# Privileged and Unprivileged Groups: An Empirical Study on the Impact of the Age Attribute on Fairness

Max Hort
University College London
London, United Kingdom
max.hort.19@ucl.ac.uk

Federica Sarro
University College London
London, United Kingdom
f.sarro@ucl.ac.uk

## ABSTRACT

Recent advances in software fairness investigate bias in the treatment of different population groups, which are devised based on attributes such as gender, race and age. Groups are divided into privileged groups (favourable treatment) and unprivileged groups (unfavourable treatment). To truthfully represent the real world and to measure the degree of bias according to age (young vs. old), one needs to pick a threshold to separate those groups.

In this study we investigate two popular datasets (i.e., German and Bank) and the bias observed when using every possible age threshold in order to divide the population into "young" and "old" groups, in combination with three different Machine Learning models (i.e., Logistic Regression, Decision Tree, Support Vector Machine). Our results show that age thresholds do not only impact the intensity of bias in these datasets, but also the direction (i.e., which population group receives a favourable outcome). For the two investigated datasets, we present a selection of suitable age thresholds. We also found strong and very strong correlations between the dataset bias and the respective bias of trained classification models, in 83% of the cases studied.

## CCS CONCEPTS

• **Social and professional topics** → *User characteristics*; • **General and reference** → **Empirical studies**.

## KEYWORDS

software fairness, bias, binary classification

**ACM Reference Format:**
Max Hort and Federica Sarro. 2022. Privileged and Unprivileged Groups: An Empirical Study on the Impact of the Age Attribute on Fairness. In *International Workshop on Equitable Data and Technology (FairWare '22 ), May 9, 2022, Pittsburgh, PA, USA.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3524491.3527308

## 1 INTRODUCTION

Nowadays, a growing number of Machine Learning (ML) models lie at the core of many software systems used world-wide, from social media and visa application systems, to facial recognition technology [26]. These algorithms often relieve users from the burden of tedious manual tasks. However, they have been found culprit of driving inequality and among others affecting human rights [47] and university admissions [6]. This is mainly due to the fact that they are designed and built on data which reflects societal bias humans may have against certain groups or individuals.

Incorporating bias would have a negative effect on software systems, as this suppresses opportunities of deprived groups or individuals [38, 39], due to sensitive attributes (e.g., race, gender, age) rather than merit. In particular, *privileged* population groups would be more likely to receive a favourable treatment than *unprivileged* population groups. Not only is such a behaviour undesired, but it can also face legal risk [17, 51, 52].

Due to its importance as a non-functional property, software fairness has recently received a lot of attention, in both software engineering [9, 16, 31, 59, 60] and machine learning literature [5, 12, 38, 41].

Among the sensitive attributes studied in software fairness literature, race and gender are categorical features, that are used to divide the population into privileged and unprivileged groups (e.g., male - female, white - non-white) [3]. The protected attribute "age" is continuous and needs to be addressed differently.[1] While there exist methods for dealing with continuous attributes (e.g., pairwise comparisons [49] and correlations [27, 46]), we focus on treating protected attributes as binary attributes, in accordance with prior works [15, 16, 35]. To divide the population into two groups (i.e., young - old) one needs to select an age threshold which divides the population as follows: everyone older than the threshold is "old"; everyone of the same age as threshold or younger is "young".

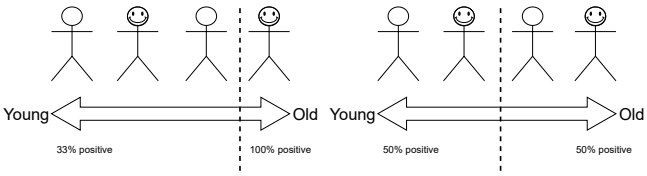
Figure 1 illustrates the impact of different age thresholds on fairness, when treating different populations groups, and the risks of selecting unsuitable thresholds. In Figure 1 (a) a high threshold is applied, dividing the population in three young individuals and one old individual. By doing so, the "old" population group, on average, receives a more favourable treatment (represented by smiling faces) than the "young" group. Using instead the age threshold shown in Figure 1 (b), we can observe that an equal treatment of the two groups is possible, when an adequate age threshold is chosen. While this is a simplified example, it signifies the importance of selecting sensible age thresholds when investigating the fairness of ML software.

Currently, there exists no systematic study focusing on the problem on how to approach the choice of sensible age thresholds when faced with new datasets, and what the impact of age thresholds has on: 1) the dataset; 2) the classification models that are trained

[1] As pointed out by Jacobs and Wallach [34], protected attributes, such as race and gender, are contested constructs.


*FairWare '22 , May 9, 2022, Pittsburgh, PA, USA*

ACM ISBN 978-1-4503-9292-1/22/05…$15.00
https://doi.org/10.1145/3524491.3527308




(a) High age threshold. (b) Medium age threshold.

**Figure 1: Example of the fairness of two different age thresholds. Smiling faces represent a favourable treatment.**

on the dataset. Therefore, we aim to provide guidelines on how to approach the protected attribute "age" from a computational point of view. Nonetheless, if regulations exist, the final choice of acceptable age thresholds is to be done by law makers and domain experts [40, 61].

In summary, the main contributions of this work are:

- a general approach on how to choose age thresholds;
- an empirical evaluation on bias in classification models with respect to age thresholds on two datasets.

The rest of the paper is organized as follows. Section 2 presents related work on software fairness research, including types of bias and an overview of methods to combat bias in classification models. The experimental design, fairness metrics and datasets are outlined in Section 3. Experiments and results are presented in Section 4 while Section 5 concludes.

## 2 RELATED WORK

In recent years, the fairness of software systems has risen in importance, and gained attention from both the software engineering [9, 16, 31, 33, 59, 60] and the machine learning research communities [5, 12, 38, 41]. To date, the software engineering community has tackled fairness at different stages of the software project lifecycle such as requirements analysis [23], design [16], and testing [1, 2, 25, 54, 55].

To improve the fairness of ML software, practitioners proposed three types of debiasing methods used at different stages of the ML development process [24]. First, bias can be prevented from reaching the model before it is trained (pre-processing) [10, 13, 15, 22, 36]. This can be achieved by data modification or removal of data points [16, 61]. Several techniques have been used to mitigate bias during the training process (in-processing) [12, 14, 32, 37, 42, 56]. For example, Zhang et al. [58] used adversarial learning, while others incorporated fairness constraints [5, 11, 41]. Lastly, bias can be combated after models have been trained (post-processing) by either modifying predictions [29, 38, 39] or modifying the classification model [37].

To quantify the fairness of classification models and potential improvements achieved by bias mitigation methods, several fairness metrics have been introduced [4]. These can be divided in two categories [53]: individual fairness (similar individuals should receive a similar treatment); group fairness (privileged and unprivileged groups should receive a similar treatment).

To foster a better understanding of fairness issues and increase the usability of fairness techniques, frameworks, such as AIF Fairness 360 (AIF360) [4] and Fairlearn [7], have been created. Among others, these provide bias mitigation methods, fairness metrics, datasets, and have been frequently used by the software engineering community [15, 16, 33].

Investigations on the effect of datasets on fairness have been carried out by Zhang and Harman [59], and Kamiran and Calders [35]. Zhang and Harman [59] investigated the influence of training data on the fairness of classification models. Particularly, rich feature sets have the ability to improve the fairness of ML models. Kamiran and Calders [35] proposed a pre-processing method called "massaging" with the goal to create an unbiased datasets with the least intrusive modifications before training classification models. Their investigation covered the German dataset (see Section 3.3), for which they chose an age threshold of 25, as a high degree of bias was observed. This age threshold is incorporated in the AIF360 framework [4]. Nonetheless, other thresholds have been used as well, such as 30 and 45 for other datasets [28, 45], and 50 for the German dataset [28].

While Kamiran and Calders [35] focus lay on proposing a novel bias mitigation method on the German dataset (Section 3.3) with the protected attribute "age", we focus our investigation entirely on the choice of age thresholds for multiple datasets. In particular, we do not only consider the German dataset, but a second dataset (Bank), which uses the same age threshold of 25 (according to the AIF360 framework [4]). In addition to measuring the bias and comparing the usability of different age thresholds (e.g., is an age threshold of 25 suitable for the Bank dataset?), we measure the impact of age thresholds on the proceeding bias of three classification models (Logistic Regression, Decision Tree, Support Vector Machine).

## 3 EMPIRICAL STUDY DESIGN

In this section, we describe the design of the analysis we carry out to investigate the impact age thresholds have on the fairness of datasets and classification models. We first introduce the research questions, followed by the subjects and the experimental procedure.

### 3.1 Research Questions

To determine the relation of the protected attribute "age" and the resulting bias in classification problems, we first investigate the bias present in datasets:

**RQ1: What is the impact of age thresholds on the bias in datasets?**

To answer this research question, we investigate the dataset fairness of two datasets (German [30] and Bank [48]) according to the dataset fairness metric *Mean Difference* (Section 3.2). In particular, we evaluate Mean Difference for each possible age threshold for the respective dataset (i.e., the ages present in the dataset). Not only does this allow us to detect the degree of bias that the datasets exhibit, when following different rules to divide the population in to "young" and "old", but also the direction of bias (i.e., which population group receives a favourable treatment).

After determining the degree of bias with respect to the age threshold within a dataset, we investigate the impact of age thresholds on the bias in classification models:

**RQ2. What is the impact of age thresholds on the bias in classification models?**
For this purpose, we train three different classification models (Logistic Regression, Decision Trees, Support Vector Machine) on two datasets (German [30] and Bank [48]). According to the experiments in RQ1, we train the classification models for every possible age threshold to measure resulting biases. This allows us to determine the relation of dataset bias and classification bias in two aspects:

- **RQ2.1** What is the impact of dataset bias on the direction of classification bias (e.g., if the dataset bias favours privileged groups, do classification models as well)?
- **RQ2.2** What is the impact of dataset bias on the degree of classification bias (e.g., does a high dataset bias lead to a high classification bias)?

## 3.2 Fairness Metrics

For our investigation, we are concerned with the disparate treatment of population groups (privileged and unprivileged). Therefore, we use group fairness metrics [16, 18, 19, 41], to determine the "age" bias in datasets. We investigate four group fairness metrics in total (one dataset metric and three classification metrics).

In the proceeding equations: $D$ denotes a group ($D = privileged$ or $D = unprivileged$); $Pr$ to denotes probability; $y$ denotes the true label of an instance and $\hat{y}$ the predictions of a classification model (used for classification metrics).

*Dataset Metrics.* Dataset metrics are used to determine bias in the instances of a dataset. Mean Difference (MD) is a dataset metric which computes differences between privileged and unprivileged group in regards to how likely it is that they receive a favourable treatment (i.e., a positive label).[2]

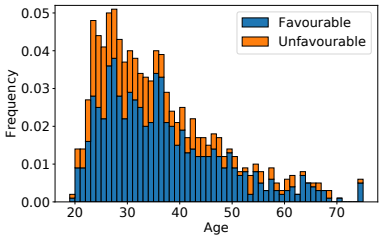
$$MD = Pr(y = 1|D = unprivileged) - Pr(y = 1|D = privileged) \quad (1)$$

*Classification Metrics.* Classification metrics are used to determine the bias of predictions made by classification models. We consider three popular classification metrics: Statistical Parity Difference (SPD) [20], Equal Opportunity Difference (EOD) [29] and Average Odds Difference (AOD) [29]. SPD (Equation 2) computes the difference in favourable and unfavourable classifications for each demographic group [20]. EOD (Equation 3) is determined by the True Positive Rate (TPR) difference [29], while AOD (Equation 4) averages TPR and False Positive Rate (FPR) differences [29].

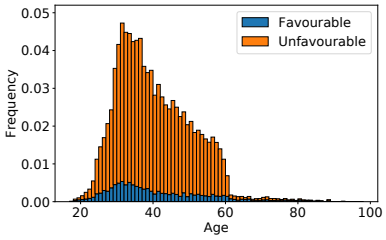
$$SPD = Pr(\hat{y} = 1|D = unprivileged) - Pr(\hat{y} = 1|D = privileged) \quad (2)$$

$$EOD = TPR_{D=unprivileged} - TPR_{D=privileged} \quad (3)$$

$$AOD = \frac{1}{2}((FPR_{D=unprivileged} - FPR_{D=privileged}) + (TPR_{D=unprivileged} - TPR_{D=privileged})) \quad (4)$$




**(a) German**




**(b) Bank**

**Figure 2: Distribution of favourable and unfavourable labels in the German [30] and Bank [48] dataset by age.**

## 3.3 Datasets

We perform our experiments on two publicly available, real-world datasets, widely studied in the fairness literature [10, 14, 16, 22, 56, 57]: the German, and Bank dataset. While there exist other datasets that have been used for fairness research, such as the Adult [43] and COMPAS [3] datasets, we only focus on those datasets that are publicly available in the AIF360 framework [4], have *age* as a protected attribute, and use a default threshold to divide the privileged and unprivileged groups.

The **German Credit Data (German)** [30] dataset contains the credit information of 1,000 individuals. A classification is made, whether individuals have a good or bad credit risk. Among others, the dataset contains additional information about the credit purpose, credit history and employment status.

The **Bank Marketing (Bank)** [48] dataset contains details of direct marketing campaign, which used phone calls, performed by a Portuguese banking institution. Given the information of potential clients, the goal is to predict whether clients subscribe to a term deposit after receiving a phone call. This is denoted by the variable "y" and "1" signals that a client subscribed to a term deposit. Further information in the dataset include education, type of job, and the number of days that passed by after the client was last contacted from a previous campaign

Table 1 provides more information about the two datasets. This includes the size of the dataset, the number of features, the favourable label, and the majority label. The default criteria to form privileged and unprivileged groups from the protected attribute "age" are

[2] Mean Difference can also be called Statistical Parity Difference. We choose to call it Mean Difference to not confuse it with the classification metric which is also called Statistical Parity Difference.

**Table 1: Dataset Information**

| Dataset | Size | Features | Favour Label | Majority Label | Priv. - Unpriv. |
|---|---|---|---|---|---|
| German | 1,000 | 20 | 1 (good credit) | 1 (70%) | $> 25$ - $\leq 25$ |
| Bank | 41,188 | 20 | 1 (yes) | 0 (87%) | $\geq 25$ - $< 25$ |

given.[3] At the time of performing our experiments, individuals with an age $> 25$ are part of the privileged group in the German datasets, whereas the individuals with an age $\geq 25$ are part of the privileged group in the Bank dataset, according to the default settings of the AIF 360 framework [4].

For the two datasets, Figure 2 provides histograms to show how many individuals receive favourable and unfavourable outcomes. When comparing the two datasets, we can see that the average age of the Bank dataset is higher than on the German dataset (40 vs. 35.5). Furthermore, the age range within the dataset is larger on the Bank dataset (17-98) in contrast to the German dataset (19-75).

## 3.4 Experimental Configuration

To carry out our experiments, we use the dataset and fairness metric implementations provided by the AIF360 framework [4]. When loading the datasets, the AIF360 framework allows the definition of rules to determine the age threshold which we use to modify the datasets in RQ1 and RQ2.

For RQ2, we use the data investigated in RQ1 to train classification models. In particular, we consider three classification models that have previously been used in fairness research: Logistic Regression (LR) [15, 16, 22, 38, 39, 41, 56], Decision Trees (DT) [38, 39], and Support Vector Machines (SVM) [15, 22, 39, 56]. We implemented each classification model with scikit-learn [50], according to their default configuration.

When training classification models (RQ2), we use random data-splits with a train-test split of 70%-30%. For each age threshold, we adjust the "age" label of the underlying dataset to "young" and "old" before training classification models. To measure the classification bias, we repeat experiments 50 times, with different train-test splits, and average the results [8, 14].

As RQ2.2 considers the degree of bias and not the direction of bias, we compute the absolute bias values. Thereby, bias is minimized at 0 and maximized at 1. Afterwards, we use the Pearson correlation coefficient [44] to determine the correlation between dataset bias and classification bias.

# 4 EMPIRICAL STUDY RESULTS

This section presents the results of our experiments to answer the research questions explained in Section 3.1.

## 4.1 RQ1: Dataset Fairness

The first research question investigates the fairness of the two datasets (German, Bank) according to the dataset fairness metric *Mean Difference* (Section 1).

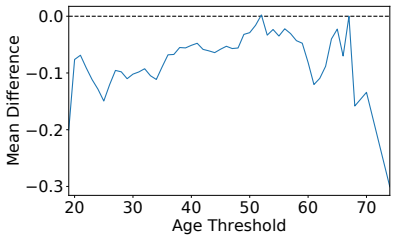
To evaluate datasets based on Mean Difference (probability that unprivileged group receives a favourable label - probability that privileged group receives a favourable label), we compute the Mean Difference for every possible age threshold to create privileged and unprivileged groups. In particular, we gather a list of unique ages that are present in the two datasets (i.e., there are 53 unique age values in the German dataset and 78 unique ages in the Bank dataset) and use each value to separate privileged and unprivileged groups. Given an age threshold $a_t$, the privileged group consists of all instances of a dataset for which $age > a_t$, the remaining instances are part of the unprivileged group. This is due to the fact that for both, the German and Bank dataset, "young" individuals are deemed (according to the default configuration of the AIF360 framework [4]). We perform this for each age that is present in the dataset except for the maximum (oldest) age, to ensure that both groups (privileged and unprivileged) are not empty. Therefore, we collect 52 measures of Mean Difference for the German datasets and 77 measures for the Bank dataset. Figure 3 illustrates the results.

When analyzing the Mean Difference of the different age thresholds, we can see that general notion of bias and privilege holds for the German dataset: **Privileged groups are more likely to receive a favourable outcome**. This is indicated by a negative Mean Difference. Only at the thresholds 52 and 67 are non-negative Mean Difference values reached (0.002 and 0). Furthermore, we observe that the default setting ($age > 25$ is old) provided by the AIF360 framework [4] which was chosen according to Kamiran and Calders [35], is a logical choice. The Mean Difference with an age threshold of 25 is $-0.15$, which is a local minimum. This divides the dataset into an unprivileged group which contains 19% of the instances, the remaining 81% are part of the privileged group. A balanced division of groups, according to the age median of 33, achieves a Mean Difference of $-0.1$ while 48% of the instances belong to the privileged and 52% to the unprivileged group. Given the purpose of the protected attribute (e.g., causing the highest disparity between privileged or unprivileged, or having groups of balanced sizes) an age threshold between 25 and 33 inclusive is reasonable.

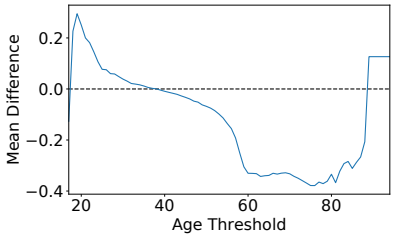
Using age thresholds of 19, 68 and 74 achieves an even higher Mean Difference than 25, however they cause imbalanced sizes of privileged and unprivileged groups (with the smaller of the two being of size 2%, 7%, 2% respectively).

The Mean Difference of age thresholds for the Bank dataset shows a different situation: **There are thresholds at which the unprivileged group is more likely to receive a favourable outcome than the privileged group**. In particular, the Mean Difference is positive within the intervals 18-38 (38 being the median age of the Bank dataset) and 89-94. We disregard the latter interval, because the size of the privileged group at an age threshold of 89 is only 10%, given a dataset size of 41,188 (Table 1). Choosing an age threshold within 18-38 would violate our conception of bias, as it does not favour the privileged group. Therefore, using a default threshold of 25, which is motivated based on the German datasets' threshold, does not represent the dataset correctly (given that the privileged group is "old"). Either the notion of privileged and unprivileged groups ought to be adjusted (i.e., "young" is a privileged group given an age threshold of 25) or the threshold value should be increased. Potential values, for which our notion of bias holds, are 47 (the 75%-percentile with a Mean Difference of -0.05) or 59 (mean difference of -0.31), which is the first threshold followed by

[3] We use the default parameter from version 0.4.0 or the AIF360 framework, last updated on the fourth of March 2021.




**(a) German Dataset**




**(b) Bank Dataset**

**Figure 3: RQ1: Mean Difference of the German and Bank dataset for each possible age threshold to divide privileged and unprivileged groups.**

a sharp decrease in Mean Difference as seen in Figure 3. The size of the privileged group at a threshold of 59 is 3%, opposed to 24% at a threshold of 47.

**To conclude:** We showed that the age threshold does not only impact the degree of bias, but also the bias direction. While an age threshold of 25, to distinguish privileged and unprivileged groups, is reasonable for the German datasets, it violates our notion of fairness on the Bank dataset, by favouring the unprivileged group. In addition to determining the bias between privileged and unprivileged group, age thresholds also impact the balance between the group sizes.

## 4.2 RQ2: Classification Fairness

Following, we carry out experiments to determine the bias of classification models, when being trained on the German and Bank datasets under different age thresholds. In particular, we investigate the relation of the bias present in the dataset and its impact when using it to train classification models.

*4.2.1 RQ2.1: Bias direction.* To answer RQ2.1, we consider the same pair of datasets as used for RQ1 as well as the same procedure to determine age thresholds. For each age threshold, the datasets are adjusted (i.e., setting the protected attribute age to "young" or "old" depending on the age of an individual and the age threshold). We then train three classification models (LR, DT, SVM) for each dataset and age threshold. Afterwards, we determine the bias degree of the classification models according to three classification metrics (SPD, AOD, EOD).

If the bias measures are $< 0$ it signifies that the privileged group is favoured, whereas if bias measures $> 0$ it shows that the unprivileged group receives a favourable treatment. If there is no bias present in the prediction made by a classification model the metric is equal to 0.

**Table 2: RQ2.1: Percentage of age thresholds for which Mean Difference and classification metrics are in the same direction (favour the same population group.**

| | German | | | Bank | | |
|---|---|---|---|---|---|---|
| | SPD | AOD | EOD | SPD | AOD | EOD |
| Logistic Regression | 100% | 80% | 97% | 96% | 80% | 77% |
| Decision Tree | 89% | 49% | 89% | 99% | 91% | 87% |
| Support Vector Machine | 6% | 6% | 14% | 97% | 87% | 81% |

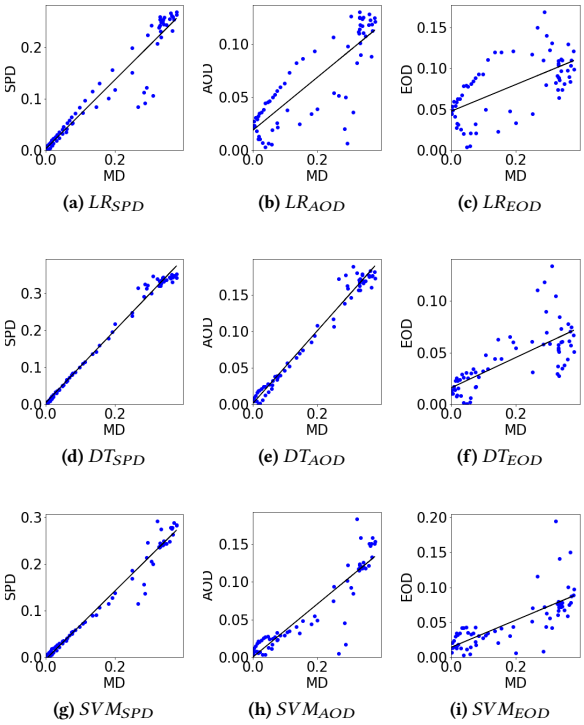
Table 2 shows the results. For the two datasets and three classification metrics, we perform experiments for every age threshold and measure the proportion of bias directions which agree with the Mean Difference. On the Bank dataset, we can observe that for every of the nine pairs (three classification values and three classification metrics) the direction of bias agrees with MD in at least 77% of the cases. For each classification model, the bias direction of SPD agrees with MD in at least 96% of the cases. This value is higher than for the two confusion matrix metrics AOD and EOD, as the computation of SPD and MD are similar.

While for 15 out of 18 evaluations, the direction of MD and classification metrics are alike (at least 77% of measures have the same direction), we are not able to confirm that an underlying dataset bias leads to classification models that are biased in the same way (e.g., a dataset that is biased towards the privileged group does not always lead to classification models that do the same). In particular, there are cases on the German dataset which on the dataset level favour the privileged group, but when trained on SVMs are more likely to favour the unprivileged group. Reasons for this disparity can be seen in the small size of the German dataset (1,000 instances) or the high degree of imbalance (87% of the instances receive an unfavourable outcome).

*4.2.2 RQ2.2: Bias intensity.* In addition to investigating the relation of bias direction in regards to dataset and classification bias, we are interested to see whether a highly biased dataset (e.g., high Mean Difference) leads to highly biased classification model, or vice versa (i.e., low dataset bias leads to fair classification models). Since we are only interested in the bias intensity and not the direction of bias, we continue our investigation with absolute bias values.

Figure 4 illustrates the relation of dataset and classification bias for the Bank dataset and Figure 5 for the German dataset. Each age threshold is represented as a point in the graphs, with the intensity of dataset bias (Mean Difference) on the x-axis and intensity of one of the classification metrics on the y-axis. In addition to the dataset-classification bias pairs, each graph displays a regression line, with the corresponding Pearson correlation coefficient [44] shown in Table 3. We follow the guidelines proposed by Evans [21], who described correlation strength as: very weak (±0.00 ±0.19), weak (±0.20 ±0.39), moderate (±0.40 ±0.59), strong (±0.60 ±0.79) and very strong (±0.80 ±1.00).

When looking at the Bank dataset, we can observe that the correlation between dataset and classification metrics are either very strong (for SPD and AOD) or strong (EOD) for all classification

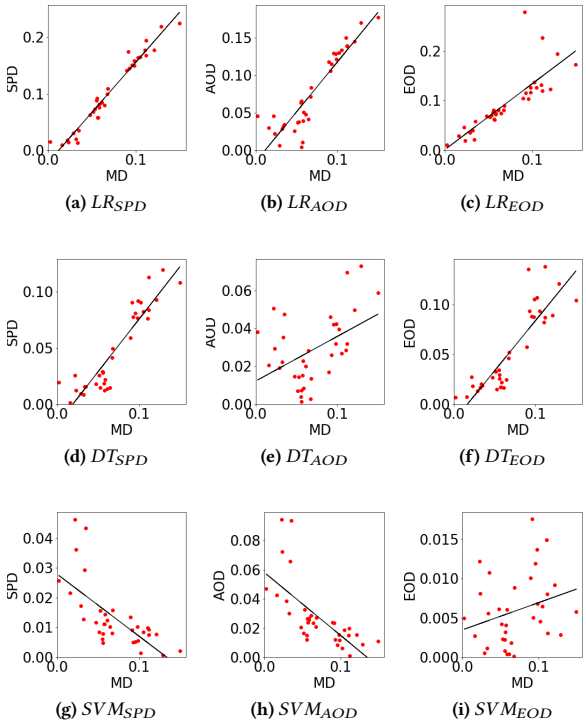



(a) $LR_{SPD}$ (b) $LR_{AOD}$ (c) $LR_{EOD}$

(d) $DT_{SPD}$ (e) $DT_{AOD}$ (f) $DT_{EOD}$

(g) $SVM_{SPD}$ (h) $SVM_{AOD}$ (i) $SVM_{EOD}$

**Figure 4: RQ2.2 Bank: Relation of Mean Difference (MD) and classification metrics (SPD, AOD, EOD). Each point represents the bias of an age threshold (dataset bias before training and classification bias after training the given classification model). A regression line is shown in black.**




(a) $LR_{SPD}$ (b) $LR_{AOD}$ (c) $LR_{EOD}$

(d) $DT_{SPD}$ (e) $DT_{AOD}$ (f) $DT_{EOD}$

(g) $SVM_{SPD}$ (h) $SVM_{AOD}$ (i) $SVM_{EOD}$

**Figure 5: RQ2.2 German: Relation of Mean Difference (MD) and classification metrics (SPD, AOD, EOD). Each point represents the bias of an age threshold (dataset bias before training and classification bias after training the given classification model). A regression line is shown in black.**

models. The Bank dataset confirms the intuition that a high bias in the dataset (according to Mean Difference) leads to a high bias in classification models that are trained on this data. A similar conclusion can be drawn for the German dataset when only considering LR and DT classification models. However, the results on SVMs do not comply with this intuition. Differently from all the other evaluations, dataset bias and classification bias are inverse-correlated for SVMs on the German dataset (i.e., a large dataset bias leads to classification models with little bias). Reasons for such observations could be the small dataset size or properties of the classification model.

## 5 CONCLUSION

Recent advances on the investigation of software fairness are conducted by dividing the population in two groups (privileged and unprivileged) based on protected attributes. Protected attributes come in the form of categorical, and continuous attributes for which thresholds need to be chosen. Our work provides choices on thresholds when dealing with continuous protected attributes (i.e., "age"), which has a direct impact on the perceived bias of software systems. We performed a detailed study on age thresholds and their impact on fairness for two frequently used datasets in fairness research.

Our findings show that age thresholds that are sufficient for one dataset (e.g., 25 for the German dataset) can not be transferred to other datasets without further considerations. Furthermore, even though the dataset bias is correlated to the bias in subsequently trained classification models (e.g., high bias in datasets leads to a high bias in classification models), we also found examples for which this is not true. Therefore, we cannot confirm the notion that a high bias in datasets corresponds with a high bias in classification models. However, instead of interpreting this as an issue (e.g., when using SVMs on the German dataset), one could investigate further classification models to check whether their classification bias is correlated to the underlying dataset bias, in future work.

While we provided potential age thresholds for the German and Bank datasets, and support the decision making process when dealing with continuous protected attributes, we note that the ultimate choice of age thresholds is up to practitioners and lawmakers.

## ACKNOWLEDGEMENTS

This research is funded by the ERC advanced fellowship grant 741278 (EPIC: Evolutionary Program Improvement Collaborators).

**Table 3: RQ2.2: Pearson correlation coefficient and the corresponding p-value for Mean Difference (MD) and classification metrics (SPD, AOD, EOD).**

| Correlation (p-value) | Bank | | | German | | |
|---|---|---|---|---|---|---|
| | SPD | AOD | EOD | SPD | AOD | EOD |
| Logistic Regression | 0.95 (0.00) | 0.83 (0.00) | 0.63 (0.00) | 0.98 (0.00) | 0.91 (0.00) | 0.82 (0.00) |
| Decision Tree | 1.00 (0.00) | 0.99 (0.00) | 0.72 (0.00) | 0.91 (0.00) | 0.46 (0.01) | 0.88 (0.00) |
| Support Vector Machine | 0.98 (0.00) | 0.92 (0.00) | 0.75 (0.00) | -0.69 (0.00) | -0.69 (0.00) | 0.28 (0.11) |

# REFERENCES


[1] Aniya Aggarwal, Pranay Lohia, Seema Nagar, Kuntal Dey, and Diptikalyan Saha. 2019. Black box fairness testing of machine learning models. In *Proceedings of the 2019 27th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*. 625–635. https://doi.org/10.1145/3338906.3338937

[2] Rico Angell, Brittany Johnson, Yuriy Brun, and Alexandra Meliou. 2018. Themis: Automatically testing software for discrimination. In *Proceedings of the 2018 26th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*. 871–875. https://doi.org/10.1145/3236024.3264590

[3] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. Machine bias. ProPublica. *See https://www. propublica. org/article/machine-bias-risk-assessments-in-criminal-sentencing* (2016).

[4] Rachel KE Bellamy, Kuntal Dey, Michael Hind, Samuel C Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilovic, et al. 2018. AI Fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. *arXiv preprint arXiv:1810.01943* (2018).

[5] R. Berk, H. Heidari, S. Jabbari, M. Joseph, M. Kearns, J. Morgenstern, S. Neel, and A. Roth. 2017. A Convex Framework for Fair Regression. *FAT-ML Workshop* (2017).

[6] Peter J Bickel, Eugene A Hammel, and J William O'Connell. 1975. Sex bias in graduate admissions: Data from Berkeley. *Science* 187, 4175 (1975), 398–404. https://doi.org/10.1126/science.187.4175.398

[7] Sarah Bird, Miro Dudík, Richard Edgar, Brandon Horn, Roman Lutz, Vanessa Milan, Mehrnoosh Sameki, Hanna Wallach, and Kathleen Walker. 2020. Fairlearn: A toolkit for assessing and improving fairness in AI. *Microsoft, Tech. Rep. MSR-TR-2020-32* (2020).

[8] Sumon Biswas and Rajan Hridesh. 2020. Do the Machine Learning Models on a Crowd Sourced Platform Exhibit Bias? An Empirical Study on Model Fairness. *arXiv preprint arXiv:2005.12379* (2020). https://doi.org/10.1145/3368089.3409704

[9] Yuriy Brun and Alexandra Meliou. 2018. Software fairness. In *Proceedings of the 2018 26th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*. 754–759. https://doi.org/10.1145/3236024.3264838

[10] Toon Calders, Faisal Kamiran, and Mykola Pechenizkiy. 2009. Building classifiers with independency constraints. In *2009 IEEE International Conference on Data Mining Workshops*. IEEE, 13–18. https://doi.org/10.1109/ICDMW.2009.83

[11] Toon Calders, Asim Karim, Faisal Kamiran, Wasif Ali, and Xiangliang Zhang. 2013. Controlling attribute effect in linear regression. In *2013 IEEE 13th international conference on data mining*. IEEE, 71–80. https://doi.org/10.1109/ICDM.2013.114

[12] Toon Calders and Sicco Verwer. 2010. Three naive Bayes approaches for discrimination-free classification. *Data Mining and Knowledge Discovery* 21, 2 (2010), 277–292. https://doi.org/10.1007/s10618-010-0190-x

[13] Flavio Calmon, Dennis Wei, Bhanukiran Vinzamuri, Karthikeyan Natesan Ramamurthy, and Kush R Varshney. 2017. Optimized pre-processing for discrimination prevention. In *Advances in Neural Information Processing Systems*. 3992–4001.

[14] L Elisa Celis, Lingxiao Huang, Vijay Keswani, and Nisheeth K Vishnoi. 2019. Classification with fairness constraints: A meta-algorithm with provable guarantees. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 319–328. https://doi.org/10.1145/3287560.3287586

[15] Joymallya Chakraborty, Suvodeep Majumder, and Tim Menzies. 2021. Bias in machine learning software: why? how? what to do? *Proceedings of the 29th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering* (2021).

[16] Joymallya Chakraborty, Suvodeep Majumder, Zhe Yu, and Tim Menzies. 2020. Fairway: a way to build fair ML software. In *Proceedings of the 28th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*. 654–665. https://doi.org/10.1145/3368089.3409697

[17] Jiahao Chen, Nathan Kallus, Xiaojie Mao, Geoffry Svacha, and Madeleine Udell. 2019. Fairness under unawareness: Assessing disparity when protected class is unobserved. In *Proceedings of the conference on fairness, accountability, and transparency*. 339–348. https://doi.org/10.1145/3287560.3287594

[18] Sam Corbett-Davies and Sharad Goel. 2018. The measure and mismeasure of fairness: A critical review of fair machine learning. *arXiv preprint arXiv:1808.00023* (2018).

[19] Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. 2017. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 797–806. https://doi.org/10.1145/3097983.3098095

[20] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*. 214–226. https://doi.org/10.1145/2090236.2090255

[21] James D Evans. 1996. *Straightforward statistics for the behavioral sciences*. Thomson Brooks/Cole Publishing Co.

[22] Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. 2015. Certifying and removing disparate impact. In *proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*. 259–268. https://doi.org/10.1145/2783258.2783311

[23] Anthony Finkelstein, Mark Harman, S Afshin Mansouri, Jian Ren, and Yuanyuan Zhang. 2009. A search based approach to fairness analysis in requirement assignments to aid negotiation, mediation and decision making. *Requirements engineering* 14, 4 (2009), 231–245. https://doi.org/10.1007/s00766-009-0075-y

[24] Sorelle A Friedler, Carlos Scheidegger, Suresh Venkatasubramanian, Sonam Choudhary, Evan P Hamilton, and Derek Roth. 2019. A comparative study of fairness-enhancing interventions in machine learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. ACM, 329–338. https://doi.org/10.1145/3287560.3287589

[25] Sainyam Galhotra, Yuriy Brun, and Alexandra Meliou. 2017. Fairness testing: testing software for discrimination. In *Proceedings of the 2017 11th Joint Meeting on Foundations of Software Engineering*. ACM, 498–510. https://doi.org/10.1145/3106237.3106277

[26] Clare Garvie and Jonathan Frankle. 2016. Facial-recognition software might have a racial bias problem. *The Atlantic* 7 (2016).

[27] Vincent Grari, Boris Ruf, Sylvain Lamprier, and Marcin Detyniecki. 2019. Fairness-Aware Neural R\'eyni Minimization for Continuous Features. *arXiv preprint arXiv:1911.04929* (2019).

[28] Sara Hajian and Josep Domingo-Ferrer. 2012. A methodology for direct and indirect discrimination prevention in data mining. *IEEE transactions on knowledge and data engineering* 25, 7 (2012), 1445–1459.

[29] Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of opportunity in supervised learning. In *Advances in neural information processing systems*. 3315–3323.

[30] Hans Hofmann. 1994. Statlog (german credit data) data set. *UCI Repository of Machine Learning Databases* 53 (1994).

[31] Jennifer Horkoff. 2019. Non-functional requirements for machine learning: Challenges and new directions. In *2019 IEEE 27th International Requirements Engineering Conference (RE)*. IEEE, 386–391. https://doi.org/10.1109/RE.2019.00050

[32] Max Hort and Federica Sarro. 2021. Did You Do Your Homework? Raising Awareness on Software Fairness and Discrimination. In *2021 36th IEEE/ACM International Conference on Automated Software Engineering (ASE)*. IEEE, 1322–1326.

[33] Max Hort, Jie M Zhang, Federica Sarro, and Mark Harman. 2021. Fairea: a model behaviour mutation approach to benchmarking bias mitigation methods. In *Proceedings of the 29th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*. 994–1006.

[34] Abigail Z. Jacobs and Hanna Wallach. 2021. Measurement and Fairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (Virtual Event, Canada) *(FAccT '21)*. Association for Computing Machinery, New York, NY, USA, 375–385. https://doi.org/10.1145/3442188.3445901

[35] Faisal Kamiran and Toon Calders. 2009. Classifying without discriminating. In *2009 2nd International Conference on Computer, Control and Communication*. IEEE,

1–6. https://doi.org/10.1109/IC4.2009.4909197
[36] Faisal Kamiran and Toon Calders. 2012. Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems* 33, 1 (2012), 1–33. https://doi.org/10.1007/s10115-011-0463-8
[37] Faisal Kamiran, Toon Calders, and Mykola Pechenizkiy. 2010. Discrimination aware decision tree learning. In *2010 IEEE International Conference on Data Mining*. IEEE, 869–874. https://doi.org/10.1109/ICDM.2010.50
[38] Faisal Kamiran, Asim Karim, and Xiangliang Zhang. 2012. Decision theory for discrimination-aware classification. In *2012 IEEE 12th International Conference on Data Mining*. IEEE, 924–929. https://doi.org/10.1109/ICDM.2012.45
[39] Faisal Kamiran, Sameen Mansha, Asim Karim, and Xiangliang Zhang. 2018. Exploiting reject option in classification for social discrimination control. *Information Sciences* 425 (2018), 18–33. https://doi.org/10.1016/j.ins.2017.09.064
[40] Faisal Kamiran and Indrė Žliobaitė. 2013. Explainable and non-explainable discrimination in classification. In *Discrimination and Privacy in the Information Society*. Springer, 155–170. https://doi.org/10.1007/978-3-642-30487-3_8
[41] Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. 2012. Fairness-aware classifier with prejudice remover regularizer. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 35–50. https://doi.org/10.1007/978-3-642-33486-3_3
[42] Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. 2018. Preventing Fairness Gerrymandering: Auditing and Learning for Subgroup Fairness *(Proceedings of Machine Learning Research, Vol. 80)*, Jennifer Dy and Andreas Krause (Eds.). PMLR, Stockholmsmässan, Stockholm Sweden, 2564–2572. http://proceedings.mlr.press/v80/kearns18a.html
[43] Ronny Kohavi and Barry Becker. 1996. Adult data set. *UCI machine learning repository* 5 (1996), 2093.
[44] Joseph Lee Rodgers and W Alan Nicewander. 1988. Thirteen ways to look at the correlation coefficient. *The American Statistician* 42, 1 (1988), 59–66.
[45] Pranay Lohia. 2021. Priority-based Post-Processing Bias Mitigation for Individual and Group Fairness. *arXiv preprint arXiv:2102.00417* (2021).
[46] Jérémie Mary, Clément Calauzenes, and Noureddine El Karoui. 2019. Fairness-aware learning for continuous attributes and treatments. In *International Conference on Machine Learning*. PMLR, 4382–4391.
[47] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2019. A survey on bias and fairness in machine learning. *arXiv preprint arXiv:1908.09635* (2019).
[48] Sérgio Moro, Paulo Cortez, and Paulo Rita. 2014. A data-driven approach to predict the success of bank telemarketing. *Decision Support Systems* 62 (2014), 22–31. https://doi.org/10.1016/j.dss.2014.03.001
[49] Harikrishna Narasimhan, Andrew Cotter, Maya Gupta, and Serena Wang. 2020. Pairwise fairness for ranking and regression. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 5248–5255.
[50] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.
[51] Dino Pedreshi, Salvatore Ruggieri, and Franco Turini. 2008. Discrimination-aware data mining. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*. 560–568. https://doi.org/10.1145/1401890.1401959
[52] Andrea Romei and Salvatore Ruggieri. 2011. A multidisciplinary survey on discrimination analysis. https://doi.org/10.1017/S0269888913000039
[53] Till Speicher, Hoda Heidari, Nina Grgic-Hlaca, Krishna P Gummadi, Adish Singla, Adrian Weller, and Muhammad Bilal Zafar. 2018. A unified approach to quantifying algorithmic unfairness: Measuring individual &group unfairness via inequality indices. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2239–2248. https://doi.org/10.1145/3219819.3220046
[54] Florian Tramer, Vaggelis Atlidakis, Roxana Geambasu, Daniel Hsu, Jean-Pierre Hubaux, Mathias Humbert, Ari Juels, and Huang Lin. 2017. FairTest: Discovering unwarranted associations in data-driven applications. In *2017 IEEE European Symposium on Security and Privacy (EuroS&P)*. IEEE, 401–416. https://doi.org/10.1109/EuroSP.2017.29
[55] Sakshi Udeshi, Pryanshu Arora, and Sudipta Chattopadhyay. 2018. Automated directed fairness testing. In *Proceedings of the 33rd ACM/IEEE International Conference on Automated Software Engineering*. 98–108. https://doi.org/10.1145/3238147.3238165
[56] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rogriguez, and Krishna P Gummadi. 2017. Fairness constraints: Mechanisms for fair classification. In *Artificial Intelligence and Statistics*. 962–970.
[57] Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. 2013. Learning fair representations. In *International Conference on Machine Learning*. 325–333.
[58] Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. 2018. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*. ACM, 335–340. https://doi.org/10.1145/3278721.3278779
[59] Jie Zhang and Mark Harman. 2021. "Ignorance and Prejudice" in Software Fairness. In *2021 IEEE/ACM 43th International Conference on Software Engineering (ICSE)*. IEEE. https://doi.org/10.1109/ICSE43902.2021.00129
[60] J. M. Zhang, M. Harman, L. Ma, and Y. Liu. 2020. Machine Learning Testing: Survey, Landscapes and Horizons. *IEEE Transactions on Software Engineering* (2020), 1–1. https://doi.org/10.1109/tse.2019.2962027
[61] Indre Žliobaite, Faisal Kamiran, and Toon Calders. 2011. Handling conditional discrimination. In *2011 IEEE 11th International Conference on Data Mining*. IEEE, 992–1001. https://doi.org/10.1109/ICDM.2011.72