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:

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 population 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.

\(^1\)As pointed out by Jacobs and Wallach [34], protected attributes, such as race and gender, are contested constructs.
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 “mas-saging” 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 into “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 (\( D = \text{privileged} \) or \( D = \text{unprivileged} \)); \( Pr \) 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).\(^2\)

\[
MD = Pr(\ y = 1|D = \text{unprivileged}) - Pr(\ y = 1|D = \text{privileged})
\]

**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].

\[
SPD = Pr(\ \hat{y} = 1|D = \text{unprivileged}) - Pr(\ \hat{y} = 1|D = \text{privileged})
\]

\[
EOD = TPR_{D=\text{unprivileged}} - TPR_{D=\text{privileged}}
\]

\[
AOD = \frac{1}{2} \left( \frac{FPR_{D=\text{unprivileged}} - FPR_{D=\text{privileged}}}{1 - TPR_{D=\text{privileged}}} + \frac{FPR_{D=\text{privileged}} - FPR_{D=\text{unprivileged}}}{1 - TPR_{D=\text{unprivileged}}} \right)
\]

\(^2\)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.

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
When comparing the two datasets, we can see that the average age we compute the absolute bias values. Thereby, bias is minimized at

With an age

The first research question investigates the fairness of the two 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, 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 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.

### 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, 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 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 \( \text{age} > a_t \), 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
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 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 ($\pm 0.00 \pm 0.19$), weak ($\pm 0.20 \pm 0.39$), moderate ($\pm 0.40 \pm 0.59$), strong ($\pm 0.60 \pm 0.79$) and very strong ($\pm 0.80 \pm 1.00$).

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 metrics. The correlation between dataset and classification metrics is either very strong (for SPD and AOD) or strong (EOD) for all classification metrics.
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).

<table>
<thead>
<tr>
<th>Metric</th>
<th>SPD (p-value)</th>
<th>Bank AOD (p-value)</th>
<th>EOD (p-value)</th>
<th>SPD (p-value)</th>
<th>German AOD (p-value)</th>
<th>EOD (p-value)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Logistic Regression</td>
<td>0.95 (0.00)</td>
<td>0.83 (0.00)</td>
<td>0.63 (0.00)</td>
<td>0.98 (0.00)</td>
<td>0.91 (0.00)</td>
<td>0.82 (0.00)</td>
</tr>
<tr>
<td>Decision Tree</td>
<td>1.00 (0.00)</td>
<td>0.99 (0.00)</td>
<td>0.72 (0.00)</td>
<td>0.91 (0.00)</td>
<td>0.46 (0.01)</td>
<td>0.88 (0.00)</td>
</tr>
<tr>
<td>Support Vector Machine</td>
<td>0.98 (0.00)</td>
<td>0.92 (0.00)</td>
<td>0.75 (0.00)</td>
<td>-0.69 (0.00)</td>
<td>-0.69 (0.00)</td>
<td>0.28 (0.11)</td>
</tr>
</tbody>
</table>

### REFERENCES


