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
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
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.)
time_in_hospital	Numerical	[1 - 14]	0	The number of days between admission and discharge
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
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
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
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
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
chlorpropamide	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
glipizide	Categorical	4	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
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
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?
diabetesMed	Binary	{Yes, No}	0	Was there any diabetic medication prescribed?
readmitted	Categorical	3	0	The number of days to inpatient readmission

has been employed for fairness-aware classification tasks in many studies [9, 15, 37, 42, 45, 110–112]. 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 summarized in Table 11.

Protected attributes: In this dataset, only the attribute *race* can be used as the 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%).

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 presented in Fig. 26.

It is easy to observe that the combined grade (attribute *combine*) has a direct effect on the class label (*promoted*). Fig. 27 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 Fig. 28, 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 with no promotion.

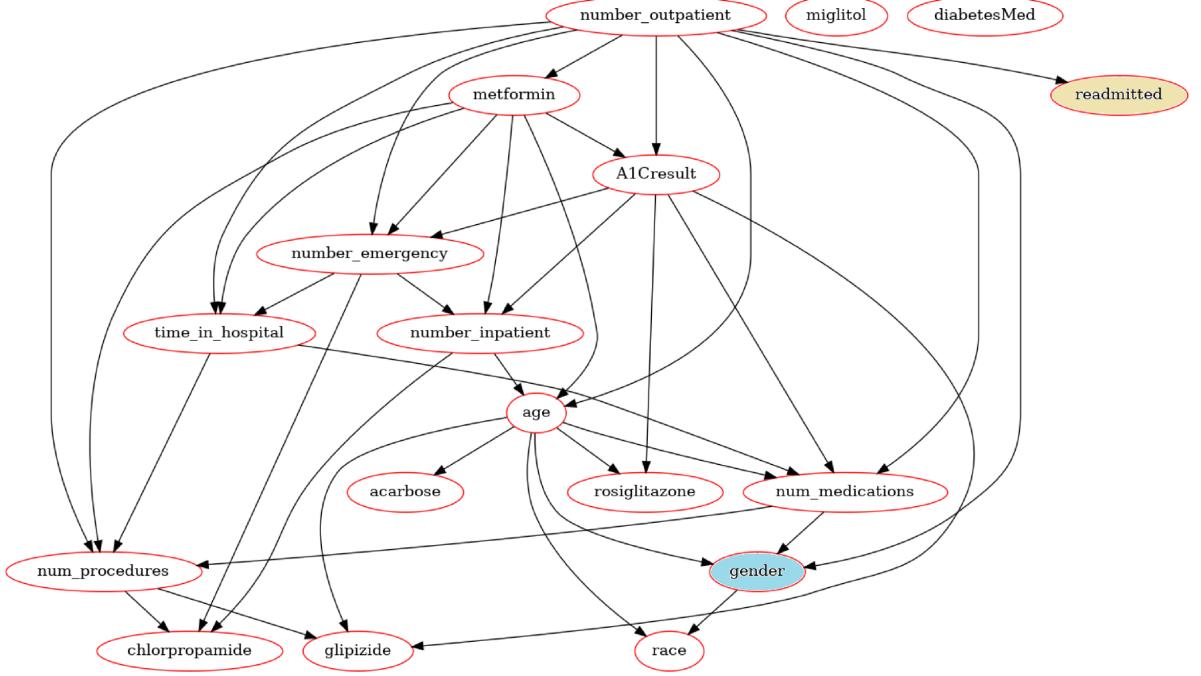


Fig. 25: Diabetes: Bayesian network (class label: *readmitted*, protected attribute: *gender*)

Table 11: Ricci: attributes characteristics

Attributes	Type	Values	#Missing values	Description
Position	Binary	{Lieutenant, Captain}	0	The desired promotion (Captain or Lieutenant)
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

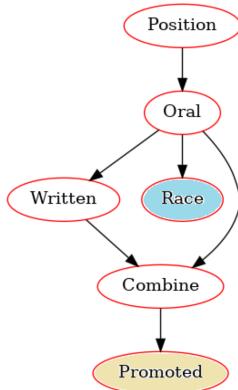


Fig. 26: Ricci: Bayesian network (class label: *promoted*, protected attribute: *race*)

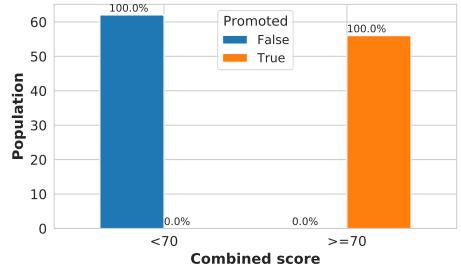


Fig. 27: Ricci: Relationship between combined score and promotion status

3.4 Educational datasets

3.4.1 Student performance dataset

The student performance dataset [113] described students' achievement in the secondary education

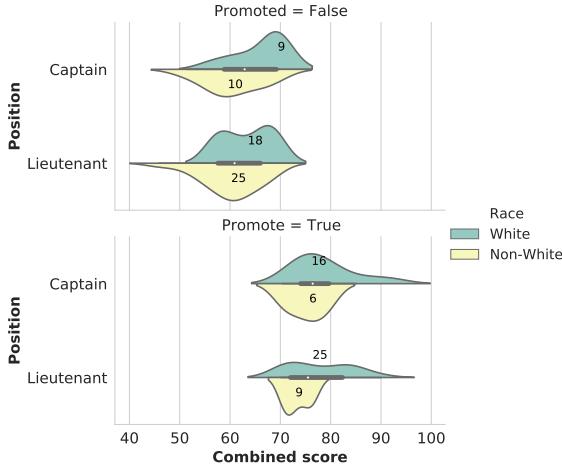


Fig. 28: Ricci: Distribution of combined score, position and promotion decision across race

of two Portuguese schools in 2005 - 2006 with two distinct subjects: Mathematics and Portuguese.¹⁵. The regression task is to predict the final year grade of the students. It is investigated in several studies [63, 67, 106, 114], with fair-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). A summary of the characteristics of the attributes is described in Table 12. To simplify the classification problem, we create a class label based on attribute $G3$, $class = \{Low, High\}$ corresponds to $G3 = \{<10, \geq 10\}$.

Protected attributes: Typically, in the literature, sex is considered as the protected attribute. In several studies [63, 67], they select age as the protected attribute.

- $sex = \{male, female\}$: 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 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 $G1 = \{<10, \geq 10\}$; $G2 = \{<10, \geq 10\}$. 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 Fig. 29 and Fig. 30, respectively.

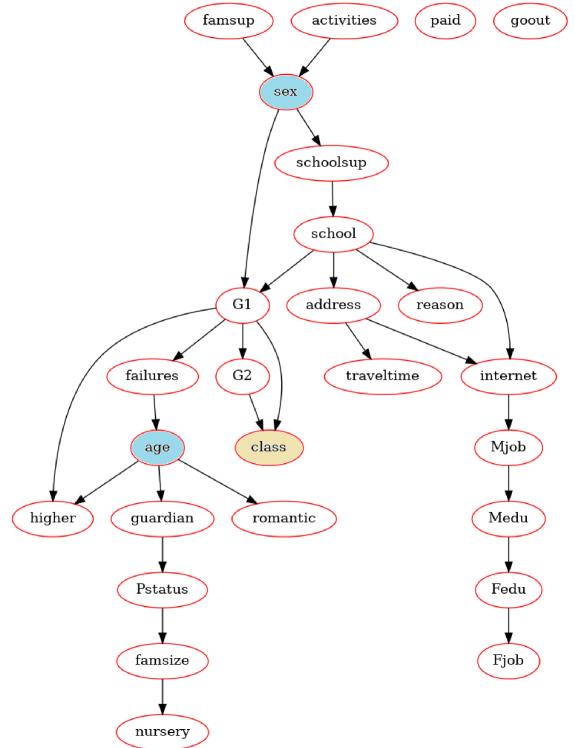


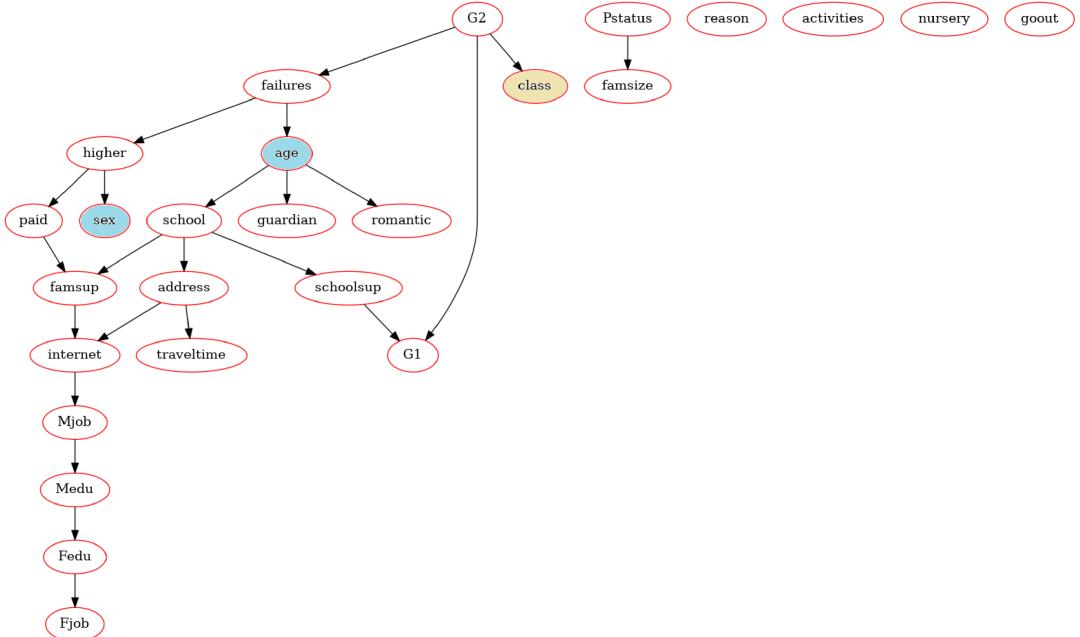
Fig. 29: Student performance - Portuguese subject: Bayesian network (class label: $class$, protected attributes: age, sex)

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. Fig. 31 and Fig. 32 reveal that the male students

¹⁵<https://archive.ics.uci.edu/ml/datasets/student+performance>

Table 12: Student performance: attributes characteristics

Attributes	Type	Values	#Missing values	Description
school	Binary	{GP, MS}	0	The student's school
sex	Binary	{Male, Female}	0	Sex
age	Numerical	[15 - 22]	0	Age (in years)
address	Binary	{U, R}	0	The address type
famsize	Binary	{LE3, GT3}	0	The family size
Pstatus	Binary	{T, A}	0	The parent's cohabitation status
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
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
freetime	Numerical	[1 - 5]	0	Free time after school
goout	Numerical	[1 - 5]	0	How often does the student go out with friends?
Dalc	Numerical	[1 - 5]	0	The workday alcohol consumption
Walc	Numerical	[1 - 5]	0	The weekend alcohol consumption
health	Numerical	[1 - 5]	0	The current health status
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

**Fig. 30:** Student performance - Mathematics subject: Bayesian network (class label: *class*, protected attributes: *age*, *sex*)

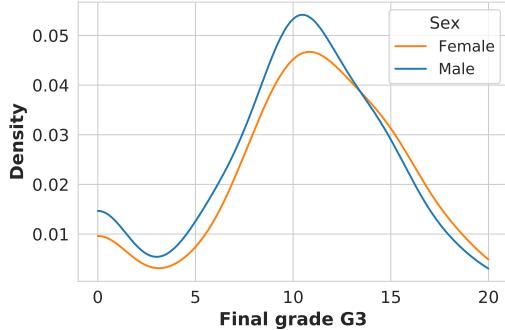


Fig. 31: Student performance - Mathematics: distribution of the final grade G3 w.r.t. Sex

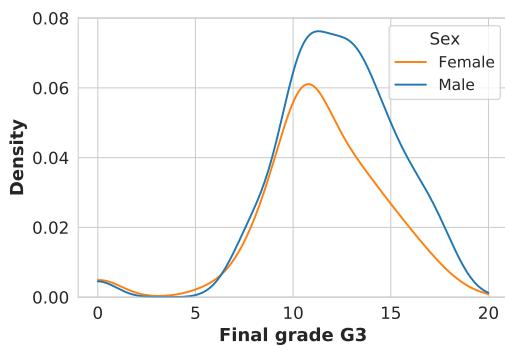


Fig. 32: Student performance - Portuguese: distribution of the final grade G3 w.r.t. Sex

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) dataset¹⁶ was collected from the OU analysis project [115] 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 studies [114, 116, 117], on fairness-aware problems. The goal is to predict the success of students.

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. The *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". The ratio of Pass:Fail is 14,655:6,907 (68%:32%).

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* = {≤100, >100}. The network is depicted in Fig. 33. The final result attribute is directly conditionally dependent on the highest education 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. We

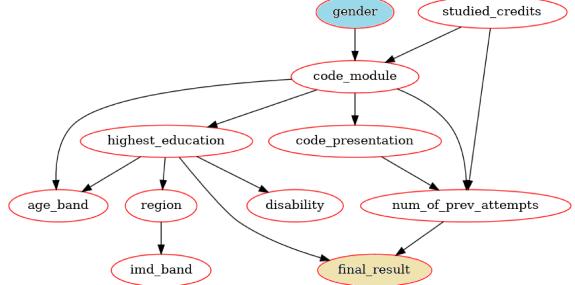


Fig. 33: OULAD: Bayesian network (class label: *final_result*, protected attributes: *gender*)

perform the analysis on the relationship of the highest education, number of previous attempts and the final result for each gender. As demonstrated in Fig. 34, 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.

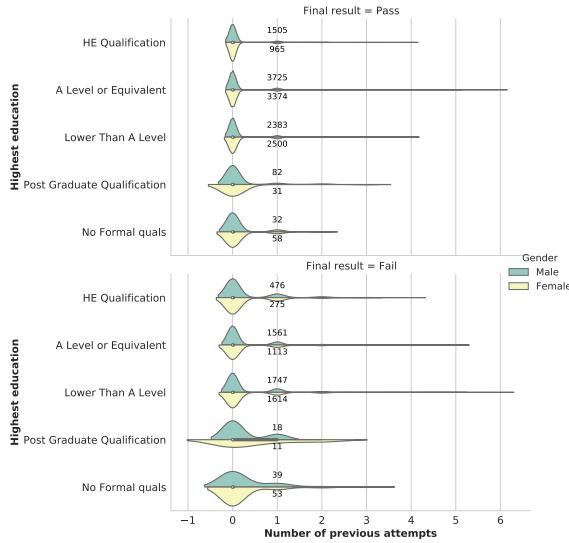
¹⁶https://analyse.kmi.open.ac.uk/open_dataset

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)

Table 14: Law school: 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
racetxt	Categorical	6	0	Race
pass_bar	Binary	{0, 1}	0	Whether the student passed the bar exam on the first try

**Fig. 34:** OULAD: Distribution of the number of previous attempts, the highest education and the final result w.r.t. gender

3.4.3 Law school dataset

The Law school dataset [118] 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 several studies [61, 67–69, 101, 106, 119–121].

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. The class label *pass_bar* = {0, 1} is used for the classification task. The ratio of *pass* (1):*non-pass* (0) is 18,505:2,293 (89%:11%).

Protected attributes: In the literature, *race* [67–69, 101, 106, 119–121] and *male* [61, 67, 119–121] 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%).

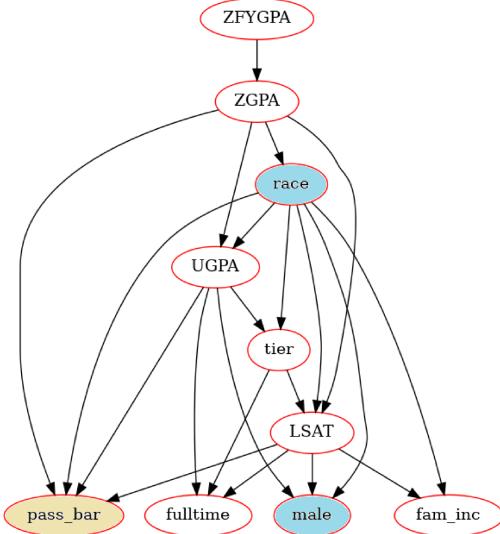


Fig. 35: Law school: Bayesian network (class label: *pass_bar*, protected attributes: *male*, *race*)

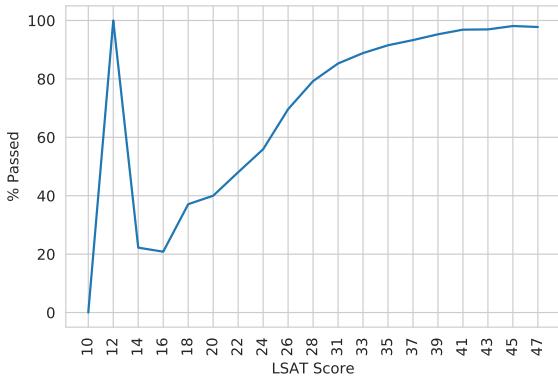


Fig. 36: Law school: The percentage of students that passed the bar exam by LSAT scores

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 Fig. 35.

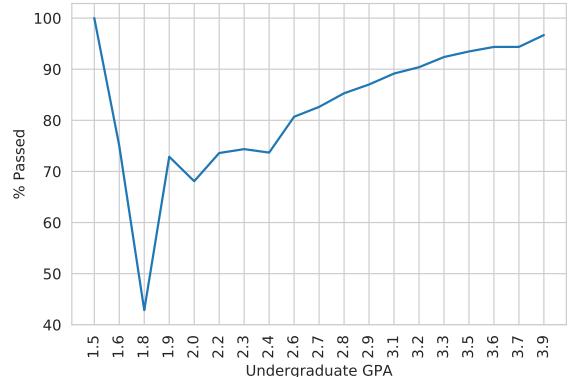


Fig. 37: Law school: The percentage of students that passed the bar exam by UGPA scores

It is easy to observe that the result of the bar exam 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 Fig. 36 and Fig. 37.

4 Experimental evaluation

This section demonstrates our experiments of a classical predictive model on all datasets and reports the results on several fairness metrics.

4.1 Evaluation setup

Predictive model. We use a very simple predictive model, namely *Logistic regression* [122], for the classification task. It is 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 Fig. 38 (in which, *prot* and *non-prot* stand for *protected*, *non-protected*, respectively), we report the performance of the predictive model on the following measures.

		Predicted class	
		Positive	Negative
Actual class	Negative	True Positive (TP) $TP_{prot} + TP_{non-prot}$	False Negative (FN) $FN_{prot} + FN_{non-prot}$
	Positive	False Positive (FP) $FP_{prot} + FP_{non-prot}$	True Negative (TN) $TN_{prot} + TN_{non-prot}$

Fig. 38: Confusion matrix

- Accuracy

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \quad (7)$$

- Balanced accuracy

$$BA = \frac{1}{2} \times \left(\frac{TP}{TP + FN} + \frac{TN}{TN + FP} \right) \quad (8)$$

- True positive rate (TPR) on protected group

$$TPR_{prot} = \frac{TP_{prot}}{TP_{prot} + FN_{prot}} \quad (9)$$

- TPR on non-protected group

$$TPR_{non-prot} = \frac{TP_{non-prot}}{TP_{non-prot} + FN_{non-prot}} \quad (10)$$

- True negative rate (TNR) on protected group

$$TNR_{prot} = \frac{TN_{prot}}{TN_{prot} + FP_{prot}} \quad (11)$$

- TNR on non-protected group

$$TNR_{non-prot} = \frac{TN_{non-prot}}{TN_{non-prot} + FP_{non-prot}} \quad (12)$$

- Statistical parity (Eq. 3)
- Equalized odds (Eq. 5)
- ABROCA (Eq. 6)

Table 15: Performance of logistic regression model on datasets

Dataset	Protected attribute	Group distribution (%)	Accuracy	Balanced accuracy	Statistical Parity	Equalized odds	ABROCA	TPR prot.	TNR non-prot.	
Adult	Gender [$s_+, s_-, \bar{s}_+, \bar{s}_-$]	[3.7, 28.8, 21.1, 46.4] [1.3, 50.7, 4.8, 43.2] [20.1, 10.9, 49.9, 19.1] [16.4, 33.7, 31.2, 18.7] [5.6, 34.2, 6.1, 54.1] [12.5, 47.8, 9.7, 30.0] [31.5, 28.7, 15.5, 24.3] [12.0, 44.8, 5.2, 38.0] [5.8, 46.3, 0.3, 47.6] [11.1, 34.1, 13.1, 41.7] [12.7, 29.7, 34.7, 22.9] [33.7, 19.0, 33.4, 13.9] [51.3, 7.7, 33.3, 7.7] [32.1, 14.2, 35.9, 17.8] [11.5, 4.4, 77.5, 6.6]	0.7864 0.9474 0.6967 0.5713 0.8149 0.8855 0.7822 0.6414 0.8432 0.9683 0.7584 0.10 0.9412 0.9282 0.6751 0.9072	0.6249 0.6031 0.0752 0.1634 0.2985 0.3746 0.0396 0.1322 0.2195 0.7011 0.5 0.3029 0.9360 0.8447 0.5 0.6260	0.1989 -0.0752 0.0748 0.1228 -0.2985 0.0261 -0.0254 0.6452 0.0913 0.4314 0.0 0.10 0.1616 0.0575 0.0490 0.0252 0.1983	0.0281 0.0403 0.1634 0.0983 0.6984 0.025 0.0220 0.0675 0.0584 0.4507 0.0 0.0 0.0177 0.0273 0.0 0.0325	0.3194 0.1825 0.8533 0.2759 0.8382 0.1527 0.0 0.5996 0.1826 0.031 0.0 0.0 0.9354 0.9633 0.10 0.9100	0.3007 0.2195 0.8533 0.2759 0.9219 0.1726 0.0 0.2058 0.0 0.44 0.0 0.0 0.9762 0.9630 0.75 0.9555	0.9521 0.9961 0.9928 0.2419 0.6871 0.9849 1.0 0.6793 0.9606 1.0 1.0 1.0 0.9762 0.9630 0.75 0.5251	0.9426 0.9928 0.2419 0.6871 0.9787 0.9787 1.0 0.9307 0.9975 0.9892 1.0 1.0 1.0 0.9762 0.9630 0.75 0.1063

Training/test set splitting. The ratio of training set and test set in our experiment is 70%:30% 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.

In general, a significant difference in terms of predictive performance and fairness measures is observed among 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, 9/15 datasets have the absolute value of the statistical parity less than 10. The *Diabetes* dataset has the best value of the statistical parity while the *Communities & Crimes* dataset shows the worst value. 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 Fig. 39. 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 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 [123] and images [124, 125] 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 among 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 among 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*, *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

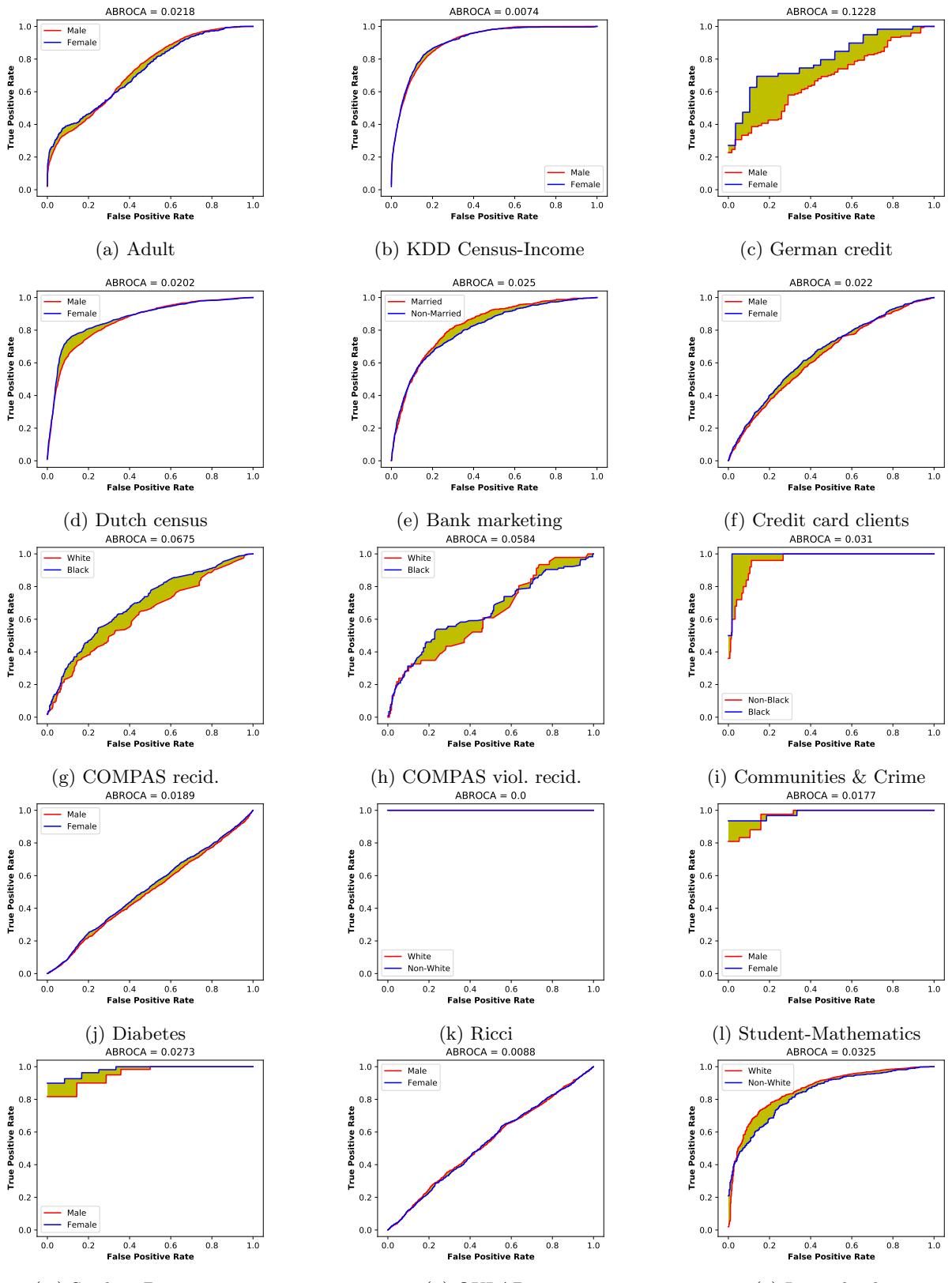


Fig. 39: ABROCA slice plot on datasets

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 [13].

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.

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).

Conflict of interest

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

ORCID

Tai Le Quy  0000-0001-8512-5854
Eirini Ntoutsi  0000-0001-5729-1003

References

- [1] Mukerjee, A., Biswas, R., Deb, K., Mathur, A.P.: Multi-objective evolutionary algorithms for the risk–return trade-off in bank loan management. *International Transactions in Operational research* **9**(5), 583–597 (2002). <https://doi.org/10.1111/1475-3995.00375>
- [2] Faliagka, E., Ramantas, K., Tsakalidis, A., Tzimas, G.: Application of machine learning algorithms to an online recruitment system. In: Proceeding of the International Conference on Internet and Web Applications And Services (2012)
- [3] Moore, J.S.: An expert system approach to graduate school admission decisions and academic performance prediction. *Omega* **26**(5), 659–670 (1998). [https://doi.org/10.1016/S0305-0483\(98\)00008-5](https://doi.org/10.1016/S0305-0483(98)00008-5)
- [4] Yeh, I.-C., Lien, C.-h.: 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 (2009). <https://doi.org/10.1016/j.eswa.2007.12.020>
- [5] Angwin, J., Larson, J., Mattu, S., Kirchner, L.: Machine bias. ProPublica, May **23** (2016)
- [6] Simonite, T.: Probing the dark side of google’s ad-targeting system. *MIT Technology Review* (2015)
- [7] Datta, A., Tschantz, M.C., Datta, A.: Automated experiments on ad privacy settings. *Proceedings on Privacy Enhancing Technologies* **2015**(1), 92–112 (2015)

- [8] Madras, D., Creager, E., Pitassi, T., Zemel, R.: 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 (2019). <https://doi.org/10.1145/3287560.3287564> //doi.org/10.1007/978-3-540-85066-3_1
- [9] Datta, A., Fredrikson, M., Ko, G., Mardziel, P., Sen, S.: Proxy non-discrimination in data-driven systems. arXiv preprint arXiv:1707.08120 (2017)
- [10] Warner, R., Sloan, R.H.: Making artificial intelligence transparent: Fairness and the problem of proxy variables. *Criminal Justice Ethics* **40**(1), 23–39 (2021). <https://doi.org/10.1080/0731129X.2021.1893932>
- [11] Yeom, S., Datta, A., Fredrikson, M.: Hunting for discriminatory proxies in linear regression models. In: Proceedings of the 32nd International Conference on Neural Information Processing Systems, pp. 4573–4583 (2018)
- [12] Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., Galstyan, A.: A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)* **54**(6), 1–35 (2021). <https://doi.org/10.1145/3457607>
- [13] Ntoutsi, E., Fafalios, P., Gadiraju, U., Iosifidis, V., Nejdl, W., Vidal, M.-E., Ruggieri, S., Turini, F., Papadopoulos, S., Krasanakis, E., et al.: Bias in data-driven artificial intelligence systems—an introductory survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* **10**(3), 1356 (2020). <https://doi.org/10.1002/widm.1356>
- [14] Caton, S., Haas, C.: Fairness in machine learning: A survey. arXiv preprint arXiv:2010.04053 (2020)
- [15] Pessach, D., Shmueli, E.: Algorithmic fairness. arXiv preprint arXiv:2001.09784 (2020)
- [16] Holmes, D.E., Jain, L.C.: Introduction to bayesian networks. In: Innovations in Bayesian Networks, pp. 1–5 (2008). <https://doi.org/10.1145/3303772.3303791>
- [17] Husmeier, D., Dybowski, R., Roberts, S.: Probabilistic Modeling in Bioinformatics and Medical Informatics. Springer, ??? (2006)
- [18] Daniel, K.: Thinking, fast and slow (2017)
- [19] Chen, Y.-C., Wheeler, T.A., Kochenderfer, M.J.: Learning discrete bayesian networks from continuous data. *Journal of Artificial Intelligence Research* **59**, 103–132 (2017). <https://doi.org/10.1613/jair.5371>
- [20] Žliobaitė, I.: Measuring discrimination in algorithmic decision making. *Data Mining and Knowledge Discovery* **31**(4), 1060–1089 (2017). <https://doi.org/10.1007/s10618-017-0506-1>
- [21] Verma, S., Rubin, J.: Fairness definitions explained. In: 2018 IEEE/ACM International Workshop on Software Fairness (FairWare), pp. 1–7 (2018). <https://doi.org/10.23919/FAIRWARE.2018.8452913>
- [22] Dwork, C., Hardt, M., Pitassi, T., Reingold, O., Zemel, R.: Fairness through awareness. In: Proceedings of the 3rd Innovations in Theoretical Computer Science Conference, pp. 214–226 (2012). <https://doi.org/10.1145/2090236.2090255>
- [23] Romei, A., Ruggieri, S.: A multidisciplinary survey on discrimination analysis. *The Knowledge Engineering Review* **29**(5), 582–638 (2014). <https://doi.org/10.1017/S0269888913000039>
- [24] Hardt, M., Price, E., Srebro, N.: Equality of opportunity in supervised learning. In: Proceedings of the 30th International Conference on Neural Information Processing Systems, pp. 3323–3331 (2016)
- [25] Gardner, J., Brooks, C., Baker, R.: Evaluating the fairness of predictive student models through slicing analysis. In: Proceedings of the LAK19 Conference, pp. 225–234 (2019). <https://doi.org/10.1145/3303772.3303791>

- [26] Kohavi, R.: Scaling up the accuracy of naive-bayes classifiers: A decision-tree hybrid. In: KDD, vol. 96, pp. 202–207 (1996)
- [27] Krasanakis, E., Spyromitros-Xioufis, E., Papadopoulos, S., Kompatsiaris, Y.: 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 (2018)
- [28] Kamiran, F., Calders, T.: Data pre-processing techniques for classification without discrimination. Knowledge and Information Systems **33**(1), 1–33 (2012). <https://doi.org/10.1007/s10115-011-0463-8>
- [29] Kamiran, F., Žliobaitė, I., Calders, T.: Quantifying explainable discrimination and removing illegal discrimination in automated decision making. Knowledge and Information Systems **35**(3), 613–644 (2013). <https://doi.org/10.1007/s10115-012-0584-8>
- [30] Calders, T., Kamiran, F., Pechenizkiy, M.: Building classifiers with independency constraints. In: IEEE International Conference on Data Mining Workshops, 2009. ICDMW'09., pp. 13–18 (2009). <https://doi.org/10.1109/ICDMW.2009.83>
- [31] Žliobaite, I., Kamiran, F., Calders, T.: Handling conditional discrimination. In: 2011 IEEE 11th International Conference on Data Mining (ICDM), pp. 992–1001 (2011). <https://doi.org/10.1109/ICDM.2011.72>
- [32] Calders, T., Verwer, S.: Three naive bayes approaches for discrimination-free classification. Data Mining and Knowledge Discovery **21**(2), 277–292 (2010). <https://doi.org/10.1007/s10618-010-0190-x>
- [33] Luong, B.T., Ruggieri, S., Turini, F.: 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 (2011). <https://doi.org/10.1145/2020408.2020488>
- [34] Kamiran, F., Calders, T., Pechenizkiy, M.: Discrimination aware decision tree learning. In: 2010 IEEE 10th International Conference on Data Mining (ICDM), pp. 869–874 (2010). <https://doi.org/10.1109/ICDM.2010.50>
- [35] Iosifidis, V., Ntoutsi, E.: Dealing with bias via data augmentation in supervised learning scenarios. Jo Bates Paul D. Clough Robert Jäschke, 24
- [36] Calmon, F., Wei, D., Vinzamuri, B., Ramamurthy, K.N., Varshney, K.R.: Optimized pre-processing for discrimination prevention. In: Advances in Neural Information Processing Systems, pp. 3992–4001 (2017)
- [37] Feldman, M., Friedler, S.A., Moeller, J., Scheidegger, C., Venkatasubramanian, S.: Certifying and removing disparate impact. In: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 259–268 (2015). <https://doi.org/10.1145/2783258.2783311>
- [38] Hajian, S., Domingo-Ferrer, J.: A methodology for direct and indirect discrimination prevention in data mining. IEEE Transactions on Knowledge and Data Engineering **25**(7), 1445–1459 (2013). <https://doi.org/10.1109/TKDE.2012.72>
- [39] Zafar, M.B., Valera, I., Rodriguez, M.G., Gummadi, K.P.: Fairness Constraints: Mechanisms for Fair Classification. In: Singh, A., Zhu, J. (eds.) Proceedings of the 20th International Conference on Artificial Intelligence and Statistics. Proceedings of Machine Learning Research, vol. 54, pp. 962–970 (2017)
- [40] Žliobaite, I.: On the relation between accuracy and fairness in binary classification. In: The 2nd FATML Workshop at ICML'15 (2015)
- [41] Calders, T., Kamiran, F.: Classification with no discrimination by preferential sampling.

- In: Proceeding 19th Machine Learning Conference Belgium and the Netherlands (2010)
- [42] Feldman, M.: Computational fairness: Preventing machine-learned discrimination (2015)
- [43] Fish, B., Kun, J., Lelkes, A.D.: Fair boosting: a case study. In: Workshop on Fairness, Accountability, and Transparency in Machine Learning (2015)
- [44] Fish, B., Kun, J., Lelkes, A.D.: A confidence-based approach for balancing fairness and accuracy. In: Proceedings of the 2016 SIAM International Conference on Data Mining, pp. 144–152 (2016). <https://doi.org/10.1137/1.9781611974348.17>
- [45] Friedler, S.A., Scheidegger, C., Venkatasubramanian, S., Choudhary, S., Hamilton, E.P., Roth, D.: A comparative study of fairness-enhancing interventions in machine learning. In: Conference on Fairness, Accountability, and Transparency, pp. 329–338 (2019). <https://doi.org/10.1145/3287560.3287589>
- [46] Ristanoski, G., Liu, W., Bailey, J.: Discrimination aware classification for imbalanced datasets. In: Proceedings of the 22nd ACM International Conference on Information & Knowledge Management, pp. 1529–1532 (2013). <https://doi.org/10.1145/2505515.2507836>
- [47] Chakraborty, J., Peng, K., Menzies, T.: Making fair ml software using trustworthy explanation. In: 2020 35th IEEE/ACM International Conference on Automated Software Engineering (ASE), pp. 1229–1233 (2020)
- [48] Quadrianto, N., Sharmanska, V.: Recycling privileged learning and distribution matching for fairness (2017)
- [49] Xu, R., Cui, P., Kuang, K., Li, B., Zhou, L., Shen, Z., Cui, W.: Algorithmic decision making with conditional fairness. In: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 2125–2135 (2020). <https://doi.org/10.1145/3394486.3403263>
- [50] Zafar, M.B., Valera, I., Gomez-Rodriguez, M., Gummadi, K.P.: Fairness constraints: A flexible approach for fair classification. *J. Mach. Learn. Res.* **20**(75), 1–42 (2019)
- [51] Choi, Y., Farnadi, G., Babaki, B., Van den Broeck, G.: 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 (2020). <https://doi.org/10.1609/aaai.v34i06.6565>
- [52] Oneto, L., Doninini, M., Elders, A., Pontil, M.: Taking advantage of multitask learning for fair classification. In: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, And Society, pp. 227–237 (2019). <https://doi.org/10.1145/3306618.3314255>
- [53] Grari, V., Ruf, B., Lamprier, S., Detyniecki, M.: Fair adversarial gradient tree boosting. In: 2019 IEEE International Conference on Data Mining (ICDM), pp. 1060–1065 (2019). <https://doi.org/10.1109/ICDM.2019.00124>
- [54] L. Cardoso, R., Meira Jr, W., Almeida, V., J. Zaki, M.: 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 (2019). <https://doi.org/10.1145/3306618.3314262>
- [55] Agarwal, A., Beygelzimer, A., Dudík, M., Langford, J., Wallach, H.: A reductions approach to fair classification. In: International Conference on Machine Learning, pp. 60–69 (2018)
- [56] Backurs, A., Indyk, P., Onak, K., Schieber, B., Vakilian, A., Wagner, T.: Scalable fair clustering. In: International Conference on Machine Learning, pp. 405–413 (2019)

- [57] Hu, T., Iosifidis, V., Liao, W., Zhang, H., Yang, M.Y., Ntoutsi, E., Rosenhahn, B.: FairNN-conjoint learning of fair representations for fair decisions. In: International Conference on Discovery Science, pp. 581–595 (2020). https://doi.org/10.1007/978-3-030-61527-7_38
- [58] Chierichetti, F., Kumar, R., Lattanzi, S., Vassilvitskii, S.: Fair clustering through fairlets. In: Proceedings of the 31st International Conference on Neural Information Processing Systems, pp. 5036–5044 (2017)
- [59] Ziko, I.M., Yuan, J., Granger, E., Ayed, I.B.: Variational fair clustering. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, pp. 11202–11209 (2021)
- [60] Haeri, M.A., Zweig, K.A.: The crucial role of sensitive attributes in fair classification. In: 2020 IEEE Symposium Series on Computational Intelligence (SSCI), pp. 2993–3002 (2020). <https://doi.org/10.1109/SSCI47803.2020.9308585>
- [61] Berk, R., Heidari, H., Jabbari, S., Joseph, M., Kearns, M., Morgenstern, J., Neel, S., Roth, A.: A convex framework for fair regression. 4th Workshop on Fairness, Accountability, and Transparency in Machine Learning (FAT/ML 2017) (2017)
- [62] Esmaeili, S., Brubach, B., Tsepenekas, L., Dickerson, J.: Probabilistic fair clustering. Advances in Neural Information Processing Systems **33** (2020)
- [63] Deepak, P., Abraham, S.S.: Fair outlier detection. In: International Conference on Web Information Systems Engineering, pp. 447–462 (2020). https://doi.org/10.1007/978-3-030-62008-0_31
- [64] Mahabadi, S., Vakilian, A.: Individual fairness for k-clustering. In: International Conference on Machine Learning, pp. 6586–6596 (2020)
- [65] Anderson, N., Bera, S.K., Das, S., Liu, Y.: Distributional individual fairness in clustering. arXiv preprint arXiv:2006.12589 (2020)
- [66] Huang, L., Jiang, S.H.-C., Vishnoi, N.K.: Coresets for clustering with fairness constraints. In: Proceedings of the 33rd International Conference on Neural Information Processing Systems, pp. 7589–7600 (2019)
- [67] Kearns, M., Neel, S., Roth, A., Wu, Z.S.: An empirical study of rich subgroup fairness for machine learning. In: Proceedings of the Conference on Fairness, Accountability, and Transparency, pp. 100–109 (2019). <https://doi.org/10.1145/3287560.3287592>
- [68] Bechavod, Y., Ligett, K.: Penalizing unfairness in binary classification. arXiv preprint arXiv:1707.00044 (2017)
- [69] Ruoss, A., Balunovic, M., Fischer, M., Vechev, M.: Learning certified individually fair representations. Advances in Neural Information Processing Systems 33 pre-proceedings (NeurIPS 2020) **33**, 7584–7596 (2020)
- [70] Zhang, B.H., Lemoine, B., Mitchell, M.: Mitigating unwanted biases with adversarial learning. In: Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, And Society, pp. 335–340 (2018). <https://doi.org/10.1145/3278721.3278779>
- [71] Iosifidis, V., Ntoutsi, E.: Adafair: Cumulative fairness adaptive boosting. In: Proceedings of the 28th ACM International Conference on Information and Knowledge Management, pp. 781–790 (2019). <https://doi.org/10.1145/3357384.3357974>
- [72] Du, M., Yang, F., Zou, N., Hu, X.: Fairness in deep learning: A computational perspective. IEEE Intelligent Systems (2020). <https://doi.org/10.1109/MIS.2020.3000681>
- [73] Zhang, W., Ntoutsi, E.: FAHT: an adaptive fairness-aware decision tree classifier. In: Proceedings of the 28th International Joint Conference on Artificial Intelligence (IJCAI-19) (2019)
- [74] Galhotra, S., Saisubramanian, S., Zilberstein, S.: Learning to generate fair clusters from demonstrations. Proceedings of

- the 2021 AAAI/ACM Conference on AI, Ethics, and Society (2021). <https://doi.org/10.1145/3461702.3462558>
- [75] Abbasi, M., Bhaskara, A., Venkatasubramanian, S.: Fair clustering via equitable group representations. In: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, pp. 504–514 (2021). <https://doi.org/10.1145/3442188.3445913>
- [76] Hébert-Johnson, U., Kim, M., Reingold, O., Rothblum, G.: Multicalibration: Calibration for the (computationally-identifiable) masses. In: International Conference on Machine Learning, pp. 1939–1948 (2018)
- [77] Martinez, N., Bertran, M., Sapiro, G.: Minimax pareto fairness: A multi objective perspective. In: International Conference on Machine Learning, pp. 6755–6764 (2020)
- [78] Abraham, S.S., Sundaram, S.S., et al.: Fairness in clustering with multiple sensitive attributes. In: 23rd International Conference on Extending Database Technology (EDBT), pp. 287–298 (2020)
- [79] Dheeru, D., Karra Taniskidou, E.: UCI Machine Learning Repository (2017). <http://archive.ics.uci.edu/ml>
- [80] Iosifidis, V., Ntoutsi, E.: Fabboo - online fairness-aware learning under class imbalance. In: International Conference on Discovery Science, pp. 159–174 (2020). https://doi.org/10.1007/978-3-030-61527-7_11
- [81] Ruggieri, S., Pedreschi, D., Turini, F.: Data mining for discrimination discovery. ACM Transactions on Knowledge Discovery from Data (TKDD) **4**(2), 9 (2010). <https://doi.org/10.1145/1754428.1754432>
- [82] Pedreschi, D., Ruggieri, S., Turini, F.: Discrimination-aware data mining. In: Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 560–568 (2008). <https://doi.org/10.1145/1401890.1401959>
- [83] Pedreschi, D., Ruggieri, S., Turini, F.: Measuring discrimination in socially-sensitive decision records. In: Proceedings of the 2009 SIAM International Conference on Data Mining, pp. 581–592 (2009). <https://doi.org/10.1137/1.9781611972795.50>
- [84] Kamiran, F., Calders, T.: Classifying without discriminating. In: 2009 2nd International Conference on Computer, Control and Communication, pp. 1–6 (2009). <https://doi.org/10.1109/IC4.2009.4909197>
- [85] Zemel, R., Wu, Y., Swersky, K., Pitassi, T., Dwork, C.: Learning fair representations. In: International Conference on Machine Learning, pp. 325–333 (2013)
- [86] Mancuhan, K., Clifton, C.: Combating discrimination using bayesian networks. Artificial Intelligence and Law **22**(2), 211–238 (2014). <https://doi.org/10.1007/s10506-014-9156-4>
- [87] Ahn, Y., Lin, Y.-R.: Fairsight: Visual analytics for fairness in decision making. IEEE transactions on visualization and computer graphics **26**(1), 1086–1095 (2019). <https://doi.org/10.1109/TVCG.2019.2934262>
- [88] Van der Laan, P.: The 2001 census in the netherlands. In: Conference The Census of Population (2000)
- [89] Moro, S., Cortez, P., Rita, P.: A data-driven approach to predict the success of bank tele-marketing. Decision Support Systems **62**, 22–31 (2014). <https://doi.org/10.1016/j.dss.2014.03.001>
- [90] Bera, S.K., Chakrabarty, D., Flores, N.J., Negahbani, M.: Fair algorithms for clustering. Advances in Neural Information Processing Systems **33** (2020)
- [91] Chouldechova, A., G'Sell, M.: Fairer and more accurate, but for whom? arXiv preprint arXiv:1707.00046 (2017)
- [92] Zhang, Z., Neill, D.B.: Identifying significant predictive bias in classifiers. arXiv preprint arXiv:1611.08292 (2016)

- [93] Chouldechova, A.: Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data* **5**(2), 153–163 (2017). <https://doi.org/10.1089/big.2016.0047>
- [94] Zafar, M.B., Valera, I., Gomez Rodriguez, M., Gummadi, K.P.: 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 (2017). <https://doi.org/10.1145/3038912.3052660>
- [95] Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., Huq, A.: 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 (2017). <https://doi.org/10.1145/3097983.3098095>
- [96] Tu, R., Zhang, X., Liu, Y., Kjellström, H., Liu, M., Zhang, K., Zhang, C.: How do fair decisions fare in long-term qualification? In: Thirty-fourth Conference on Neural Information Processing Systems (2020)
- [97] Slack, D., Friedler, S.A., Givental, E.: Fairness warnings and fair-maml: learning fairly with minimal data. In: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, pp. 200–209 (2020). <https://doi.org/10.1145/3351095.3372839>
- [98] Grgić-Hlača, N., Zafar, M.B., Gummadi, K.P., Weller, A.: Beyond distributive fairness in algorithmic decision making: Feature selection for procedurally fair learning. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 32 (2018)
- [99] Lahoti, P., Gummadi, K.P., Weikum, G.: Operationalizing individual fairness with pairwise fair representations. *Proceedings of the VLDB Endowment* **13**(4), 506–518 (2019). <https://doi.org/10.14778/3372716.3372723>
- [100] Heidari, H., Ferrari, C., Gummadi, K.P., Krause, A.: 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 (2018)
- [101] Russell, C., Kusner, M.J., Loftus, J.R., Silva, R.: When worlds collide: integrating different counterfactual assumptions in fairness. *Advances in Neural Information Processing Systems* **30**. Pre-proceedings **30** (2017)
- [102] Kearns, M., Neel, S., Roth, A., Wu, Z.S.: Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In: International Conference on Machine Learning, pp. 2564–2572 (2018)
- [103] Narasimhan, H., Cotter, A., Gupta, M., Wang, S.: Pairwise fairness for ranking and regression. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, pp. 5248–5255 (2020). <https://doi.org/10.1609/aaai.v34i04.5970>
- [104] Sharifi-Malvajerdi, S., Kearns, M., Roth, A.: Average individual fairness: Algorithms, generalization and experiments. *Advances in Neural Information Processing Systems* **32**, 8242–8251 (2019)
- [105] Calders, T., Karim, A., Kamiran, F., Ali, W., Zhang, X.: Controlling attribute effect in linear regression. In: 2013 IEEE 13th International Conference on Data Mining (ICDM), pp. 71–80 (2013). <https://doi.org/10.1109/ICDM.2013.114>
- [106] Chzhen, E., Denis, C., Hebiri, M., Oneto, L., Pontil, M.: Fair regression via plug-in estimator and recalibration with statistical guarantees. *Advances in Neural Information Processing Systems* **33** (2020)
- [107] Strack, B., DeShazo, J.P., Gennings, C., Olmo, J.L., Ventura, S., Cios, K.J., Clore, J.N.: Impact of HbA1c measurement on hospital readmission rates: analysis of 70,000 clinical database patient records. *BioMed research international* **2014** (2014). <https://doi.org/10.1155/2014/2014>

- //doi.org/10.1155/2014/781670
- [108] Sattigeri, P., Hoffman, S.C., Chenthamarakshan, V., Varshney, K.R.: Fairness GAN. arXiv preprint arXiv:1805.09910 (2018)
- [109] Supreme Court of the United States: Ricci v. destefano. In: 557 U.S. 557, 174 (2009)
- [110] Ignatiev, A., Cooper, M.C., Siala, M., Hebrard, E., Marques-Silva, J.: Towards formal fairness in machine learning. In: International Conference on Principles and Practice of Constraint Programming, pp. 846–867 (2020). https://doi.org/10.1007/978-3-030-58475-7_49
- [111] Schelter, S., He, Y., Khilnani, J., Stoyanovich, J.: Fairprep: Promoting data to a first-class citizen in studies on fairness-enhancing interventions. In: EDBT (2020). <https://doi.org/10.5441/002/edbt.2020.41>
- [112] Valdivia, A., Sánchez-Monedero, J., Casillas, J.: How fair can we go in machine learning? Assessing the boundaries of accuracy and fairness. International Journal of Intelligent Systems **36**(4), 1619–1643 (2021). <https://doi.org/10.1002/int.22354>
- [113] Cortez, P., Silva, A.M.G.: Using data mining to predict secondary school student performance. Proceedings of 5th FUTURE BUSINESS TECHNOLOGY Conference (FUBUTEC 2008), 5–12 (2008)
- [114] Le Quy, T., Roy, A., Friege, G., Ntoutsi, E.: Fair-capacitated clustering. In: Proceedings of The 14th International Conference on Educational Data Mining (EDM21)., pp. 407–414 (2021)
- [115] Kuzilek, J., Hlosta, M., Zdrahal, Z.: Open university learning analytics dataset. Scientific data **4**, 170171 (2017). <https://doi.org/10.1038/sdata.2017.171>
- [116] Riazy, S., Simbeck, K.: Predictive algorithms in learning analytics and their fairness. DELFI 2019 (2019). https://doi.org/10.18420/delfi2019_305
- [117] Riazy, S., Simbeck, K., Schreck, V.: Fairness in learning analytics: Student at-risk prediction in virtual learning environments. In: CSEDU (1), pp. 15–25 (2020). <https://doi.org/10.5220/0009324100150025>
- [118] Wightman, L.F.: Lsac national longitudinal bar passage study. lsac research report series (1998)
- [119] Kusner, M., Loftus, J., Russell, C., Silva, R.: Counterfactual fairness. In: Proceedings of the 31st International Conference on Neural Information Processing Systems, pp. 4069–4079 (2017)
- [120] Lahoti, P., Beutel, A., Chen, J., Lee, K., Prost, F., Thain, N., Wang, X., Chi, E.: Fairness without demographics through adversarially reweighted learning. In: 34th Conference on Neural Information Processing Systems (2020)
- [121] Yang, F., Cisse, M., Koyejo, S.: Fairness with overlapping groups. arXiv preprint arXiv:2006.13485 (2020)
- [122] Hosmer Jr, D.W., Lemeshow, S., Sturdivant, R.X.: Applied Logistic Regression vol. 398. John Wiley & Sons, ??? (2013)
- [123] Zhao, J., Wang, T., Yatskar, M., Ordonez, V., Chang, K.-W.: 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 (2018). <https://doi.org/10.18653/v1/N18-2003>
- [124] Buolamwini, J., Gebru, T.: Gender shades: Intersectional accuracy disparities in commercial gender classification. In: Conference on Fairness, Accountability and Transparency, pp. 77–91 (2018)
- [125] Merler, M., Ratha, N., Feris, R.S., Smith, J.R.: Diversity in faces. arXiv preprint arXiv:1901.10436 (2019)