A survey on datasets for fairness-aware machine learning

Tai Le Quy*, Arjun Roy†, Vasileios Iosifidis‡, and Eirini Ntoutsi§

1L3S Research Center, Leibniz University Hannover, Germany
2Institute of Computer Science, Free University Berlin, Germany

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

Abstract

As decision-making increasingly relies on machine learning 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 between 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.

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

*Corresponding author. tai@l3s.de 0000-0001-8512-5854
†arjun.roy@fu-berlin.de 0000-0002-4279-9442
‡iosifidis@l3s.de 0000-0002-3005-4507
§eirini.ntoutsi@fu-berlin.de 0000-0001-5729-1003
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 (Mukerjee, Biswas, Deb, & Mathur, 2002), recruiting (Faliagka, Ramantas, Tsakalidis, & Tzimas, 2012), school admission (Moore, 1998), risk prediction (Yeh & Lien, 2009). There are many advantages of using algorithmic decision-making as computers can 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 (Angwin, Larson, Mattu, & Kirchner, 2016), a software used by the U.S. courts to assess the risk of recidivism; in particular, it has been found that black defendants were predicted with a higher risk of recidivism than their actual risk compared to white defendants. Another example refers to search algorithms in job search websites; it has been found that such algorithms exhibit gender-bias as they display higher-paying jobs to male applicants compared to female ones (Simonite, 2015; Datta, Tschantz, & Datta, 2015).

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

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

---

1https://qz.com/1034972/the-data-that-changed-the-direction-of-ai-research-and-possibly-the-world/
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 of 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 dataset collection strategy and introduce Bayesian networks as a tool for learning the structure from the data. In addition, we provide a summary of fairness measures we will use for the quantitative evaluation.

2.1 Strategy for collecting datasets

To identify the relevant datasets, we use Google Scholar\(^2\) with “fairness datasets” as the primary query term along with other terms like “bias”, “discrimination”, “public” to narrow down the search. After identifying the related datasets, we use Google Scholar to find the related papers which satisfy the following conditions: 1) The public dataset is used in the experiments, and 2) The learning tasks, i.e., classification, clustering, are related to fairness problems. To restrict the investigation of the related work, we consider only important works as assessed by the number of citations, quality of publication venue, i.e., published in ranked conferences, journals. We consider datasets that have been used in at least three fairness-related papers. Datasets that are not publicly available via some known repository like the UCI machine learning repository\(^3\), Kaggle\(^4\), etc., are not taken into consideration.

2.2 Bayesian network

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

\(^2\)https://scholar.google.com/
\(^3\)https://archive.ics.uci.edu
\(^4\)https://www.kaggle.com
the form of conditional independence relations. Formally:

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

Learning the structure of a BN from the dataset $\mathcal{D}$ is an optimization problem (Husmeier, Dybowski, & Roberts, 2006), namely to learn an optimal BN model $\mathcal{M}^*$ which maximizes the likelihood of generating $\mathcal{D}$. A set of parameters of any BN model $\mathcal{M}$, denoted by $\hat{\mathcal{M}}$, is the set of edges $E$ which represents the conditional independence relationship between the attribute set $\mathcal{V}$. Moreover, between the possible models $\mathcal{M}$, the less complex one, i.e., the one with the least $\hat{\mathcal{M}}$, 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 $P(D \mid \mathcal{M})$. However, we aim to investigate the factors (protected/non-protected attributes) that determine the class attribute’s prediction probability. Therefore, we also employ a constraint on the class attribute to be a leaf node in our learning objective. Formally the problem is defined as:

$$\max_{\mathcal{M}^*} \{ P(D \mid \mathcal{M}) - \gamma \hat{\mathcal{M}} \}$$
subject to $y \in \mathcal{L}$  \hspace{1cm} (2)$$

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

A high conditional probability often refers to a strong correlation (Daniel, 2017). Attribute $X_i$ is strongly correlated with $X_j$ if there exists a direct edge between $X_i$ and $X_j$, for any pair of attributes $X_i, X_j \in \mathcal{X}$. Intuitively, the correlation is comparatively weaker with ancestors that are not immediate parents, i.e., indirect edges. In addition, the attributes which do not have any incoming or outgoing edge (direct/indirect connection) with $X_i$, the correlation between them will be negligible. As a consequence, if we find any direct/indirect edge from any protected 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 (Chen, Wheeler, & Kochenderfer, 2017). Therefore, we need to discretize the continuous numeric data attributes into meaningful categorical attributes to keep the complexity of learning the BN model in a polynomial time. We describe the discretization procedure for each dataset in Section 3.

### 2.3 Fairness metrics

Measuring bias in ML models comprises the first step to bias elimination. Fairness depends on context; thus, a large variety of fairness measures exists. Only in the computer science research area, more than 20 measures of fairness have been introduced thus far (Žliobaitė, 2017; Verma & Rubin, 2018). 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 $S \in \{s, \pi\}$ with $s$ and $\pi$ denoting the protected and non-protected groups, respectively. We use the notation $s_+$ ($s_-$), $\pi_+$ ($\pi_-$) to denote the protected and non-protected groups for the positive (negative, respectively) class.

### 2.3.1 Statistical parity
Statistical parity (shortly Parity) (Dwork, Hardt, Pitassi, Reingold, & Zemel, 2012; Romei & Ruggieri, 2014) reports on the percentage difference between two populations w.r.t. the positive class. It is formally defined as follows:

$$Parity = \frac{|\{x \in D \mid S = \pi, y = +\}| - |\{x \in D \mid S = s, y = +\}|}{|\{x \in D \mid S = s\}|}$$  \hspace{1cm} (3)

The value domain is: $Parity \in [-1, 1]$, 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 (Hardt, Price, & Srebro, 2016) (shortly Eq.Odds) measures the difference in prediction errors between the protected and non-protected groups. Let $\Delta FPR$ and $\Delta FNR$ denote the differences in false positive rates and false negative rates, respectively between the protected and non-protected groups, defined as follows:

$$\Delta FPR = P(y \neq \hat{y} \mid S = s_\pi) - P(y \neq \hat{y} \mid S = s_s)$$
$$\Delta FNR = P(y \neq \hat{y} \mid S = s_\pi) - P(y \neq \hat{y} \mid S = s_s)$$ \hspace{1cm} (4)

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

Equalized odds aims at minimizing both $\Delta FPR$ and $\Delta FNR$ and is defined as:

$$Eq.Odds = |\Delta FPR| + |\Delta FNR|$$ \hspace{1cm} (5)

The value domain is: $Eq.Odds \in [0, 2]$, 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 (Gardner, Brooks, & Baker, 2019). It is based on the Receiver Operating Characteristics (ROC) curve. ABROCA measures the divergence between the protected ($ROC_s$) and non-protected group ($ROC_\pi$) 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:

$$\int_0^1 |ROC_s(t) - ROC_\pi(t)| \, dt$$ \hspace{1cm} (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 social datasets (Section 3.3) and educational datasets (Section 3.4). A summary of the statistics of the different datasets is provided in Table 1.

### 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                | 15/6/2                       | Binary | Finance | 1:3.03            | Sex, race, age      | Income       | 1994-1995        | The US             |
| KDD Cen.Income | 299,285     | 284,556               | 32/7/3                       | Binary | Finance | 1:15.30           | Sex, race           | Income       | 1994-1995        | The US             |
| German credit  | 1,000       | 1,000                 | 13/1/7                       | Binary | Finance | 2:31.1            | Sex, age            | Credit score  | 1973-1975        | Germany           |
| Dutch census   | 189,725     | 60,420                | 10/3/0                       | Binary | Finance | 1:1.1             | Sex                | Occupation    | 2001              | The Netherlands    |
| Bank marketing | 45,211      | 45,211                | 6/4/7                        | Binary | Finance | 1:7.95            | Age, marital status| Deposit payment| 2000-2011        | Portugal           |
| Credit card    | 30,000      | 30,000                | 6/2/14                       | Binary | Finance | 1:3.52            | Sex, marriage, education| Default payment| 2005              | Taiwan            |
| COMPAS recid   | 7,234       | 6,172                 | 31/9/24                      | Binary | Criminology | 1:20 | Race | Two years recid | 2013-2014 | The US |
| COMPAS recid   | 4,743       | 4,020                 | 31/9/24                      | Binary | Criminology | 1:2.17 | Race | Two years recid | 2013-2014 | The US |
| CommunityCrime | 1,064       | 1,064                 | 4/0/123                      | Multi  | Crime | 1:13.13           | Black                | Violent crime rate | 1993 | The US |
| Diabetes       | 101,786     | 101,786               | 31/10                        | Binary | Physician | 1:1.13 | Gender | Readmitted in 30 days | 1999-2000 | The US |
| Risi           | 118         | 118                   | 0/1/3                        | Binary | Society | 1:1.11 | Race | Promotion        | 2003 | The US |
| Student-Mathematics | 649 | 649 | 4/11/38 | Binary | Education | 1:0.04 | Sex, age | Final grade | 2000-2008 | Portugal |
| Student-Portuguese | 649 | 649 | 4/11/38 | Binary | Education | 1:0.49 | Sex, age | Final grade | 2000-2008 | Portugal |
| OLSAD         | 52,502      | 23,662                | 7/2/3                        | Multi  | Education | - | Gender | Outcome | 2013-2014 | England |
| Law School     | 20,700      | 20,700                | 3/1/6                        | Binary | Education | 8:0.7 | Male, Race | Pass the bar exam | 1991 | The US |

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

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

3.1 Financial datasets

3.1.1 Adult dataset

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

**Dataset characteristics:** The dataset consists of 48,842 instances, each described via 15 attributes, of which 6 are numerical, 7 are categorical and 2 are binary attributes. An overview of attribute characteristics is shown in Table 2. We discard the attribute fnlwgt (final weight) as the suggestions of related work (B. H. Zhang, Lemoine, & Mitchell, 2018; Kamiran & Calders, 2012; Calders, Kamiran, & Pechenizkiy, 2009; Calders & Kamiran, 2010). Missing values exist in 3,620 (7.41%) records. Many researchers remove the instances containing missing values (Zafar, Valera, Rogriguez, & Gummadi, 2017; Iosifidis & Ntoutsi, n.d., 2019; Choi, Farnadi, Babaki, & Van den Broeck, 2020) in their experiments; other researches consider the whole dataset or do...
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                 |
| finwgt           | Numerical     | [13,492 - 1,490,400]        | 0               | The final weight                                 |
| education        | Categorical   | 16                          | 0               | The highest level of education                   |
| educational-num | Numerical     | 1 - 16                      | 0               | The highest level of education achieved in numerical form |
| marital-status   | Categorical   | 7                           | 0               | The marital status                               |
| occupation       | Categorical   | 14                          | 2,809           | The general type of occupation                   |
| relationship     | Categorical   | 6                           | 0               | Represents what this individual is relative to others |
| race             | Categorical   | 5                           | 0               | Race                                             |
| sex              | Binary        | Male, Female                | 9               | The biological sex of the individual             |
| capital-gain     | Numerical     | [0 - 99,999]                | 0               | The capital gains for an individual              |
| capital-loss     | Numerical     | [0 - 4,350]                 | 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   | US, non-US                  | 657             | The country of origin for an individual          |
| income           | Binary        | (≤50K, >50K)                | 0               | Whether or not an individual makes more than $50,000 annually |

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 literature:

- **sex = \{male, female\}**: the dataset is dominated by male instances. The ratio of male:female is 32,650:16,192 (66.9%:33.1%).

- **race = \{white, black, asian-pac-islander, amer-indian-eskimo, other\}**: Typically, race is used as a binary attribute in the related work (Luong, Ruggieri, & Turini, 2011; Chakraborty, Peng, & Menzies, 2020; Zafar, Valera, Rodriguez, & Gummadi, 2017): race = \{white, non-white\}. The dataset is dominated by white people, the white:non-white ratio is 38,903:6,319 (86%:14%). In our analysis we also encode race as a binary attribute.

- **age = [17-90]**. Typically, age is used as a categorical attribute in the related work. In our analysis we also discretize age as (Zafar, Valera, Rodriguez, & Gummadi, 2017): age = \{25-60, <25 or >60\}. The dataset is dominated by the [25–60] years old group, the ratio is 35,066:10,156 (77.5%:22.5%).

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

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

---

7Please 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 (Hébert-Johnson, Kim, Reingold, & Rothblum, 2018; Martinez, Bertran, & Sapiro, 2020; Abraham, Sundaram, et al., 2020).
As demonstrated in Figure 1, there is a direct dependency between income and education as well as between sex and education. Therefore, we explore in more detail the distribution of the population w.r.t. education, income and sex in Figure 2a. As expected, highly educated people have a high income. However, in the high education segment and for the high-income class, the number of males is at least 5 times higher than that of females showing an under-representation of high education women in the high-income class. Based on the dependence of hours per week attribute on sex, we plot the weekly working hours w.r.t income and sex (Figure 2b). The number of males who work more than 40 hours per week is more than 7 times higher than that of the females.

Figure 2: Adult: relationships of education, weekly working hours, education and income attributes
Interestingly, there are many outgoing edges from the relationship and age attributes in the BN. We show the distribution of sex in each class based on the age (x-axis) and the relationship status (y-axis) in Figure 3. A first observation is that a great amount of young (less than 25 years old) or old (more than 60 years old) people do not receive more than 50K. “Unmarried” people have an income higher than 50K when they are older than 45 years, while people in the “Own-child” group can have a high income when they are young. In general, there are more males than females for almost all relationship statuses for the high-income group.

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

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

3.1.2 KDD Census-Income dataset

The KDD Census-Income\(^8\) dataset (Dheeru & Karra Taniskidou, 2017) was collected from Current Population Surveys implemented by the U.S. Census Bureau from 1994 to 1995. The dataset has been considered in numerous related works (Appendix A). The prediction task is to decide if a person receives more than 50 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 dataset’s authors (Dheeru & Karra Taniskidou, 2017) are: “the goal field was drawn from the total person income field rather than the adjusted gross income and may, therefore, behave differently than the original adult goal field”.

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

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

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

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

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

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

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

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

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

As mentioned, race could also be considered as the protected attribute. Based on the data,
the income of *non-white* people is significantly different from the income of the *white* group. Only 3.2% of the *non-white* group have an income above 50K, compared to 6.7% for the *white* group. Furthermore, since *age* has a conditional dependence on *marital-status* attribute, we investigate the relationship between these attributes, the protected attribute *sex* and the class label *income* in Figure 6. As shown in this figure, males comprise the majority of the high-income group, especially for certain population segments like the *Married-civilian spouse present* segment where the number of males is 5 times higher than that of females. Interestingly, the number of widows is 1.7 times higher than the number of widowers in terms of high income. Regarding the *age* effect, most people have a high income when they are over 40 years old. With respect to the protected attributes, there is no edge between *race* and *sex*, which suggests the
researchers should perform their fairness-aware models on both these protected attributes.

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

3.1.3 German credit dataset
The German credit\textsuperscript{10} dataset (Dheeru & Karra Taniskidou, 2017) consists of samples of bank account holders. The dataset is used for risk assessment prediction, i.e., to determine whether it is risky to grant credit to a person or not. Customers are classified into two classes: \{Good, Bad\}. The dataset is frequently employed in fairness-aware learning research (Appendix A).

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

Table 4: German credit: attributes characteristics

| Attributes                      | Type         | Values | #Missing values | Description                                                                 |
|---------------------------------|--------------|--------|-----------------|-----------------------------------------------------------------------------|
| checking-account                | Categorical  | 4      | 0               | The status of existing checking account                                      |
| duration                        | Numerical    | [4 - 72]| 0               | The duration of the credit (month)                                          |
| credit-history                  | Categorical  | 5      | 0               | The credit history                                                          |
| purpose                         | Categorical  | 10     | 0               | Purpose                                                                     |
| credit-amount                   | Numerical    | [250 - 4,242]| 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-provide-maintenance-for | Numerical | [1 - 2]| 0 | Number of people being liable to provide maintenance for                    |
| telephone                       | Binary       | {Yes, No}| 0 | Telephone number                                                           |
| foreign-worker                  | Binary       | {Yes, No}| 0 | Is the individual a foreign worker?                                        |
| class-label                     | Binary       | {Good, Bad}| 0 | Class                                                                       |

Protected attributes: In all studies, sex is considered as the protected attribute. Age can also be considered as the protected attribute after binarization into \{young, old\} by age thresholding at 25 (Kamiran & Calders, 2009; Friedler et al., 2019).

\textsuperscript{10}https://archive.ics.uci.edu/ml/datasets/statlog+(german+ credit+data)
• sex = \{ male, female \}. The dataset is dominated by male instances, the ratio of male:female is 690:310 (69%:31%). The percentage of women identified as bad customers is 35.2% while that of men is only 27.7%.

• age = \{ \leq 25, > 25 \}: The dataset is dominated by people older than 25 years, the ratio is 810:190 (81%:19%). We discover that there is a discrimination on the age of customers. There are 42.1% of young people are recognized as bad customers while this proportion in old 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 Figure 7; class-label is set as a leaf node.

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

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

(a) Distribution of class label on status of checking account
(b) Relationship between credit amount and installment rate

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

3.1.4 Dutch census dataset

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

**Dataset characteristics:** The dataset includes 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.

| Attributes         | Type      | Values                        | #Missing values | Description                                                                 |
|--------------------|-----------|-------------------------------|----------------|-----------------------------------------------------------------------------|
| sex                | Binary    | {Male, Female}                | 0              | The biological sex of the person                                           |
| age                | Categorical | 22                          | 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                                                    |

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

**Protected attributes:** In the related work, they consider attribute sex = {male, female} as the protected attribute, male:female ratio is 30,147:30,273 (49.9%:50.1%).
Bayesian network: We use all attributes in the dataset to generate the Bayesian network. As illustrated in Figure 9, the leaf node *occupation* is conditionally dependent on *economic status*, *education level* and *sex* attributes. In fact, 62.6% of males (18,860 out of 30,147) have a high-level occupation, while this proportion on females group is only 32.7%. In addition, people with high education are doing prestigious jobs and vice versa, as depicted in Figure 10. For example, 89.5% of people having *tertiary* level are working in high-level jobs while around 80% of people with *lower secondary* degrees are doing low-level work. Interestingly, *age* has a direct effect on many attributes.

![Bayesian Network Illustration](image)

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

### 3.1.5 Bank marketing dataset

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

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

| Attributes   | Type     | Values                  | #Missing values | #Missing values |
|--------------|----------|-------------------------|----------------|----------------|
| age          | Numerical| [18 - 95]               | 0              |                |
| job          | Categorical | 12                      | 0              |                |
| marital      | Categorical | 3                      | 0              |                |
| education    | Categorical | 4                      | 0              |                |
| default      | Binary    | [Yes, No]               | 0              |                |
| balance      | Numerical | -8,019 - 102,127        | 0              |                |
| housing      | Binary    | [Yes, No]               | 0              |                |
| loan         | Binary    | [Yes, No]               | 0              |                |
| contact      | Categorical | 3                      | 0              |                |
| day          | Numerical | [1 - 31]                | 0              |                |
| month        | Categorical | 12                      | 0              |                |
| duration     | Numerical | [0 - 4,918]             | 0              |                |
| campaign     | Numerical | [1 - 63]                | 0              |                |
| pdays        | Numerical | [1 - 87]                | 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? |

**Protected attributes:** In the literature, *marital-status* can be considered as the protected attribute (Backurs et al., 2019; Hu et al., 2020; Chierichetti, Kumar, Lattanzi, & Vassilvitskii, 2017; Ziko, Yuan, Granger, & Ayed, 2021; Bera, Chakrabarty, Flores, & Negahbani, 2019). Besides, in several studies (Krasanakis, Spyromitros-Xioufis, Papadopoulos, & Kompatsiaris, 2019).

---

11[https://archive.ics.uci.edu/ml/datasets/Bank+Marketing](https://archive.ics.uci.edu/ml/datasets/Bank+Marketing)
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.

- $\text{age} = \{25-60, <25 \text{ or } >60\}$: 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%).

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

Figure 11 visualizes the Bayesian network of the Bank marketing dataset. The class label $y$, as illustrated in Figure 11, is conditionally dependent on $\text{poutcome}$, $\text{month}$ and $\text{duration}$ attributes. An insight about the relationship between the last contact $\text{duration}$ and class label $y$ is described in Figure 12. The ratio of clients who will make a deposit subscription is proportional to the duration of the last contact. When the talk is taken place in less than 2 minutes, 98.5% of people will not make the deposit subscription. However, if a marketing staff can maintain the talk with customers over 10 minutes, 48.4% of customers will say ‘Yes’. Interestingly, in the Bayesian network, both protected attributes $\text{age}$ and $\text{marital}$ have no effect on the class label $y$. However, the protected attributes are connected together by an in-direct edge, which could be a reason for a similar accuracy of fairness-aware models of the related work (Hu et al., 2020) and (Krasanakis et al., 2018).
3.1.6 Credit card clients dataset
The credit card clients dataset (Yeh & Lien, 2009) investigated the customers’ default payments in Taiwan in October 2005. The goal is to predict whether a customer will face the default situation in the next month or not. The data have been used for default payment prediction in several studies (Appendix A).

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

| 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)      |
| pay0       | Categorical   | 11                              | 0               | The repayment status in September 2005 |
| pay2       | Categorical   | 11                              | 0               | The repayment status in August 2005    |
| pay3       | Categorical   | 11                              | 0               | The repayment status in July 2005      |
| pay4       | Categorical   | 11                              | 0               | The repayment status in May 2005       |
| pay5       | Categorical   | 10                              | 0               | The repayment status in May 2005       |
| pay6       | Categorical   | 10                              | 0               | The repayment status in May 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 - 893,586]            | 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 - 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           |

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

- sex = \{male, female\}: the dataset is dominated by females, the ratio of male:female is 11,888:18,112 (39.6%:60.4%).

12https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients
marriage = \{\text{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 = \{\text{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 = \{\leq 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) = \{\leq 50,000, 50,001-200,000, >200,000\} (corresponding to the low, medium, high levels); history of the past payments pay_0, ..., pay_6 = \{\text{pay duly, 1-3 months, > 3 months}\}. The Bayesian network is presented in Figure 13.

![Bayesian Network Diagram](image)

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

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

**Summary of the financial datasets:** In general, the financial datasets are the most diverse in terms of both spatial and temporal criteria compared to datasets in other domains. There are many protected attributes are chosen in the related work. In which, sex is the most prevalent protected attribute, followed by race, age, marriage and education. The interesting point is the protected attributes are often related to each other (it can be a strong relationship or a weak
relationship. This is a benefit for fairness-aware studies by maintaining fairness on a protected attribute may have a positive effect on fairness on other protected attributes.

### 3.2 Criminological datasets

#### 3.2.1 COMPAS dataset

The COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) (Angwin et al., 2016) is a recent dataset, compared to the rest of the datasets in our work, which was released by ProPublica\(^\text{13}\) in 2016 based on the Broward County data (collected from Jan 2013 to Dec 2014). Defendant’s answers to the COMPAS screening survey are used to generate the recidivism risk scores. The data have been used for crime recidivism risk prediction by a plethora of works (Appendix A). Risk of recidivism (denoted as COMPAS recid.) and Risk of violent recidivism (denoted as COMPAS viol. recid.) subsets are most widely used in the literature. The former has a classification task to predict if an individual is rearrested within two years after the first arrest. The latter predicts if an individual is rearrested for a violent crime within two years.

**Dataset characteristics:** COMPAS recid. and COMPAS viol. recid. datasets contain 7,214 and 4,743 samples, respectively. Each defendant is described by 52 attributes\(^\text{14}\) (31 categorical, 6 binary, 14 numerical and a null attribute), as shown in Table 8 and Table 17 (Appendix B). Missing data is a common phenomenon in both subsets. There are 6,395 rows (88.6%) containing missing values in the COMPAS recid. subset while this number in the COMPAS viol. revid. subset is 3,748 instances (79%). Based on (Angwin et al., 2016), we clean the dataset by removing the missing data, such as "violent recid" = NULL or the change date of a crime (attribute days.b.screening_arrest) was not within 30 days when he or she was arrested. The cleaned datasets used in our analysis contain 6,172 (COMPAS recid.) and 4,020 (COMPAS viol. recid.) records.

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

**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 on which individual was brought into custody), and several ID-related attributes. A new

\(^{13}\)https://www.propublica.org/datastore/dataset/compas-recidivism-risk-score-data-and-analysis

\(^{14}\)Table 8 describes attributes used in the Bayesian network and data analysis
Table 8: COMPAS recid: attributes characteristics

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

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

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

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

Figure 18: COMPAS: Relationship between recidivism and priors count

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

### 3.2.2 Communities and Crime dataset

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

#### Table 9: Communities and Crime: attributes characteristics

| Attributes         | Type      | Values          | #Missing values | Description                                                                 |
|--------------------|-----------|-----------------|-----------------|-----------------------------------------------------------------------------|
| pctpctblack        | Numerical | [0.0 - 1.0]     | 0               | The percentage of population that is African American                        |
| pctWInvInc         | Numerical | [0.0 - 1.0]     | 0               | The percentage of households with investment, rent income in 1989            |
| pctWPubAssast      | Numerical | [0.0 - 1.0]     | 0               | The percentage of households with public assistance income in 1989          |
| NumUnderPov        | Numerical | [0.0 - 1.0]     | 0               | The number of people under the poverty level                                |
| PctPopUndrPov      | Numerical | [0.0 - 1.0]     | 0               | The percentage of people under the poverty level                            |
| PctUnemployed      | Numerical | [0.0 - 1.0]     | 0               | The percentage of people 16 and over, in the labor force, and unemployed    |
| MalePctDiv         | Numerical | [0.0 - 1.0]     | 0               | The percentage of males who are divorced                                    |
| FemalePctDiv       | Numerical | [0.0 - 1.0]     | 0               | The percentage of females who are divorced                                   |
| PctDiv             | Numerical | [0.0 - 1.0]     | 0               | The percentage of population who are divorced                               |
| PctPerFam          | Numerical | [0.0 - 1.0]     | 0               | The mean number of people per family                                        |
| PctKids2Par        | Numerical | [0.0 - 1.0]     | 0               | The percentage of kids in family housing with two parents                  |
| PctYoungKids2Par   | Numerical | [0.0 - 1.0]     | 0               | The percentage of kids 4 and under in two parent households                |
| PctTeen2Par        | Numerical | [0.0 - 1.0]     | 0               | The percentage of kids age 12-17 in two parent households                  |
| NumMilleg          | 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                               |
| PctPersOwnOcc      | Numerical | [0.0 - 1.0]     | 0               | The percentage of people in owner occupied households                      |
| HousVacant         | Numerical | [0.0 - 1.0]     | 0               | The number of vacant households                                            |
| PctHousOwnOcc      | Numerical | [0.0 - 1.0]     | 0               | The percentage of households owner occupied                                |
| PctVacantBoarded   | Numerical | [0.0 - 1.0]     | 0               | The percentage of vacant housing that is boarded up                        |
| NumInShelters      | Numerical | [0.0 - 1.0]     | 0               | The number of people in homeless shelters                                  |
| NuminStreet        | Numerical | [0.0 - 1.0]     | 0               | The number of homeless people counted in the street                        |
| ViolentCrimesPerPop| Numerical | [0.0 - 1.0]     | 0               | The total number of violent crimes per 100,000 population                   |

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

There is a very high proportion (84%) of missing values in 25 attributes, as demonstrated in Table 19. Based on the suggestions from the literature, 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%).

**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. Hence, we use the median value 0.5 as a

---

15 http://archive.ics.uci.edu/ml/datasets/communities+and +crime

16 Table 9 contains attributes used in the Bayesian network
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 Figure 20. In which, the percentage of kids born to never married \(PctIlleg\) and the percentage of kids in family housing with two parents \(PctKids2Par\) have a direct impact on the class label and the race. Looking into the dataset, we discover that 92.4% of the communities are dominated by “black” people, where the percentage of kids in family housing with two parents less than 50%), while only 55.6% of the communities are dominated by “black” people, where the percentage of kids in family housing with two parents greater than 50%.

**Summary of the criminological datasets:** In summary, the criminological datasets were collected recently and were only surveyed in the US. Race was the only protected attribute considered by the studies in ensuring the fairness of ML models. The bias w.r.t race was seen in the data, which could be challenging for the models. Furthermore, the datasets contain a lot of attributes, hence, it requires a careful selection of attributes for the models.
3.3 Healthcare and social datasets

3.3.1 Diabetes dataset

The diabetes dataset (Strack et al., 2014) 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 (Appendix A).

Dataset characteristics: The dataset contains 101,766 patients described by 50 attributes (10 numerical, 7 binary and 33 categorical). Characteristics of all attributes are summarized in Table 10 and Table 20 (in Appendix B). The attributes encounter_id and patient_nbr should not be considered in the learning tasks since they are the ID of the patients. Typically, weight, payer_code, medical_specialty attributes are removed because they 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.

| Attributes          | Type          | Values | #Missing values | Description                                                                 |
|---------------------|---------------|--------|----------------|-----------------------------------------------------------------------------|
| race                | Categorical   | 6      | 2,773          | Race (Caucasian, Asian, African American, Hispanic, and other)              |
| gender              | Categorical   | 3      | 0              | Gender (male, female, and unknown/invalid)                                  |
| age                 | Categorical   | 10     | 0              | Grouped in 10-year intervals                                               |
| time_in_hospital    | Numerical     | [1 - 14] | 0           | The number of days between admission and discharge                         |
| num_procedures      | Numerical     | [0 - 6] | 0              | The number of procedures (other than lab tests) performed during the encounter|
| num_medications     | Numerical     | [1 - 81] | 0            | The number of distinct generic names administered during the encounter      |
| number_outpatient   | Numerical     | [0 - 42] | 0              | The number of outpatient visits of the patient in the year preceding the encounter|
| number_emergency    | Numerical     | [0 - 76] | 0              | The number of emergency visits of the patient in the year preceding the encounter|
| number_inpatient    | Numerical     | [0 - 21] | 0              | The number of inpatient visits of the patient in the year preceding the encounter|
| A1Cresult           | Categorical   | 4      | 0              | Whether the drug was prescribed or there was a change in the dosage         |
| metformin           | Categorical   | 4      | 0              | Whether the drug was prescribed or there was a change in the dosage         |
| chlorpropamide      | Categorical   | 4      | 0              | Whether the drug was prescribed or there was a change in the dosage         |
| glipizide           | Categorical   | 4      | 0              | Whether the drug was prescribed or there was a change in the dosage         |
| rosiglitazone       | Categorical   | 4      | 0              | Whether the drug was prescribed or there was a change in the dosage         |
| acarbose            | Categorical   | 4      | 0              | Whether the drug was prescribed or there was a change in the dosage         |
| miglitol            | Categorical   | 4      | 0              | Whether the drug was prescribed or there was a change in the dosage         |
| diabetesMed         | Binary        | (Yes, No) | 0           | Was there any diabetic medication prescribed?                               |
| readmitted          | Categorical   | 3      | 0              | The number of days to inpatient readmission                                |

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%.

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 Figure 21.

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

17https://archive.ics.uci.edu/ml/datasets/diabetes+130-us+hospitals+for+years+1999-2008
18Table 10 describes attributes used in the Bayesian network
3.3.2 Ricci dataset

The Ricci\textsuperscript{19} dataset was generated by the Ricci v.DeStefano case (Supreme Court of the United States, 2009) in which they investigated the results of a promotion exam within a fire department in Nov 2003 and Dec 2003. Although it is a relatively small dataset, it has been employed for fairness-aware classification tasks in many studies (Appendix A). The classification task is to predict whether an individual obtains a promotion based on the exam results.

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

| 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      |

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

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

\textsuperscript{19}https://www.key2stats.com/data-set/view/690
It is easy to observe that the combined grade (attribute combine) has a direct effect on the class label (promoted). Figure 23 illustrates the relationship between the combined grade and the promotion status. 100% of people whose combined oral and written exams are equal to or above 70 are promoted. Besides, as depicted in Figure 24, the number of promotions are granted for “white” people is higher than that for “non-white” people. The opposite trend is true in the group of candidates with no promotion.
### 3.4 Educational datasets

#### 3.4.1 Student performance dataset

The student performance dataset (Cortez & Silva, 2008) described students' achievement in the secondary education of two Portuguese schools in 2005 - 2006 with two distinct subjects: Mathematics and Portuguese. The regression task is to predict the final year grade of the students. It is investigated in several researches (Appendix A) 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 simply the classification problem, we create a class label based on attribute G3, \( \text{class} = \{\text{Low, High}\} \) corresponds to \( G3 = \{<10, \geq 10\} \).

![Table 12: Student performance: attributes characteristics](https://archive.ics.uci.edu/ml/datasets/student+performance)

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

- \( \text{sex} = \{\text{male, female}\} \): the dataset is dominated by female students. The ratios of

---

https://archive.ics.uci.edu/ml/datasets/student-performance
male:female are 208:187 (52.7%:47.3%) and 383:266 (59%:41%) for the Mathematics subject and Portuguese subject, respectively.

• age = {<18, ≥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, ≥10}; G2 = {<10, ≥10}. Due to the computation of the Bayesian network generator and the correlation coefficient with the class label (with a threshold of 0.02), we select 26 variables for the network. The Bayesian networks of the dataset on Portuguese and Mathematics subjects are visualized in Figure 25 and Figure 26, respectively.

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

The class label attribute is conditionally dependent on the grade G2 in both subsets (Mathematics and Portuguese subjects). This is explained by a very high correlation coefficient (above 90%) between G2 and G3 variables. In addition, we investigate the distribution of the final grade G3 on sex because the attribute sex has an indirect relationship with the class label. Figure 27 reveals that the male students tend to receive high scores in the Portuguese subject, while the scores of Math are relatively evenly distributed across both sexes.
Figure 26: Student performance - Mathematics subject: Bayesian network (class label: class, protected attributes: age, sex)

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

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

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%).

21https://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                                         |
| age_band          | Categorical | 10                                            | 1111           | The index of multiple deprivation (IMD) band of the place where the student lived |
| imd_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)                     |

**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 Figure 28. 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.

![Figure 28: OULAD: Bayesian network (class label: final_result, protected attributes: gender)](image)

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

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

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

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

Dataset characteristics: The dataset contains information of 20,798 students characterized by 12 attributes (3 categorical, 3 binary and 6 numerical attributes). An overview of all attributes is depicted in Table 14. 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 (89.5:11).

| 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                                                  |
| upga       | Numerical  | [1.5 - 4.0]             | 0               | The student’s undergraduate GPA                                           |
| zgpa       | Numerical  | [-3.35 - 3.48]          | 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                                    |
| race       | Categorical| 6                      | 0               | Race                                                                       |
| pass_bar   | Binary     | \{0, 1\}                | 0               | Whether the student passed the bar exam on the first try                   |

Table 14: Law school: attributes characteristics

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

• male \(\{1, 0\}\). “Male” is the majority group. The ratio of male (1):female (0) is 11,675:9,123 (56.1%:43.9%).

• race \(\{\text{white}, \text{black}, \text{Hispanic}, \text{Asian}, \text{other}\}\). As introduced in the related work, we encode race \(\{\text{white}, \text{non-white}\}\) based on the original attribute. Non-white is the minority group with the ratio white:non-white is 17,491:3,307 (84.9:16%).

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

It is easy to observe that the 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 Figure 31.
Summary of the educational datasets: The educational datasets were collected in many countries around the world. There are several protected attributes are considered in the related work. Gender is the popular protected attribute, followed by age and race. The common learning tasks in the educational data is the predictions of students’ outcome or their grades. Therefore, many machine learning tasks are applied to the datasets, such as classification, regression, or clustering.

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 (Hosmer Jr, Lemeshow, & Sturdivant, 2013), 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 Figure 32 (in which, prot and non-prot stand for protected, non-protected, respectively), we report the performance of the predictive model on the following measures.

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

Figure 32: The confusion matrix

- **Accuracy**
  \[
  ACC = \frac{TP + TN}{TP + TN + FP + FN} \tag{7}
  \]

- **Balanced accuracy**
  \[
  BA = \frac{1}{2} \left( \frac{TP}{TP + FN} + \frac{TN}{TN + FP} \right) \tag{8}
  \]

- **True positive rate (TPR) on protected group**
  \[
  TPR_{prot} = \frac{TP_{prot}}{TP_{prot} + FN_{prot}} \tag{9}
  \]

- **TPR on non-protected group**
  \[
  TPR_{non-prot} = \frac{TP_{non-prot}}{TP_{non-prot} + FN_{non-prot}} \tag{10}
  \]

- **True negative rate (TNR) on protected group**
  \[
  TNR_{prot} = \frac{TN_{prot}}{TN_{prot} + FP_{prot}} \tag{11}
  \]
• TNR on non-protected group

\[
TNR_{\text{non-prot}} = \frac{TN_{\text{non-prot}}}{TN_{\text{non-prot}} + FP_{\text{non-prot}}}
\]  
(12)

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

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.

Table 15: Performance of logistic regression model on datasets

| Dataset                  | Protected attribute | Group distribution (%) | Accuracy | Balanced accuracy | Statistical Parity | Equalized odds | ABROCA | TPR prot. | TPR non-prot. | TNR prot. | TNR non-prot. |
|--------------------------|---------------------|------------------------|----------|------------------|-------------------|----------------|--------|-----------|--------------|-----------|---------------|
| Adult                    | Sex                 | [5.7, 28.8, 21.1, 44.4] | 0.7884   | 0.6249           | 0.1998           | 0.0381         | 0.0318 | 0.3194    | 0.3007       | 0.9521    | 0.9426        |
| KDD Census-Income        | Sex                 | [1.3, 50.7, 4.8, 43.2] | 0.9474   | 0.6031           | -0.0752          | 0.0403         | 0.0074 | 0.1825    | 0.2195       | 0.9961    | 0.9628        |
| German credit            | Sex                 | [20.1, 10.9, 49.9, 19.1] | 0.9667   | 0.5713           | 0.0748           | 0.1334         | 0.1228 | 0.9831    | 0.8533       | 0.2759    | 0.2429        |
| Dutch census             | Sex                 | [16.4, 33.1, 32.1, 18.7] | 0.8149   | 0.8138           | -0.2985          | 0.3746         | 0.0202 | 0.6684    | 0.8382       | 0.9219    | 0.6871        |
| Bank marketing           | Marital             | [5.8, 34.2, 61.4, 51.4] | 0.8855   | 0.5720           | -0.0366          | 0.1611         | 0.0225 | 0.1527    | 0.1726       | 0.9649    | 0.9737        |
| Credit card clients      | Sex                 | [12.5, 47.8, 61.3, 30.0] | 0.7822   | 0.5               | -0.0254          | 0.0            | 0.0220 | 0.0       | 1.0          | 1.0       | 1.0           |
| COMPAS recid.            | Race                | [31.5, 28.7, 15.5, 24.3] | 0.6414   | 0.6099           | 0.1322           | 0.6452         | 0.0675 | 0.5996    | 0.2538       | 0.6791    | 0.9307        |
| COMPAS viol. recid.      | Race                | [32.0, 44.9, 52.8, 36.8] | 0.9432   | 0.5541           | 0.0913           | 0.2195         | 0.0564 | 0.1826    | 0.0          | 0.9606    | 0.9975        |
| Communities & Crime      | Black               | [5.8, 46.3, 3.0, 27.6]  | 0.9683   | 0.7011           | 0.4314           | 0.4507         | 0.0311 | 0.312     | 1.0          | 1.0       | 1.0           |
| Diabetes                 | Gender              | [11.1, 34.1, 15.1, 41.7] | 0.7504   | 0.5               | -0.0057          | 0.0            | 0.0169 | 0.0       | 1.0          | 1.0       | 1.0           |
| Ricci                    | Race                | [12.7, 20.7, 34.7, 22.9] | 1.0      | 1.0               | 0.3029           | 0.0            | 0.0   | 1.0       | 1.0          | 1.0       | 1.0           |
| Student - Mathematics    | Sex                 | [33.7, 10.0, 33.4, 11.9] | 0.9412   | 0.9160           | -0.0665          | 0.1616         | 0.0177 | 0.9354    | 0.9762       | 0.9630    | 0.8421        |
| Student - Portuguese     | Sex                 | [9.3, 7.7, 13.3, 7.7]   | 0.9292   | 0.8447           | 0.0575           | 0.0400         | 0.0271 | 0.9633    | 0.959        | 0.7143    | 0.656         |
| OULAD                    | Gender              | [52.1, 14.2, 35.9, 17.6] | 0.6751   | 0.5               | -0.0252          | 0.0            | 0.0088 | 1.0       | 1.0          | 0.0       | 0.0           |
| Law School               | Race                | [11.5, 4.4, 77.5, 6.6]  | 0.9072   | 0.6260           | 0.1983           | 0.5403         | 0.0325 | 0.9100    | 0.9955       | 0.5251    | 0.1063        |

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

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

5 Conclusion and outlook

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

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

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

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

Fifth, collecting new datasets is always a requirement of data scientists. The surveyed datasets were all collected quite a long time in the past with an average age of about 20 years. The oldest dataset was obtained 48 years ago, while the newest dataset was identified from 7 years ago. Of course, the newer the data, the more up-to-date with the trends of the modern society, so the analysis and application of fairness-aware algorithms on the new datasets will reflect the manifestations of the social behaviors more realistic. On the other hand, the old datasets are of reference value in comparing and contrasting the movement and variation of fairness in the same or different domains. The datasets are collected in the US and European countries where the data protection laws are in place. However, the general policies on data quality or collection still need to be studied and proposed (Ntoutsi et al., 2020).
To conclude, fairness-aware ML has attracted many recently in various domains from criminology, healthcare, finance to education. This paper reviews the most popular datasets used in fairness-aware ML researches. We explore the relationship of the variables as well as analyze their correlation concerning protected attributes and the class label. We believe our analysis will be the basis for developing frameworks or simulation environments to evaluate fairness-aware algorithms. In another aspect, an excellent understanding of well-known datasets can also inspire researchers to develop synthetic data generators because finding a suitable real-world dataset is never a simple task.

**Funding Information**

Ministry of Science and Education of Lower Saxony, German, project ID: 51410078

**Acknowledgements**

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

**References**

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

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

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

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

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

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

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

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

Berk, R., Heidari, H., Jabbari, S., Joseph, M., Kearns, M., Morgenstern, J., . . . Roth, A. (2017). A convex framework for fair regression. *4th Workshop on Fairness, Accountability, and Transparency in Machine Learning (FAT/ML 2017).*
Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency* (pp. 77–91).

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

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

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

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

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

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

Chen, Y.-C., Wheeler, T. A., & Kochenderfer, M. J. (2017). Learning discrete bayesian networks from continuous data. *Journal of Artificial Intelligence Research, 59*, 103–132. doi: https://doi.org/10.1613/jair.5371

Chhabra, A., Masalkovaite, K., & Mohapatra, P. (2021). An overview of fairness in clustering. *IEEE Access*. doi: https://doi.org/10.1109/ACCESS.2021.3114099

Chierichetti, F., Kumar, R., Lattanzi, S., & Vassilvitskii, S. (2017). Fair clustering through fairlets. In *Proceedings of the 31st International Conference on Neural Information Processing Systems* (pp. 5036–5044).

Choi, Y., Farnadi, G., Babaki, B., & Van den Broeck, G. (2020). Learning fair naive bayes classifiers by discovering and eliminating discrimination patterns. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 34, pp. 10077–10084). doi: https://doi.org/10.1609/aaai.v34i06.6565

Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data, 5*(2), 153–163. doi: https://doi.org/10.1089/big.2016.0047

Chzhen, E., Denis, C., Hebiri, M., Oneto, L., & Pontil, M. (2020). Fair regression via plug-in estimator and recalibration with statistical guarantees. *Advances in Neural Information Processing Systems, 33*.

Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., & Huq, A. (2017). Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 797–806). doi: https://doi.org/10.1145/3097983.3098095

Cortez, P., & Silva, A. M. G. (2008). Using data mining to predict secondary school student performance. *Proceedings of 5th FUTURE BUSINESS TECHNOLOGY CONFERENCE (FUBUTEC 2008)*, 5–12.

Daniel, K. (2017). *Thinking, fast and slow*. Farrar Straus Giroux.

Datta, A., Fredrikson, M., Ko, G., Mardziel, P., & Sen, S. (2017). Use privacy in data-driven systems: Theory and experiments with machine learnt programs. In *Proceedings of the*
Hajian, S., & Domingo-Ferrer, J. (2013). A methodology for direct and indirect discrimination prevention in data mining. *IEEE Transactions on Knowledge and Data Engineering, 25*(7), 1445–1459. doi: https://doi.org/10.1109/TKDE.2012.72

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

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

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

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

Hosmer Jr, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression* (Vol. 398). John Wiley & Sons.

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

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

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

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

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

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

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

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

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

Kamiran, F., Calders, T., & Pechenizkiy, M. (2010). Discrimination aware decision tree learning. In *2010 IEEE 10th International Conference on Data Mining (ICDM)* (pp. 869–874). doi: https://doi.org/10.1109/ICDM.2010.50
Kamiran, F., Žliobaitė, I., & Calders, T. (2013). Quantifying explainable discrimination and removing illegal discrimination in automated decision making. *Knowledge and Information Systems, 35*(3), 613–644. doi: https://doi.org/10.1007/s10115-012-0584-8

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

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

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

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

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

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

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

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

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

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

Luong, B. T., Ruggieri, S., & Turini, F. (2011). k-nn as an implementation of situation testing for discrimination discovery and prevention. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge discovery and Data mining* (pp. 502–510). doi: https://doi.org/10.1145/2020408.2020488

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

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

Mancuhan, K., & Clifton, C. (2014). Combating discrimination using bayesian networks. *Artificial Intelligence and Law, 22*(2), 211–238. doi: https://doi.org/10.1007/s10456-014-9156-4

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

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias
and fairness in machine learning. ACM Computing Surveys (CSUR), 54(6), 1–35. doi: https://doi.org/10.1145/3457607
Moore, J. S. (1998). An expert system approach to graduate school admission decisions and academic performance prediction. Omega, 26(5), 659–670. doi: https://doi.org/10.1016/S0305-0483(98)00008-5
Moro, S., Cortez, P., & Rita, P. (2014). A data-driven approach to predict the success of bank telemarketing. Decision Support Systems, 62, 22–31. doi: https://doi.org/10.1016/j.dss.2014.03.001
Mukerjee, A., Biswas, R., Deb, K., & Mathur, A. P. (2002). Multi-objective evolutionary algorithms for the risk–return trade-off in bank loan management. International Transactions in Operational research, 9(5), 583–597. doi: https://doi.org/10.1111/1475-3995.00375
Narasimhan, H., Cotter, A., Gupta, M., & Wang, S. (2020). Pairwise fairness for ranking and regression. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 34, pp. 5248–5255). doi: https://doi.org/10.1609/aaai.v34i04.5970
Ntoutsi, E., Fafalios, P., Gadiraju, U., Iosifidis, V., Nejdl, W., Vidal, M.-E., . . . others (2020). Bias in data-driven artificial intelligence systems—an introductory survey. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 10(3), e1356. doi: https://doi.org/10.1002/widm.1356
Oneto, L., Donini, M., Elders, A., & Pontil, M. (2019). Taking advantage of multitask learning for fair classification. In Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society (pp. 227–237). doi: https://doi.org/10.1145/3306618.3314255
Pedreschi, D., Ruggieri, S., & Turini, F. (2009). Measuring discrimination in socially-sensitive decision records. In Proceedings of the 2009 SIAM International Conference on Data Mining (pp. 581–592). doi: https://doi.org/10.1137/1.9781611972795.50
Pedreschi, D., Ruggieri, S., & Turini, F. (2008). Discrimination-aware data mining. In Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 560–568). doi: https://doi.org/10.1145/1401890.1401959
Pitoura, E., Stefanidis, K., & Koutrika, G. (2021). Fairness in rankings and recommendations: An overview. The VLDB Journal. doi: https://doi.org/10.1007/s00778-021-00697-y
Quadrianto, N., & Sharmanska, V. (2017). Recycling privileged learning and distribution matching for fairness.
Riazy, S., & Simbeck, K. (2019). Predictive algorithms in learning analytics and their fairness. DELFI 2019. doi: https://dx.doi.org/10.18420/delfi2019_305
Riazy, S., Simbeck, K., & Schreck, V. (2020). Fairness in learning analytics: Student at-risk prediction in virtual learning environments. In CSEDU (1) (pp. 15–25). doi: https://doi.org/10.5220/0009324100150025
Ristanoski, G., Liu, W., & Bailey, J. (2013). Discrimination aware classification for imbalanced datasets. In Proceedings of the 22nd ACM International Conference on Information & Knowledge Management (pp. 1529–1532). doi: https://doi.org/10.1145/2505515.2507836
Romei, A., & Ruggieri, S. (2014). A multidisciplinary survey on discrimination analysis. The Knowledge Engineering Review, 29(5), 582–638. doi: https://doi.org/10.1017/S0269889914000039
Ruggieri, S., Pedreschi, D., & Turini, F. (2010). Data mining for discrimination discovery. ACM Transactions on Knowledge Discovery from Data (TKDD), 4(2), 9. doi: https://doi.org/10.1145/1754428.1754432
Ruoss, A., Balunovic, M., Fischer, M., & Vechev, M. T. (2020). Learning certified individually fair representations. In Proceedings of the 33rd International Conference on Neural Information
Russell, C., Kusner, M. J., Loftus, J. R., & Silva, R. (2017). When worlds collide: integrating different counterfactual assumptions in fairness. Advances in Neural Information Processing Systems 30. Pre-proceedings, 30.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Yang, F., Cisse, M., & Koyejo, S. (2020). Fairness with overlapping groups: a probabilistic perspective. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, & H. Lin (Eds.), Advances in Neural Information Processing Systems (Vol. 33, pp. 4067–4078). Curran Associates, Inc.
Yeh, I.-C., & Lien, C.-h. (2009). The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. *Expert Systems with Applications, 36*(2), 2473–2480. doi: https://doi.org/10.1016/j.eswa.2007.12.020

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

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

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

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

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

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

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

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

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

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

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

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

1. Adult dataset

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

2. KDD Census-Income dataset

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

3. German credit dataset

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

4. Dutch census dataset

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

5. Bank marketing dataset

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

6. Credit card clients dataset

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

7. COMPAS dataset
(Krasanakis et al., 2018; Calmon et al., 2017; Chouldechova, 2017; Zafar, Valera, Gomez Rodriguez, & Gummadi, 2017; Corbett-Davies, Pierson, Feller, Goel, & Huq, 2017; Friedler et al., 2019; Tu et al., 2020; Quadrianto & Sharmanska, 2017; Xu et al., 2020; Zafar et al., 2019; Slack, Friedler, & Givental, 2020; Choi et al., 2020; Grgić-Hlača, Zafar, Gummadi, & Weller, 2018; Oneto et al., 2019; L. Cardoso et al., 2019; Agarwal et al., 2018; Haeri & Zweig, 2020; Laloti, Gummadi, & Weikum, 2019; Heidari, Ferrari, Gummadi, & Krause, 2018; Russell et al., 2017; Berk et al., 2017; Ruoss et al., 2020; Du et al., 2020).

8. Communities & Crime dataset
(Kamiran & Calders, 2012; Kamiran et al., 2013, 2010; Lahoti et al., 2019; Kearns, Neel, Roth, & Wu, 2018; Narasimhan, Cotter, Gupta, & Wang, 2020; Slack et al., 2020; Sharifi-Malvajerdi, Kearns, & Roth, 2019; Heidari et al., 2018; Calders, Karim, Kamiran, Ali, & Zhang, 2013; Chzhen et al., 2020; Kearns et al., 2019; Berk et al., 2017; Ruoss et al., 2020; Galhotra et al., 2021).

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

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

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

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

13. Law School dataset
(Chzhen et al., 2020; Kearns et al., 2019; Kusner et al., 2017; Russell et al., 2017; Lahoti et al., 2020; Bechavod & Ligett, 2017; Berk et al., 2017; Yang et al., 2020; Ruoss et al., 2020).
B Datasets’ characteristics

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

| Attributes                      | Type       | Values | #Missing values | Description                                                                 |
|---------------------------------|------------|--------|-----------------|-----------------------------------------------------------------------------|
| enroll-in-edu-inst-last-wk      | Categorical| 3      | 0               | An individual enrolled in an educational institute last week?               |
| major-industry                 | Categorical| 24     | 0               | The major industry code                                                     |
| major-occupation                | Categorical| 15     | 0               | The major occupation code                                                   |
| hispanic-origin                 | Categorical| 9      | 1,279           | The Hispanic origin                                                         |
| reason-union                   | Categorical| 6      | 0               | Member of a labor union                                                     |
| region-previous                 | Categorical| 6      | 0               | The region of previous residence                                            |
| state-previous                  | Categorical| 50     | 1036            | The state of previous residence                                             |
| migration-code-change-in-msa    | Categorical| 10     | 149,642         | Migration code-change in MSA                                               |
| migration-code-change-in-reg    | Categorical| 9      | 149,642         | Migration code-change in region                                             |
| migration-code-move-within-reg | Categorical| 10     | 149,642         | Migration code-move within region                                           |
| live-hour-1-year-ago            | Categorical| 3      | 0               | Live in this house 1 year ago                                               |
| 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                                          |
| fill-questionnaire              | Categorical| 3      | 0               | Fill the questionnaire for veteran’s admin                                  |

Table 17: COMPAS recid: attributes characteristics (continued)

| Attributes                      | Type       | Values | #Missing values | Description                                                                 |
|---------------------------------|------------|--------|-----------------|-----------------------------------------------------------------------------|
| name                            | Categorical| 7,156  | 0               | First and last name of the defendant                                        |
| first                           | Categorical| 2,800  | 0               | First name                                                                 |
| last                            | Categorical| 3,950  | 0               | Last name                                                                   |
| date                            | Categorical| 6,800  | 0               | The date on which the decile score was given                                |
| case_number                     | Categorical| 7,192  | 22              | The case number for original crime                                          |
| offense_date                    | Categorical| 927    | 1,159           | The date of follow-up crime                                                 |
| arrest_date                     | Categorical| 580    | 6,077           | The arrest date for original crime                                          |
| c_days_from_compas              | Numerical   | [0 - 9,485] | 22           | Between the COMPAS screening and the original crime offense date            |
| c_charge_desc                   | Categorical| 437    | 29              | Description of charge for original crime                                    |
| ic_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] | 4,988       | Between the follow-up crime and the arrest date (days)                      |
| r_offense_date                  | Categorical| 1,076  | 3,743           | The date of follow-up crime                                                 |
| r_charge_desc                   | Categorical| 340    | 3,801           | Description of charge for follow-up crime                                   |
| jail_in                         | Categorical| 972    | 4,988           | The jail entry date for follow-up crime                                     |
| jail_out                        | Categorical| 938    | 4,988           | The jail exit date for follow-up crime                                      |
| violent_recid                   | Binary      | {0, 1} | 7,214           | The binary indicator of violent follow-up crime                             |
| violent_in_custody             | Categorical| 819    | 6,395           | The case number for violent follow-up crime                                 |
| violent_charge_degree           | Categorical| 9      | 6,395           | Charge degree for violent follow-up crime                                   |
| violent_offense_date            | Categorical| 570    | 6,395           | The date of offense for violent follow-up crime                             |
| violent_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                             |
| screening_date                  | Categorical| 690    | 0               | Repeat column of decile score                                              |
| 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                             |
| event                           | 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                      |
| start                           | Numerical   | [0 - 937] | 0            | No information                                                             |
| end                             | Numerical   | [0 - 1,186] | 0            | No information                                                             |

47
| Attributes | Type     | Values | #Missing values | Description |
|------------|----------|--------|-----------------|-------------|
| state      | Categorical | 46     | 0              | The US state (by number) |
| county     | Categorical | 109    | 1174           | The numeric code for county |
| communityname | Categorical | 1,828  | 0              | The community name |
| fold       | Numerical   | [1 - 10] | 0              | The fold number for non-random 10 fold cross validation |
| population | Numerical   | [0 - 1] | 0              | The population for community |
| householdsize | Numerical  | [0 - 1] | 0              | The mean people per household |
| ractOctWhite | Numerical  | [0 - 1] | 0              | The percentage of population that is Caucasian |
| ractPctAsian | Numerical  | [0 - 1] | 0              | The percentage of population that is of Asian heritage |
| ractPctHisp | Numerical   | [0 - 1] | 0              | The percentage of population that is of Hispanic heritage |
| agePct1219 | Numerical   | [0 - 1] | 0              | The percentage of population that is 12-19 in age |
| agePct2029 | Numerical   | [0 - 1] | 0              | The percentage of population that is 12-29 in age |
| agePct3064 | Numerical   | [0 - 1] | 0              | The percentage of population that is 16-24 in age |
| agePct65+  | Numerical   | [0 - 1] | 0              | The percentage of population that is 65 and over in age |
| numbUrb   | Numerical   | [0 - 1] | 0              | The number of people living in areas classified as urban |
| pctUrban  | Numerical   | [0 - 1] | 0              | The percentage of people living in areas classified as urban |
| medianlncome | Numerical   | [0 - 1] | 0              | The median household income |
| pctWWage  | Numerical   | [0 - 1] | 0              | The percentage of households with wage or salary income in 1989 |
| pctWWFamSelf | Numerical  | [0 - 1] | 0              | The percentage of households with farm or self employment income in 1989 |
| pctWWSec   | Numerical   | [0 - 1] | 0              | The percentage of households with social security income in 1989 |
| pctWRetire | Numerical   | [0 - 1] | 0              | The percentage of households with retirement income in 1989 |
| medFamInc | Numerical   | [0 - 1] | 0              | The median family income |
| perCapInc | Numerical   | [0 - 1] | 0              | Per capita income (national income divided by population size) |
| whitePerCap | Numerical  | [0 - 1] | 0              | Per capita income for Caucasians |
| blackPerCap | Numerical   | [0 - 1] | 0              | Per capita income for African Americans |
| indianPerCap | Numerical   | [0 - 1] | 0              | Per capita income for native Americans |
| AsianPerCap | Numerical   | [0 - 1] | 0              | Per capita income for people with Asian heritage |
| otherPerCap | Numerical   | [0 - 1] | 0              | Per capita income for people with ‘other’ heritage |
| HispPerCap | Numerical   | [0 - 1] | 0              | Per capita income for people with Hispanic heritage |
| PctLess9thGrad | Numerical | [0 - 1] | 0              | The percentage of people 25 and over with less than a 9th grade education |
| PctNoHighGrad | Numerical | [0 - 1] | 0              | The percentage of people 25 and over that are not high school graduates |
| PctBarMore | Numerical | [0 - 1] | 0              | The percentage of people 25 and over with a bachelors degree or higher education |
| PctUnemployed | Numerical | [0 - 1] | 0              | The percentage of people 16 and over, in the labor force, and unemployed |
| PctEmploy | Numerical | [0 - 1] | 0              | The percentage of people 16 and over who are employed |
| PctEmpManu | Numerical | [0 - 1] | 0              | The percentage of people 16 and over who are employed in manufacturing |
| PctEmpProfServ | Numerical | [0 - 1] | 0              | The percentage of people 16 and over who are employed in professional services |
| PctOccPmgProf | Numerical | [0 - 1] | 0              | The percentage of people 16 and over who are employed in management |
| MalePctNavMar | Numerical | [0 - 1] | 0              | The percentage of males who have never married |
| PctYearsHHSocOcc | Numerical | [0 - 1] | 0              | The percentage of all occupied households that are large (6 or more) |
| PctHouseLettRd | Numerical | [0 - 1] | 0              | The percentage of all occupied households that are large (6 or more people) |
| PctPerOccupHos | Numerical | [0 - 1] | 0              | The mean persons per household |
| PctPerOwnOccHous | Numerical | [0 - 1] | 0              | The mean persons per owner occupied household |
| PctPentRentOccHous | Numerical | [0 - 1] | 0              | The mean persons per rental household |
| PctPentHousDens | Numerical | [0 - 1] | 0              | The percentage of persons in dense housing (more than 1 person per room) |
| PctHouseLessBR | Numerical | [0 - 1] | 0              | The percentage of housing units with less than 3 bedrooms |
| MaxNumBR | Numerical | [0 - 1] | 0              | The median number of bedrooms |
| PctHouseOccup | Numerical | [0 - 1] | 0              | The percentage of housing occupied |
| PctVacMore6Mos | Numerical | [0 - 1] | 0              | The percentage of vacant housing that has been vacant more than 6 months |
| MedYrHousBuilt | Numerical | [0 - 1] | 0              | The median year housing units built |
| PctHousNoPhone | Numerical | [0 - 1] | 0              | The percentage of occupied housing units without phone (in 1990) |
| PctWDFullPlumb | Numerical | [0 - 1] | 0              | The percentage of housing without complete plumbing facilities |
| OwnOccLowQuart | Numerical | [0 - 1] | 0              | Owner-occupied housing - lower quartile value |
| OwnOccHighQuart | Numerical | [0 - 1] | 0              | Owner-occupied housing - upper quartile value |
| RentLowQ | Numerical | [0 - 1] | 0              | Rental housing - lower quartile rent |
| RentMedian | Numerical | [0 - 1] | 0              | Rental housing - median rent |
### Table 19: Communities and Crime: attributes characteristics (continued)

| Attributes                       | Type     | Values          | #Missing values | Description                                      |
|----------------------------------|----------|-----------------|-----------------|--------------------------------------------------|
| RentHighQ                        | Numerical| [0.0 - 1.0]      | 0               | Rental housing - upper quartile rent              |
| MedRent                          | Numerical| [0.0 - 1.0]      | 0               | The median gross rent                            |
| MedRentPctHouseInc               | Numerical| [0.0 - 1.0]      | 0               | The median gross rent as a percentage of household income |
| MedOwnCostPctPcmkInc             | Numerical| [0.0 - 1.0]      | 0               | The median owners cost (without a mortgage) as a percentage of household income |
| PctForeignBom                    | Numerical| [0.0 - 1.0]      | 0               | The percentage of people foreign born            |
| PctBomSameState                  | 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) |
| LemasSwmFT                       | Numerical| [0.0 - 1.0]      | 1,675           | The number of sworn full-time police officers    |
| LemasSwmFTPerPop                 | Numerical| [0.0 - 1.0]      | 1,675           | The number of sworn full-time police officers in field operations |
| LemasSwmTFieldOps                | Numerical| [0.0 - 1.0]      | 1,675           | The number of sworn full-time police officers in field operations per 100,000 population |
| LemasSwmTFieldPerPop             | 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                    |
| LemasTostReqPerPop               | Numerical| [0.0 - 1.0]      | 1,675           | The total requests for police per 100,000 population |
| PolcReqPerOffice                 | Numerical| [0.0 - 1.0]      | 1,675           | The total requests for police per police officer |
| PolcPctPop                       | 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 |
| OfficeAssignDrugUnits            | Numerical| [0.0 - 1.0]      | 1,675           | The number of officers assigned to special drug units |
| NumKindsDrugGiev                 | Numerical| [0.0 - 1.0]      | 1,675           | The number of different kinds of drugs seized   |
| PolcAVPtWorkOut                  | Numerical| [0.0 - 1.0]      | 1,675           | Police average overtime worked                  |
| LandArea                         | Numerical| [0.0 - 1.0]      | 0               | Land area in square miles                        |
| PctUsePubTrans                   | Numerical| [0.0 - 1.0]      | 0               | The percentage of people using public transit for commuting |
| PolcCars                         | Numerical| [0.0 - 1.0]      | 1,675           | The number of police cars                        |
| PolcOpenBudg                     | Numerical| [0.0 - 1.0]      | 1,675           | Police operating budget                         |
| LemasPctPolicOnPat               | 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                              |
| LemasPctOfficDrugIn             | Numerical| [0.0 - 1.0]      | 1,675           | The percentage of officers assigned to drug units |
| PolcBudyPerPop                   | Numerical| [0.0 - 1.0]      | 1,675           | Police operating budget per population          |

### Table 20: Diabetes: attributes characteristics (continued)

| Attributes                        | Type     | Values          | #Missing values | Description                                      |
|-----------------------------------|----------|-----------------|-----------------|--------------------------------------------------|
| encounter_id                      | Numerical| [12,522 - 443,897,222] | 0 | Encounter’s unique identifier                     |
| patient_nbr                       | Numerical| [139 - 189,502,619] | 0 | Patient’s unique identifier                       |
| weight                            | Categorical| [10 - 98,569]      | 0 | Weight (pounds)                                   |
| admission_type_id                 | Categorical| [8 - 17]          | 0 | The admission type (emergency, urgent, etc.)      |
| discharge_disposition_id          | Categorical| [26 - 49,949]     | 0 | Discharge disposition (discharged to home, expired, etc.) |
| admission_source_id               | Categorical| [17 - 1]          | 0 | The admission source (physician referral, emergency room, etc.) |
| payer_code                        | Categorical| [18 - 40,256]     | 0 | Payer code (Medicare, self-pay, etc.)             |
| medical_specialty                 | Categorical| [73 - 749]        | 0 | The specialty of the admitting physician          |
| num_lab_procedures                | Numerical| [1 - 132]        | 0 | The number of lab tests performed during the encounter |
| diag_1                            | Categorical| [717 - 749]       | 21 | The primary diagnosis                             |
| diag_2                            | Categorical| [717 - 749]       | 21 | Secondary diagnosis                              |
| diag_3                            | Categorical| [717 - 749]       | 21 | Additional secondary diagnosis                    |
| number_diagnoses                  | Numerical| [1 - 16]         | 0 | The number of diagnoses entered to the system     |
| max glide level                   | Categorical| [4 - 7]           | 0 | The range of the results or if the test was not taken |
| repglinide                        | Categorical| [4 - 7]           | 0 | Whether the drug was prescribed or there was a change in the dosage |
| nataglinide                       | Categorical| [4 - 7]           | 0 | Whether the drug was prescribed or there was a change in the dosage |
| glimepiride                       | Categorical| [4 - 7]           | 0 | Whether the drug was prescribed or there was a change in the dosage |
| acetoheaximide                    | Categorical| [4 - 7]           | 0 | Whether the drug was prescribed or there was a change in the dosage |
| glyburide                         | Categorical| [4 - 7]           | 0 | Whether the drug was prescribed or there was a change in the dosage |
| tolbutamide                       | Categorical| [4 - 7]           | 0 | Whether the drug was prescribed or there was a change in the dosage |
| pioglitazone                      | Categorical| [4 - 7]           | 0 | Whether the drug was prescribed or there was a change in the dosage |
| voglirazone                       | Categorical| [4 - 7]           | 0 | Whether the drug was prescribed or there was a change in the dosage |
| examiclde                         | Categorical| [4 - 7]           | 0 | Whether the drug was prescribed or there was a change in the dosage |
| sitagliptin                       | Categorical| [4 - 7]           | 0 | Whether the drug was prescribed or there was a change in the dosage |
| insulin                           | Categorical| [4 - 7]           | 0 | Whether the drug was prescribed or there was a change in the dosage |
| glyburide-metformin               | Categorical| [4 - 7]           | 0 | Whether the drug was prescribed or there was a change in the dosage |
| glipizide-metformin               | Categorical| [4 - 7]           | 0 | Whether the drug was prescribed or there was a change in the dosage |
| glimepiride-pioglitazone          | Categorical| [4 - 7]           | 0 | Whether the drug was prescribed or there was a change in the dosage |
| metformin-rosiglitazone           | Categorical| [4 - 7]           | 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?      |