Vrije Universiteit Amsterdam

VU VRIJE UNIVERSITEIT AMSTERDAM

EY

Shape the future with confidence

Master's Thesis Business Analytics

Fairness-aware Machine Learning for Employee Attrition Prediction: An Intersectional Subgroup Analysis

Author: Lukas Noppert (2710850)

First supervisor: Dr. Shujian Yu
Daily supervisor: Adem Gökalp (EY)
Second reader: Dr. Rikkert Hindriks

July 27, 2026

Abstract

ML models have been increasingly used in the setting of employment, with concerns arising about the fairness of such models. This thesis considers employee attrition prediction and evaluates the unfairness with respect to gender, age, and education in baseline machine learning models. Subsequently, fairness-aware machine learning models are applied and evaluated. The existing literature on fairness-aware machine learning mostly focuses on a single sensitive attribute, and the employment setting is rarely investigated. This research addresses this gap by evaluating four fairness-aware ML models that can handle multiple sensitive attributes. Reweighing, Differential Fairness Method, Reductions Based Approach, and the novel proposed Fair Ensemble are considered. These methods are evaluated on a primary dataset and, as a robustness check, on the IBM dataset. The existing methods considering intersectional subgroups all reduced unfairness substantially relative to the baselines, with a maximum drop in F1 score of 0.005 on the primary dataset. A second conclusion is that no method was observed to yield superior performance compared to the other methods. Among the existing methods, the Differential Fairness Method reduced the intersectional fairness metrics the most, at a cost of a 0.004 reduction in F1 score for the primary dataset. However, no significant difference was found between the intersectional fairness metrics values obtained by the Differential Fairness Method and those obtained by the Reductions Based Approach. The novel Fair Ensemble improved fairness on the primary dataset but underperformed compared to the best single-attribute method. Moreover, it increased the unfairness in the IBM dataset relative to the logistic regression baseline. The Reductions Based Approach performed best with respect to point estimates on the IBM dataset, indicating that it is the most robust to a shift in data distribution, though no method dominated on both datasets.

Contents

Chapter 1

Introduction

Employee attrition is an important HR task. Identifying which employee is likely to leave allows companies to take proactive measures. This, in turn, could reduce costs significantly. In 2025, the average cost per hire for non-executives in the United States is $\$5,475$, and for executive roles, it was $\$35,879$ [50]. Machine Learning (ML) has proven to work in this domain, as ensemble approaches achieved an accuracy of 93% [48].

However, ML models are known to be sensitive to bias. The models learn patterns in the data. If this data reflects systematic historical patterns of human bias, the decisions made by these models might become unfair. An example is the COMPAS software, which was used to aid in decision-making regarding the release of prisoners. It was later revealed that this software included a bias against convicts of African-American heritage [36]. Having such a bias in a deployed model is obviously undesirable from an ethical standpoint and can have severe consequences for the people involved. This example illustrates that when such models are deployed, either directly or indirectly, in a decision-making process, the outcomes of these models must be fair, especially when sensitive data is leveraged. Simply removing the sensitive variables is not enough to ensure fairness. The models can still rely on correlated variables known as proxies. For example, zip code can act as a proxy for race, allowing the model to indirectly discriminate even when the sensitive variable is removed [46]. This is concerning in Human Resources (HR), where the data contains multiple potentially sensitive attributes such as gender, age, and education. The presence of these sensitive attributes and interactions introduces additional complexity that single sensitive attribute analysis might fail to capture.

Despite this, most fairness-aware ML research has focused on analyzing single binary sensitive attributes [43], and the HR domain has received little attention. In the field of fairness-aware ML, most research is conducted either on the Adult dataset or on the COMPAS dataset (77% and 51% of the research, respectively) [24]. The Adult dataset contains data about the income of adults, where the attribute gender is often specified as the sensitive feature. The COMPAS dataset consists of criminal records, where race and gender are mostly treated as sensitive features [24]. The lack of research is remarkable, given that HR data often consists of multiple sensitive features. However, real-world HR datasets are rarely publicly available. Besides the ethical concerns, organizations are often not allowed to share personal data due to regulations such as the General Data Protection Regulation (GDPR) in the European Union. Moreover, concerns about trustworthiness, reputational risk, and misuse could be motivations for not sharing such data. As a result, studies on employee attrition prediction often use synthetic datasets. Furthermore, this limited availability of real data could explain why fairness-aware ML has received little attention in the HR field.

Biases in ML models can arise from different sources, including historical bias, measurement bias, representation bias, algorithmic bias, and more [52]. In HR datasets, historical bias is a possible concern, and fairness-aware ML is useful for mitigating this type of bias. Fairness-aware ML models addressing this can be split into three categories: pre-processing, in-processing, and post-processing methods. Pre-processing involves altering the training

data before training to make the data fairer. In-processing involves introducing fairness during training. Lastly, post-processing aims to establish fairness after training is finished. In the scope of this thesis, only pre-processing and in-processing techniques will be considered. Since post-processing techniques operate on the model outputs, they have been shown to generally underperform compared to pre-processing and in-processing techniques [44].

Motivated by this gap, this thesis focuses on how fairness-aware ML can be effectively applied to HR data, focusing on multiple and multiclass sensitive attributes: gender, age, and education. These attributes are evaluated not only at the group-level but also across their intersectional subgroups, which is rarely considered in the literature. Among the proposed models is a novel Fair Ensemble approach, which combines multiple fairness-aware models to improve fairness across groups and intersectional subgroups. This thesis therefore asks: *"What trade-offs do pre- and in-processing fairness mitigation methods provide between predictive performance and intersectional fairness in employee attrition prediction?"*

To aid in answering the research question, the following sub-questions are addressed:

1. *How do baseline employee attrition prediction models perform in terms of predictive performance and intersectional fairness?*

2. *How do pre- and in-processing fairness mitigation methods reduce unfairness for various sensitive attributes and their intersectional subgroups in employee attrition prediction?*

3. *What trade-offs between predictive performance and intersectional fairness can be identified when applying the fairness mitigation methods compared to the baseline employee attrition prediction models?*

Ernst and Young (EY) is a large professional services firm with over 400,000 employees, operating in various industries. EY is pursuing AI implementation for its clients across multiple sectors, such as public services, healthcare, and financial services. Moreover, EY also promotes the implementation of AI within its internal functions, for instance, consulting, assurance, and tax. Furthermore, employee retention is important for the firm. With 400,000 employees, the total cost of hiring new employees can be substantial. As discussed, ML models have proven to be useful and can possibly identify which employees are likely to leave. However, AI models are known to amplify historical biases, which makes fairness and ethical concerns extremely important when deploying these models. This thesis will investigate fairness-aware ML models for employee attrition prediction to provide EY with potentially valuable insights in this domain.

The remaining part of the thesis is structured as follows. Chapter 2 provides an overview of the relevant literature. Subsequently, Chapter 3 describes the data used. Thereafter, Chapter 4 discusses the methods, fairness metrics, and evaluation metrics. Chapter 5 discusses the experimental setup. Chapter 6 shows the results, and Chapter 7 discusses and analyzes these results. Lastly, Chapter 8 provides a conclusion.

Chapter 2

Literature review

Fairness-aware ML is a research field with hundreds of different methods, all aiming to introduce fairness while maintaining predictive performance [24]. Moreover, there are numerous fairness metrics, all considering different notions of fairness [8]. To clarify the concepts, this section provides an overview of all the components in the fairness-aware pipeline, backed by relevant methods from the literature.

First, Section 2.1 discusses various sources of bias that are potentially present in this research. Subsequently, Section 2.2 reviews important fairness metrics. Then, Section 2.3 discusses the accuracy-fairness trade-off. Section 2.4 reviews prior research on fairness-aware ML. Section 2.5 discusses the ML models often used for employee attrition prediction, and lastly, Section 2.6 concludes the section by briefly discussing the research gap.

2.1 Sources of Bias

In ML models, bias can arise from several sources. Moreover, these biases can arise in several parts of the modeling pipeline: from data to algorithm, from algorithm to user, and from user to data. In the context of HR data, all three categories are relevant, as unfairness can arise in all three parts. In addition, within these categories, the sources of bias can be broken down even further. In light of the categorization by Mehrabi *et al.* [36], the most relevant biases for this research are briefly discussed below. For a more detailed overview of different sources of bias, the reader is encouraged to read the papers [36], and [51].

1. **Measurement bias**: the bias that arises when the measurement of a feature is inaccurate or when the feature is measured unevenly [36, p.4]. In the context of fairness, the mismeasurement of features could encode bias. For example, performance ratings could be influenced by an individual's perceptions towards others instead of actual performance, and hence could lead to biased predictions.

2. **Representation bias**: this type of bias occurs if the sampled population does not correctly represent the real world [36, p.5]. In the context of this thesis, this could lead to unfair model performance across different groups. This could result in certain subgroups being disadvantaged.

3. **Aggregation bias**: aggregation bias arises when inferences are drawn about an individual based on population level observations. An example of aggregation bias that can occur within the scope of this thesis is that outcomes can seem fair at the gender level. However, at a more granular level, the outcomes might be unfair with respect to, for instance, young women. This example is a case of Simpson's Paradox, which occurs when trends from the population fade when the data is examined within subgroups, or vice versa [36, p.5].

4. **Algorithmic bias**: algorithmic bias emerges only from the algorithm; the bias is not present in the data, but the algorithm still exhibits bias [36, p.6]. Such bias can be

caused by the choices of hyperparameter values and other implementation strategies. The danger of this type of bias can always occur, regardless of the data.

5. **Historical bias**: historical bias stems from systematic biases embedded in the data [36, p.5]. For example, since particular subgroups have been privileged more, this bias is embedded in the data. This bias arises even if the data is measured perfectly. A model trained on such data may learn these patterns and could produce unfair outcomes. Note that historical bias and measurement bias are related, as historical bias can arise from the systematic mismeasurement of certain groups.

2.2 Fairness Metrics

To assess whether a model is fair, one needs to use a definition of fairness. To transfer the idea of fairness to ML models, a substantial number of metrics have been proposed. In total, there are over 20 fairness metrics introduced in the literature, where certain metrics resemble one another [53]. Additionally, the same fairness metric can carry different names, e.g., Demographic Parity (DP) and Statistical Parity Difference (SPD). However, these metrics can be distinguished as group level fairness metrics and individual level fairness metrics.

The fairness metrics introduced in this thesis serve two purposes; firstly, they are needed to grasp the different kinds of fairness metrics and the framework for selecting the appropriate fairness metric in Section 4.2. Secondly, certain fairness metrics introduced in this section are used for evaluation.

To understand the different fairness metrics, notation is introduced first. Aligning with this thesis, binary classification and categorical sensitive attributes are assumed.

- $Y \in \{0, 1\}$: true label, where 1 is the desired output.
- $\hat{Y} \in \{0, 1\}$: predicted label.
- $\mathcal{Y}$: set of possible values for the target variable.
- $\mathcal{S}$: the set of possible values for the sensitive attribute, with s_i being the value of the i-th value in $\mathcal{S}$.
- $\mathcal{G}$: all intersectional subgroups derived from the sensitive attributes, where g_i is the i-th group in $\mathcal{G}$.

In the remainder of this subsection, various important fairness metrics for both categories will be highlighted. Note that the random variable S and specific value s_i are used in formulas when dealing with a single sensitive attribute, and the random variable G and specific group g_i are used in formulas when referring to intersectional subgroups derived from all sensitive attributes jointly.

2.2.1 Group Level Fairness Metrics

Group level fairness metrics focus on disparities between statistical values in the predictions, defining fairness over the population divided into groups.

2.2.1.1 Overall Group Fairness Metrics

The most popular fairness metric is Demographic Parity (DP), which is used in at least 136 publications [24]. Assuming S is a binary sensitive attribute, DP is attained when the probability of being assigned to class 1 is the same for both groups of the sensitive attribute. In mathematical terms, that is $P(\hat{Y} = 1|S = 0) = P(\hat{Y} = 1|S = 1)$. The difference between these probabilities can be used as a measure of disparity, which is often referred to as DP or SPD. The discrimination score by Calders and Verwer [5] is a difference based metric commonly used to quantify DP and is equivalent to SPD/DP. Furthermore, DP naturally extends to a multi-class sensitive attribute setting.

Another similar group based fairness metric is Disparate Impact (DI). DI takes the same quantities as DP; however, instead of taking the difference, DI calculates the ratio. DI can also be straightforwardly extended to multiclass sensitive variables. Furthermore, DI is used to compute the p-rule. The p-rule is defined as: $\min_{g_i,g_j \in \mathcal{G}} (\frac{P(\hat{Y}=1|G=g_i)}{P(\hat{Y}=1|G=g_j)}, \frac{P(\hat{Y}=1|G=g_j)}{P(\hat{Y}=1|G=g_i)})$. In the literature, the p-rule is used with respect to the four-fifths rule, where this quantity needs to be at least 0.8. For example, Foulds *et al.* [21] adjust the level of unfairness up to the point that the p-rule will be 0.8 at a minimum.

Additionally, in a vast amount of research, Equalized Odds (EO) is used as a fairness metric. EO requires access not only to the predicted value $\hat{Y}$ but also to the true label Y. To realize EO, the True Positive Rate (TPR) and False Positive Rate (FPR) must be equal across sensitive attribute groups [23]. EO, in mathematical terms, is defined as: $P(\hat{Y} = 1|S = 1, Y = y) = P(\hat{Y} = 1|S = 0, Y = y) \quad \forall\, y \in \mathcal{Y}$. The differences in TPR and FPR between groups can be used as a fairness metric. The metric can either be reported as the difference in TPR and FPR or as the maximum of the differences in TPR and FPR. For this metric, it holds that it can be rewritten so that it applies to multiclass variables as well.

In the same paper, Equality of Opportunity (EoO) is discussed. EoO is similar to EO; however, it only enforces the true positive rates to be equal for all sensitive attribute groups. EoO can hence be seen as a relaxation of EO.

2.2.1.2 Subgroup Fairness Metrics

The previous section discussed fairness metrics concerning a single sensitive attribute; however, in real world scenarios, multiple sensitive attributes often exist. An example of this is the HR context of this thesis. In more recent years, the field has adopted this vision and proposed subgroup fairness metrics, which capture fairness across different subgroup levels.

Kearns *et al.* [31] recognized that fairness metrics can seem fair overall but can still be unfair for certain subgroups. To prevent this fairness gerrymandering, two traditional fairness metrics are extended. The authors define statistical parity over all subgroups, together with a weighting term that takes the size of the subgroup into account. For a specific $g_i \in \mathcal{G}$, this can be written as:

$$P(G = g_i)(|P(\hat{Y} = 1) - P(\hat{Y} = 1|G = g_i)|).$$

This quantity needs to be below a chosen threshold γ for all subgroups $g_i \in \mathcal{G}$ to be considered fair. Here $P(G = g_i)$ weakens the effect of smaller subgroups. This metric is called Statistical Parity Subgroup Fairness (SPSF). SPSF quantifies the disparity between

the probability of a positive outcome for a certain subgroup compared to the population as a whole.

Furthermore, the authors extend the notion of EO's FPR part in a similar fashion. Their version is called False Positive Subgroup Fairness (FPSF) and for subgroup $g_i \in \mathcal{G}$ is defined as:

$$P(G = g_i, Y = 0)(|P(\hat{Y} = 1|Y = 0) - P(\hat{Y} = 1|G = g_i, Y = 0)|).$$

Similarly, as to SPSF, this quantity needs to be below γ for all $g_i \in \mathcal{G}$ to be considered fair. Moreover, there is again a weighting term and a fairness term measuring overall disparity for a certain subgroup. The weighting term diminishes the impact of subgroups with a low probability of having the class label 0. Since FPR is calculated only for instances with a true label of 0, this weighting term reduces the influence of small subgroups with relatively few observations labeled as 0. The latter term determines the amount of overall disparity in FPR for a subgroup $g_i \in \mathcal{G}$.

Additionally, Foulds *et al.* [21] define Differential Fairness (DF). Where differential fairness weighs all subgroups evenly, thus enforcing fairness across all different subgroups regardless of their size. Let $\theta \in \Theta$ be the distribution that generates the data. A classifier is said to be ϵ-DF fair if the following holds for its predictions:

$$e^{-\epsilon} \leq \frac{P(\hat{Y} = y|G = g_i, \theta)}{P(\hat{Y} = y|G = g_j, \theta)} \leq e^{\epsilon} \quad \forall y \in \mathcal{Y},\ g_i, g_j \in \mathcal{G}.$$

In this definition, the amount of unfairness can be controlled by the value of ϵ.

2.2.2 Individual Level Fairness Metrics

Individual fairness metrics approach fairness from a different standpoint. In the field of individual fairness, instances that are similar should receive similar outcomes [37]. An example of such a fairness metric is Fairness Through Awareness (FTA), presented by Dwork *et al.* [16]. Let f be a classifier, x_i and x_j be the i-th and j-th observations in the dataset X, and $D(\cdot, \cdot)$ be a distance metric for the labels, while $d(\cdot, \cdot)$ is a distance metric for the observations. FTA then requires

$$D(f(x_i), f(x_j)) \leq d(x_i, x_j), \quad \forall x_i, x_j \in X.$$

This inequality states that the difference between predictions should be smaller than or equal to the difference between instances. This implies that individuals with similar features should obtain the same predictions. However, as the authors themselves noted, defining the distance metric for features can be challenging.

Additionally, the counterfactual fairness metric proposed by Kusner *et al.* [32] is a different approach to defining fairness. The proposed fairness metric is strongly associated with causality. The authors say a classifier is counterfactually fair if there is no difference between the outcome of the predictor observed in the real world and a counterfactual world, where only the sensitive attribute and its descendants differ. This definition, however, requires defining how the descendants of the sensitive attributes change when the sensitive attribute is altered. This can be done using the structural equations of a Structural Causal Model (SCM), however, specifying a SCM can prove to be difficult [3]. Furthermore, Barocas *et al.* [3] note that using a structural causal model where the causal assumptions

do not hold can lead to invalid conclusions. Let U be the set of exogenous variables. Furthermore, let $Z = X \setminus S$ denote the set of non-sensitive features. In mathematical terms, a classifier is counterfactually fair if

$$P\Big(\hat{Y}_{G\leftarrow g_i}(U) = y \mid Z = z, G = g_i\Big) = P\Big(\hat{Y}_{G\leftarrow g_j}(U) = y \mid Z = z, G = g_i\Big) \quad \forall y \in \mathcal{Y}, g_i, g_j \in \mathcal{G}.$$

In this equation $\hat{Y}_{G\leftarrow g_j}$ means the predicted label $\hat{Y}$ that would be obtained in a counterfactual world where an instance's group and its descendants are recomputed after changing group membership to g_j while keeping U fixed. The values of the descendants are inferred from the structural equations of the SCM.

2.2.3 Relations Between Fairness Metrics

As there are numerous fairness metrics, this section discusses the relationships between the fairness metrics discussed in this thesis.

2.2.3.1 Equivalences and Differences

The fairness metrics discussed above bear various similarities and differences. For instance, DI takes the same quantities as DP; however, by taking the ratio instead of the difference, the interpretation changes. Moreover, EoO is a direct relaxation of EO. A more significant difference is found between DP and EO. DP does not take the true label Y into account. Furthermore, DP requires that the probability of being assigned to the positive class remains the same, regardless of the sensitive attribute. EO, on the other hand, requires that people with a true positive class and a true negative class have similar classification statistics. This spreads the notion of fairness over TPR and FPR instead of the overall positive predictions.

Additionally, the subgroup fairness metrics closely relate to the overall group based notions. SPSF, for example, slightly tweaks DP so that it can measure unfairness with respect to multiple sensitive attributes. Additionally, FPSF extends EO's FPR part to multiple sensitive variables. On the other hand, DF extends DI to multiple sensitive attributes and is hence also closely related to DP. A classifier that is performing well with respect to DP will likely perform well with respect to DF, as both use the same probabilities. DP takes the difference between probabilities, whereas DF uses a ratio.

On the other hand, individual fairness metrics bear little similarity to (sub)group fairness metrics. The individual fairness metrics discussed require the specification of additional metrics or assumptions, such as a distance metric or an adequate SCM, which can prove to be difficult. Given this, group based fairness metrics tend to be predominantly used in the fairness literature.

Beyond the similarities and differences in the quantities used, fairness metrics can be assigned to one of two additional categories: bias preserving and bias transforming. Bias preserving metrics are always satisfied with a perfect classifier, while bias transforming metrics are not necessarily satisfied with a perfect classifier [54]. In other words, bias preserving metrics can reflect biases present in the data, measuring the extent to which the error rates are equal across sensitive groups. By using the true label, these fairness metrics take into account the underlying data. If this data contains societal biases, these metrics preserve that bias. On the other hand, bias transforming metrics do not use the distribution of the data. Extending the table by Wachter *et al.* [54], Table 1 is obtained.

Metric	Level	Type	Requires Y	Multiclass S
DP	Group	Bias transforming	×	✓
DI/p-rule	Group	Bias transforming	×	✓
EO	Group	Bias preserving	✓	✓
EoO	Group	Bias preserving	✓	✓
SPSF	Subgroup	Bias transforming	×	✓
FPSF	Subgroup	Bias preserving	✓	✓
DF	Subgroup	Bias transforming	×	✓
FTA	Individual	Bias transforming	×	✓
Counterfactual Fairness	Individual	Bias transforming	×	✓

Table 1: Overview and comparison of fairness metrics discussed in this section.

Wachter *et al.* [54] note that bias preserving methods are not inherently wrong to use; however, one needs to understand the difference between these metrics. Bias preserving methods can still be useful in preventing ML models from injecting bias themselves. Furthermore, the authors provide guidance on how to choose which type of fairness metrics to use. This thesis will discuss these guidelines further in Section 4.2.

2.2.3.2 Incompatibilities

Viewing fairness metrics from a different standpoint has led to certain fairness metrics contradicting one another. For instance, Chouldechova [13] showed that although a classifier can seem fair with respect to predictive bias, it may remain unfair concerning FNR and FPR rates. Furthermore, fairness metrics are unable to capture both group and individual fairness concurrently [9]. Additionally, Barocas *et al.* [3] proved that given mild assumptions, for binary classification, DP and EO cannot hold simultaneously.

2.3 The Utility-Fairness Trade-off

As seen earlier, there is a large variety of fairness metrics. Improving fairness with respect to these metrics often comes at the cost of reducing the accuracy of the classifier [9]. Fairness-aware ML techniques often try to find a balance in this trade-off [6, 21, 28, 31]. The challenge in fairness-aware ML thus lies not only in achieving fairness but also in finding an appropriate trade-off between model performance and fairness.

2.4 Fairness-Aware ML Methods

Given the utility-fairness trade-off, various fairness-aware ML models have been proposed to mitigate unfairness while maintaining predictive performance. Moreover, fairness-aware ML is a relatively new research field, with interest in the subject commencing approximately in 2009 [24]. The earliest techniques mostly focus on a single binary sensitive attribute, with little expandability to multivariate multi-class sensitive features. Furthermore, the paper by Chen *et al.* [12] from 2024 notes that the majority of available methods

deal with a single protected attribute at a time. Additionally, in the context of this thesis, not only is the binary sensitive attribute of gender considered, but age and education are also included. This highlights the need for techniques capable of handling several multi-class variables.

In the literature, fairness-aware ML methods are commonly divided into three different categories: pre-processing, in-processing, and post-processing. Pre-processing techniques aim to alter the data before training the model. The goal is to remove unfairness from the data, which should lead to the subsequently trained model being fairer. In-processing methods are designed to enforce fairness during the training of the model. This can be done by enforcing fairness constraints during training, for instance. Lastly, post-processing approaches intervene in the output of the model. These approaches can, for example, determine unfair predictions and modify them to create fairer predictions [9]. As discussed in the introduction, this thesis will not consider post-processing techniques, as prior research found them to be less effective than pre- and in-processing techniques [44].

2.4.1 Pre-processing

As per Hort *et al.* [24], pre-processing is a highly researched fairness-introducing category, with over 100 publications on different methods. These approaches operate by modifying the data before training the model, with the goal of reducing dependencies between sensitive attributes and the target variable.

2.4.1.1 Instance and Label Transformation

The paper by Kamiran and Calders [28] offers three pre-processing techniques and is one of the most widely used benchmarks in the literature. By 2024, it had been used 37 times as a benchmark, making it the second most used benchmark in the field [24]. Their methods focus on a single binary sensitive attribute, but they can be extended to multiple multiclass sensitive features.

The first method introduced is called massaging; massaging involves flipping specific labels in the training dataset. The probabilities of a positive class outcome are used to make a decision on which instances to flip. In addition, a method called reweighing is proposed; here, all instances in the dataset are assigned a weight. These weights are determined by the observed label and group membership. Instances that belong to over-represented group-label combinations, with respect to what is expected, receive a lower weight. On the other hand, instances that belong to under-represented group-label combinations receive a higher weight. Lastly, an alternative sampling technique is proposed, which is designed for cases where reweighing is not applicable. However, since reweighing is directly applicable in this thesis, sampling is not considered.

The methods are tested on several datasets, including the Adult Income dataset, which classifies whether an adult has an income over or under $50,000. For this dataset, the proposed methods are shown to decrease DP without substantially compromising predictive performance. Reweighing yields considerable fairness improvement, obtaining approximately 3% DP as opposed to the 18% baseline discrimination, while the accuracy drops by approximately 1%. A major advantage of these methods is that they are model agnostic. However, flipping the labels in massaging comes with concerns about data integrity.

The methods proposed by Kamiran and Calders [28] offer straightforward techniques to mitigate bias that can be used in intersectional subgroup settings, making them practical baselines. Massaging additionally raises concerns with respect to data integrity, as the flipping of labels introduces artificial patterns in the data.

2.4.1.2 Distribution Transformation

An approach capable of handling multivariate, multi-class sensitive attributes is the method proposed by Calmon *et al.* [6]. In the paper, the probabilistic transformation of the dataset is formulated as a mathematical optimization problem with multiple constraints. The objective function minimizes the difference between the distribution of the original dataset and the transformed dataset. The constraints impose restrictions on, for example, the degree of unfairness. The resulting transformed dataset is then used for training the classifier. However, a disadvantage of the method is that it only works with discrete sensitive features.

A similar feature based drawback arises in the method by Feldman *et al.* [20]. The paper introduces two data repair approaches: Combinatorial and Geometric repair. Combinatorial repair moves the non-sensitive features towards the median in the dataset while preserving the relative rank in the group. Geometric repair works by decreasing the distance between the distributions of the sensitive groups. However, these techniques only work for numeric non-sensitive features.

Such feature type constraints are not compatible with the mixed feature types that this thesis uses and are therefore not considered further.

2.4.1.3 Representation Learning

A conceptually different pre-processing approach is representation learning. Representation learning aims to learn a new, fair representation of the data in a latent space. For example, Zemel *et al.* [56] encode the latent representation such that as much information as possible from the original dataset is preserved while removing information about group membership. Consequently, a classifier is trained on the latent representation of the data. However, as applied, this technique considers only binary sensitive attributes. Nonetheless, this approach is widely regarded as influential and has sparked numerous research studies on similar approaches, such as the method by Madras *et al.* [35].

Newer models have taken representation learning into the field of Deep Learning (DL). With the emergence of these models, more advanced frameworks have been used, such as Variational Autoencoders (VAE) [15, 34, 41]. However, these complex models are generally implemented for a single binary sensitive attribute or require independence assumptions between multiple sensitive attributes. Given these limitations, such models are not considered.

2.4.2 In-processing

This subsection highlights several representative in-processing methods, which are further divided into more granular categories. The methods are particularly discussed for their ability to handle multiple multi-class sensitive attributes.

2.4.2.1 Regularization Based Methods

In-processing methods often make use of regularization or constrained optimization to ensure the fairness of the classifier [16, 30, 55].

Regularization based methods are a straightforward solution to enforce fairness in ML algorithms. However, Kamishima *et al.* [30] only deal with a single binary sensitive variable. Additionally, the method by Kamiran *et al.* [29] operates on non-intersectional group fairness metrics. These limitations are addressed by Foulds *et al.* [21]. The study by Foulds *et al.* [21] explicitly uses regularization with intersectional notions of fairness to ensure subgroup fairness. This approach is particularly relevant to this thesis, given that the method is capable of handling multiple multi-class sensitive attributes. In their paper, the fairness metrics SPSF and DF act as regularizers. These fairness constraints are added to the learning objective and are weighted by a hyperparameter λ. The classifier used is a simple 3-layer neural network. Moreover, the method exhibits strong performance on the Adult Income dataset; the DF decreases from 1.646 to 0.379, while the accuracy only drops by roughly 1.6 percentage points.

2.4.2.2 Adversarial Learning

Adversarial learning draws inspiration from a deep learning architecture, specifically a Generative Adversarial Network (GAN). In the paper by Zhang *et al.* [57] adversarial debiasing is introduced. Adversarial debiasing makes use of a predictor that predicts the target variable Y from the data X. The adversary takes the output of the predictor and attempts to predict the sensitive feature. If the adversary has low accuracy, the model is perceived as fair. Although the proposed method can be easily modified to use multi-class sensitive variables, adversarial debiasing is restricted to a single sensitive attribute.

2.4.2.3 Reductions Based Methods

Another in-processing method that is capable of handling multiple sensitive variables is the technique proposed by Agarwal *et al.* [1]. The authors rewrite the constrained optimization problem in Lagrangian form. Subsequently, they note that this is a zero-sum game and propose an algorithm that solves this problem. The algorithm uses cost sensitive classifiers and exponentiated gradient updates, among other techniques, to iteratively enforce fairness and ensure predictive performance. The obtained randomized classifier lies on the Pareto frontier for a given set of classifiers and a given level of unfairness. Moreover, the method is compatible with multiple classification algorithms and supports several fairness metrics, including DP and EO. Given these advantages of the reductions based approach, the fairness-aware ML model introduced by Agarwal *et al.* [1] is suited for the setting of this research.

2.5 Employee Attrition Prediction

The trade-off between accuracy and fairness plays a significant role in fairness-aware ML. Hence, this section reviews ML approaches used in employee attrition prediction, identifying possible baselines for this thesis.

In the literature on employee attrition prediction, mostly the same set of machine learning methods is researched and evaluated. These include Logistic Regression (LR), Decision

Trees (DT), Random Forest (RF), Extreme Gradient Boosting (XGBoost), Neural Network (NN), and more [19, 26, 58]. Fallucchi *et al.* [19] compared 8 different classification models using 5 different evaluation metrics with the IBM HR dataset. The authors find that a Gaussian Naïve Bayes (GNB) classifier obtained the highest recall and F1 score, 0.541 and 0.446, respectively. GNB, therefore, minimizes false negatives, correctly capturing a substantial proportion of people who are leaving. LR follows closely, with an F1 score of 0.445. Moreover, Raza *et al.* [48] find that the ensemble model Extra Trees Classifier (ETC) performs best for employee attrition prediction on the IBM HR dataset. ETC outperforms LR, obtaining a recall score and F1 score of 0.93, while LR achieves a recall score of 0.72 and an F1 score of 0.73. However, Qutub *et al.* [47] find that ensemble methods, including XGBoost and RF, do not outperform LR with respect to accuracy, recall, and F1 score. Lastly, Zhao *et al.* [58] use three different datasets and find that XGBoost performs, on average, the best with respect to the Receiver Operating Characteristic (ROC). The ROC captures the trade-off between true positive and false positive rates. Additionally, NNs are found to perform solidly overall.

The variety in performances on the same dataset across different studies likely reflects differences in pre-processing the data. Raza *et al.* [48], for instance, use Synthetic Minority Oversampling Technique (SMOTE) and perform feature engineering, while the paper by Fallucchi *et al.* [19] uses neither of these. This likely leads to differences in results for LR and makes comparison difficult. Given the variety of models present and the differences in performance, this thesis will assess LR and XGBoost as base classifiers. These methods are selected due to their compatibility with the fairness-aware ML methods in this thesis, and their consistent appearance in the attrition prediction literature. These baseline models additionally differ, as one is a linear approach and the other is an ensemble-based approach.

None of the employee attrition prediction studies reviewed in this section incorporate fairness considerations. This is particularly concerning given that attrition prediction models can directly inform consequential HR decisions and, hence, potentially are categorized as high-risk systems in the AI Act. This motivates this thesis, which explicitly integrates fairness alongside predictive performance.

2.6 Research Gap

As discussed in this chapter, the majority of fairness-aware ML methods focus on dealing with a single binary sensitive attribute. Although more recent methods are designed for subgroup fairness, the application to HR data remains scarce. Considering that HR data contains several potentially sensitive attributes, such as gender, age, and education, this gap is particularly relevant. While Foulds *et al.* [21] apply intersectional fairness in classification, their method does not compare methods against one another in the employee attrition context. This thesis addresses this gap by applying and comparing pre- and in-processing methods to employee attrition prediction, where group and subgroup level fairness metrics are considered alongside predictive performance. Additionally, this research tests a novel approach, Fair Ensemble, which potentially provides a straightforward way to obtain fairness at the subgroup level. Furthermore, legislation like the EU AI Act can classify an employee attrition prediction model as a high-risk system and impose restrictions on bias mitigation, yet it does not provide guidance on how to achieve it or what level of unfairness is acceptable. This motivates the empirical comparison of fairness mitigation methods.

Chapter 3

Data

This chapter discusses the datasets and their properties. There is limited availability of real-world HR datasets, hence, research on employee attrition prediction is often conducted on the synthetic HR dataset constructed by IBM[1]. This IBM dataset will be used for a robustness check in this thesis. The primary dataset used for training is a second synthetic dataset[2]. Both datasets are designed to aid researchers and HR practitioners in identifying patterns that trigger employee attrition. Moreover, both datasets contain the demographic features required for the fairness analysis of this thesis.

Section 3.1 describes the data, and Section 3.2 discusses which features are selected as sensitive attributes. Section 3.3 describes how the datasets are aligned. Subsequently, Section 3.4 describes the data pre-processing pipeline. Moreover, Section 3.5 shows the Exploratory Data Analysis (EDA), and Section 3.6 highlights the differences between the datasets and their implications. Lastly, Section 3.7 provides the concluding remarks of this chapter.

3.1 Data Description

The reason two datasets are used in this thesis is the limited number of observations in the IBM dataset. The IBM dataset consists of 1,470 instances and 35 features, which is too small for certain learning architectures used in this thesis. Hence, the need for another dataset. The primary dataset has 24 features and comes with a predetermined $80-20$ split between the training and test sets. The train set consists of 59,598 observations, while the test set contains 14,900 instances. This makes the dataset suitable for NNs, for instance. Moreover, the IBM dataset is not used for training. However, since it is often utilized in the literature on employee attrition prediction, using it for a robustness check allows for verification of generalizability.

3.2 The Sensitive Variables

Choosing sensitive variables requires normative assumptions about which attributes need protection. In this thesis, gender, age, and education are treated as sensitive variables since differential treatment based on these attributes may be socially undesirable. Moreover, these attributes are frequently considered sensitive attributes in fairness-aware ML [1, 5, 20, 27, 33].

Gender: In the field of employment, there are known disparities concerning gender, such as the gender wage gap [4]. Another example is that some women may leave due to responsibilities for taking care of a newborn, which can lead to an employment break

[1] Available on Kaggle: https://www.kaggle.com/datasets/pavansubhasht/ibm-hr-analytics-attrition-dataset/data

[2] Available on Kaggle: https://www.kaggle.com/datasets/stealthtechnologies/employee-attrition-dataset/data?suggestionBundleId=428

and eventually a lower wage [17]. The data may contain such factors, and a regular ML model is likely to learn these patterns. Furthermore, the EU Equal Treatment Directive (2006/54/EC) states that discrimination on the grounds of sex is prohibited in employment [18]. Hence, predictions should be fair with respect to the variable of gender.

Age: Age can also be viewed as a sensitive attribute. In HR datasets, such as the IBM dataset, age often correlates with monthly income and job level. A model that uses age may penalize employees for being experienced or nearing retirement, disadvantaging certain demographic groups specifically. Moreover, if a model predicts that older employees have a higher chance of leaving the company due to retirement proximity or higher salaries, organizations might reduce their investments in these employees. This could result in a self-fulfilling prophecy that disadvantages people in a higher age group. Lower investment in an employee could result in that employee leaving the company. This theoretical concern is in line with the statistical discrimination theory from Phelps [45]. Additionally, similar to gender, there is a tendency for age discrimination in HR settings. For example, research shows that age discrimination is associated with higher rates of job separation. Neumark and Johnson [38] found the result to be significant even when controlling for retirement. Furthermore, EU law provides a framework banning discrimination based on age in employment settings [14].

Education: Unlike gender and age, education is not generally considered a protected characteristic in anti discrimination law. However, education is not equally accessible to everyone and, therefore, can be a proxy for socioeconomic status [40]. Consequently, evaluating fairness across education levels allows this thesis to examine whether fairness interventions affect individuals from educational backgrounds differently. Additionally, adding education to the subgroups provides a substantial increase in the number of subgroups, which might uncover disparities that would not be visible from gender and age alone. Therefore, education is included as a sensitive attribute for fairness analysis.

These attributes are not only considered in isolation, but the intersectional groups for these attributes are also analyzed. Combining these attributes leads to more fine-grained subgroups, which can uncover unfairness not visible at the group level [31].

3.3 Aligning Datasets

Since the datasets do not contain exactly the same attributes, the features of the datasets must be aligned to ensure a consistent input space for the model. After aligning the datasets on common columns, both datasets contain 12 columns.

Moreover, the aligned features do not have the same range of values; hence, various features need to be coarsened to make both datasets compatible with the models. The training data has five levels of education: high school, associate degree, bachelor's degree, master's degree, and PhD. However, the IBM dataset contains only 48 observations for education level 5. Hence, the PhD level is grouped together with a master's degree for the primary dataset, and for the IBM dataset education level 4 and 5 are grouped. This leads to a slight loss of information, however, since there are significantly fewer people with a PhD, it makes sense to aggregate this group with the Master's degree group. The resulting encoding also yields a more evenly distributed number of observations, which is desired in the training data, as this gives the model equal exposure to each group. Furthermore, the variable job level consists of three values in the training dataset and five in the IBM dataset. Similarly, this variable is reduced to three values in the IBM dataset.

Lastly, the feature job role was removed, as the job roles defined in the two datasets show insufficient overlap. The aligned datasets provide a consistent feature set for the models.

3.4 Pre-processing the Data

Data pre-processing is important in ML, as model performance depends on the data on which it is trained. The goal of pre-processing is to obtain a suitable dataset for the model to improve performance [22]. Both datasets are synthetic, providing clean data with no missing values and requiring minimal pre-processing. Therefore, the pre-processing consists only of encoding features and removing outliers. The sensitive attribute age is converted from a numerical feature to a categorical feature because most fairness-aware ML methods use categorical sensitive attributes. Furthermore, age is often converted into bins within the employee attrition prediction setting [19]. The three bins are defined as 18-30, 31-45, and 46+. These groups can be viewed as the early stage, middle stage, and end stage of a career. Moreover, the resulting bins yield roughly similar numbers of observations across age bins for the training data, which improves the reliability of subgroup fairness estimates.

Moreover, the target variable attrition is encoded as 0 for leaving and 1 for staying. This aligns with the literature on fairness-aware ML, where the label 1 corresponds to the desired outcome.

Furthermore, the two numeric features, monthly income and distance from home, are tested for outliers using the Interquartile Range (IQR) method. Only the monthly income is identified as containing outliers. The IQR method calculates the first (Q_1) and third (Q_3) quartiles and uses these quartiles to determine whether an observation is considered an outlier. The following formula denotes the computations for the lower and upper bounds.

$$\text{lower bound} = Q_1 - 1.5(Q_3 - Q_1)$$
$$\text{upper bound} = Q_3 + 1.5(Q_3 - Q_1)$$

A datapoint is considered an outlier when its value is either below the lower bound or above the upper bound. The IQR identified and removed 50 outliers from the monthly income attribute in the training dataset. Identifying and removing outliers is performed solely on the training set, as the test set is left intact to evaluate the whole population and to prevent data leakage. However, removing outliers might eliminate unique individuals from certain subgroups, which could reduce the representation of those subgroups and thus introduce bias. Given that only 50 outliers were removed from a training set of $59,598$ observations, this risk is negligible in this setting. Nonetheless, this decision is a choice with potential implications for fairness in the resulting models.

Table 2 describes the attributes of the features in the dataset after pre-processing.

Variable Name	Description	Data Type
Attrition	Whether the employee has left the company	Categorical
DistanceFromHome	Distance from home to workplace	Numerical
Gender	Employee's gender	Categorical
JobLevel	Seniority level	Ordinal
JobSatisfaction	Job satisfaction level	Ordinal
MonthlyIncome	Monthly salary	Numerical
OverTime	Works overtime	Categorical
PerformanceRating	Performance rating (4 levels)	Ordinal
EnvironmentSatisfaction	Work environment satisfaction	Ordinal
MaritalStatus	Marital status (single, married, divorced)	Categorical
Age	Age bins (18-30, 31-45, 46+)	Ordinal
Education	Education level (4 levels)	Ordinal

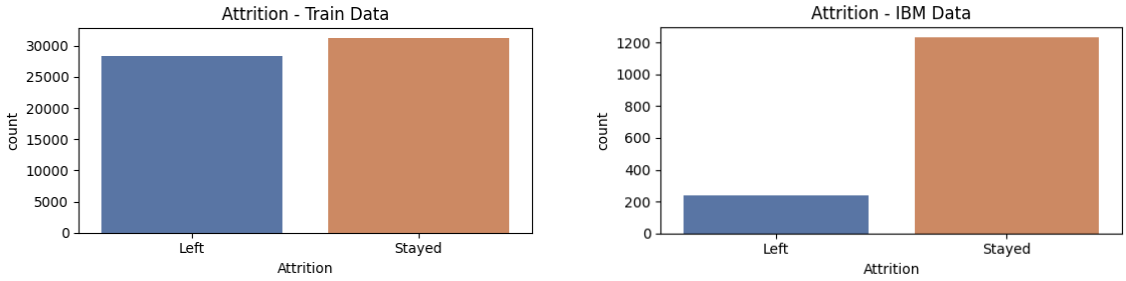
Table 2: Features after pre-processing for both datasets.

MaritalStatus is one-hot encoded, as the three classes have no natural ordering. The ordinal variables are kept with integer representations, as the order can be interpreted and used by the classifiers. The pre-processed data is used for EDA in the next section.

3.5 Exploratory Data Analysis

This EDA section provides plots and tables that aid in understanding the data and identifying its distribution.

Figure 1 shows the distribution of attrition for the training dataset and the IBM dataset. The training dataset shows little difference between the proportion of leavers and the proportion of stayers. Such a distribution is desired so that the models have an equal amount of data for both outcomes, which helps the models learn patterns associated with both outcomes. On the other hand, the IBM dataset exhibits a substantial class imbalance, which may affect the stability of performance and fairness metrics, especially for groups that have a relatively small number of attrition cases.



(a) Training dataset

(b) IBM dataset

Figure 1: Comparison of attrition distributions across datasets.

In Table 3, the distribution of the sensitive variables and the attrition rate for all sensitive variables are shown. Moreover, the approximately even distribution of the age bins in the training data is in line with the binning discussed in Section 3.4. This equal distribution is not present in the IBM dataset. However, since the model is trained on the training data, an approximately equal distribution is of greater importance there.

Furthermore, the attrition rate varies within each sensitive variable. For the training data, each sensitive attribute has a difference of at least 5 percentage points, with the biggest difference being 10 percentage points for gender. Such a substantial disparity in attrition rates may be learned by baseline ML models. Hence, a model trained on the training dataset is expected to show unfairness with respect to gender. Yet, the IBM dataset shows little difference for gender and education, where the maximum difference for age is 15 percentage points. The direction of the disparity changes. For the IBM dataset, the attrition rate is higher for males compared to females, highlighting the difference between the datasets. Therefore, a model trained on the IBM data is likely to exhibit unfairness with respect to age. Both datasets, therefore, exhibit differing disparities in attrition rates across the sensitive attributes, which may affect generalizability.

Additionally, the maximum deviation from the overall attrition rate is 5 percentage points for the training dataset and 10 percentage points for the IBM dataset. Furthermore, the average deviation is only 2.56 percentage points for the training dataset and 2.89 percentage points for the IBM dataset, which is modest. These averages at the group level can mask disparities present in intersectional subgroups, which are discussed next.

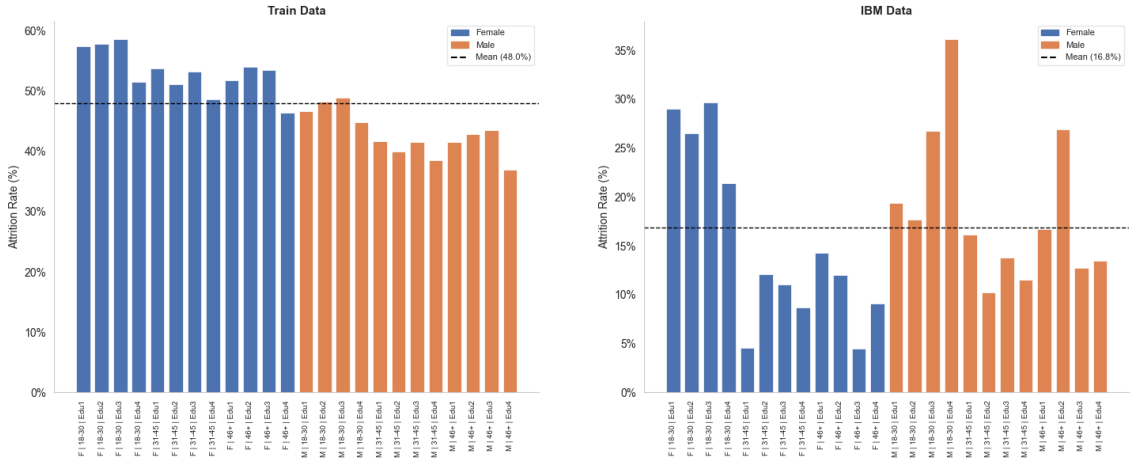
Attribute	**Group**	**Train dataset**		**IBM dataset**	
		Share (%)	**Attrition Rate (%)**	**Share (%)**	**Attrition Rate (%)**
Total	-	-	48	-	16
Gender	Male	55	43	60	17
	Female	45	53	40	15
Age	18–30	33	51	30	26
	31–45	34	45	50	11
	46+	33	46	19	12
Education	Level 1 (High School)	20	48	12	18
	Level 2 (Associate degree)	25	48	19	16
	Level 3 (Bachelor)	30	49	39	17
	Level 4 (Master & PhD)	25	44	30	14

Table 3: Distribution of sensitive attributes and attrition rate per group across both datasets.

To evaluate the properties of the intersectional subgroups, Figure 2 is presented. Figure 2 shows the attrition rate for each subgroup. The subgroups are encoded as Gender | Age | Education, where F | $18-30$ | Edu1 means females aged $18-30$ with education level 1. Particularly remarkable is the systematic difference in attrition rates between males and females in the training set, which is in line with Table 3. However, there are also differences within the male and female subgroups, indicating the need for intersectional subgroup fairness. For the IBM data, the attrition rate for the age bin $18-30$ is systematically above the mean attrition rate, which aligns with the IBM data in Table 3. In addition, the IBM data also shows differences within sensitive variables. For example, there is a 14.9

percentage points difference between males aged 46+ with education level 2 and females aged 46+ with education level 2. Additionally, Figure 2 shows that the IBM dataset has extreme outliers; for example, the subgroup of males aged $18-30$ with education level 4 has an attrition rate that is two times larger than the dataset average. This demonstrates large disparities within sensitive attributes and amplifies the need for intersectional subgroup analysis.

The findings from Table 3 and Figure 2 illustrate that the data can appear approximately balanced at the overall sensitive attribute level but imbalanced at the subgroup level. This indicates that, in this setting, a model trained to prevent unfairness at the overall level might overlook unfairness at the subgroup level, further motivating the intersectional analysis that this thesis conducts.



(a) Training dataset (b) IBM dataset

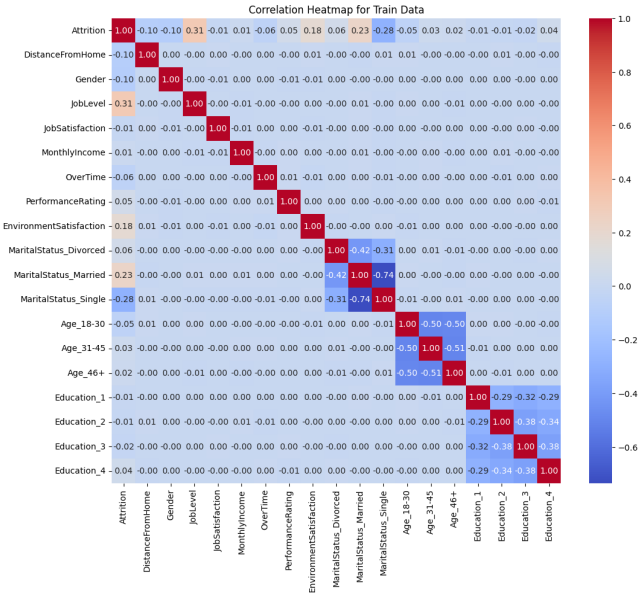
Figure 2: Intersectional subgroups attrition rate across both datasets.

Figure 3 shows the correlation heatmap of the training dataset and the IBM dataset. For a more extensive analysis, the multiclass sensitive variables are one-hot encoded. The correlation heatmap of the training data shows that job level and marital status have a large absolute correlation with attrition. Moreover, in the IBM dataset, the age bin $18-30$ and overtime have a strong negative correlation with attrition. These differences suggest that the datasets represent different parts of the true population.

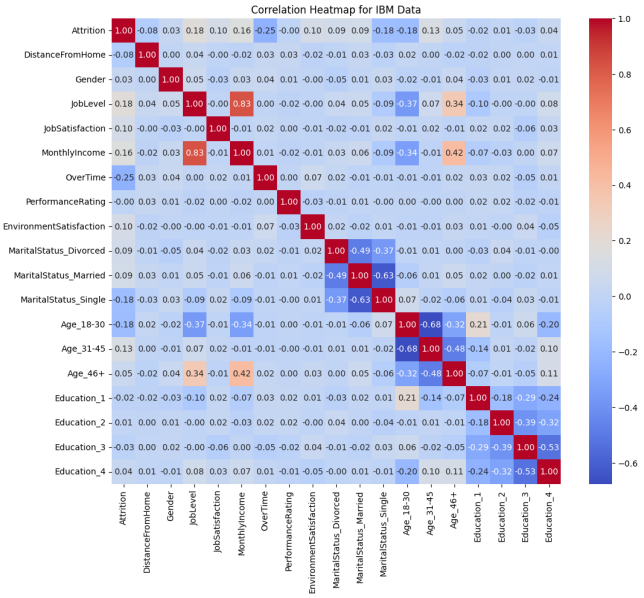
Table 4 reports the mean of key job related features per subgroup. The job-related features are selected based on the correlation heatmaps shown in Figure 3, as well as prior literature identifying key predictors for employee attrition [19]. Additionally, features that are unlikely to serve as a proxy for sensitive attributes, such as distance from home, are not evaluated in the table. Table 4 shows approximately equal values across all subgroups. This reveals that the values of these job-related variables cannot be used as proxies for sensitive attributes. Hence, this suggests that unfairness through the use of proxies is unlikely for a model trained on the training data. Nonetheless, for the IBM dataset, these values vary significantly, as can be seen in Table 23 in the Appendix. Hence, a model trained on the IBM dataset could discriminate even if the sensitive variables were removed.

Additionally, n is the number of observations for each subgroup. In the training data, the number of observations is at least $1,700$ for each subgroup, making the calculations

of fairness metrics for these subgroups reliable. On the other hand, for the IBM dataset, certain subgroups have only 7 observations. This could lead to unstable computations for these subgroups, therefore, the outcomes for these subgroups should be handled with care. Lastly, in the Appendix, one can find similar figures and tables for the test set, which exhibit patterns similar to those of the training data.



(a) Training dataset



(b) IBM dataset

Figure 3: Correlation heatmap for both datasets.

Gender	Age	Edu.	n	Avg Inc. ($)	Avg Job Lvl	OT (%)	Avg Env. Sat.
F	18–30	1	1,798	7,314	1.82	33.4	2.59
F	18–30	2	2,241	7,363	1.78	30.7	2.59
F	18–30	3	2,591	7,230	1.82	33.4	2.60
F	18–30	4	2,192	7,311	1.81	33.1	2.58
F	31–45	1	1,729	7,390	1.79	32.4	2.61
F	31–45	2	2,224	7,278	1.79	33.1	2.61
F	31–45	3	2,716	7,284	1.80	33.1	2.60
F	31–45	4	2,348	7,311	1.77	33.6	2.57
F	46+	1	1,769	7,189	1.80	32.1	2.59
F	46+	2	2,264	7,317	1.78	33.4	2.58
F	46+	3	2,681	7,287	1.79	32.5	2.60
F	46+	4	2,280	7,338	1.80	32.5	2.60
M	18–30	1	2,116	7,304	1.80	30.9	2.59
M	18–30	2	2,700	7,326	1.80	31.9	2.61
M	18–30	3	3,228	7,316	1.80	33.5	2.57
M	18–30	4	2,711	7,203	1.80	32.8	2.59
M	31–45	1	2,171	7,274	1.81	34.2	2.62
M	31–45	2	2,838	7,337	1.82	31.1	2.59
M	31–45	3	3,307	7,318	1.82	32.8	2.60
M	31–45	4	2,774	7,303	1.79	33.3	2.61
M	46+	1	2,157	7,314	1.80	34.4	2.60
M	46+	2	2,638	7,285	1.80	32.1	2.60
M	46+	3	3,288	7,271	1.79	31.2	2.60
M	46+	4	2,787	7,271	1.81	32.3	2.64

Note: Edu. denotes education level (1 = high school, 2 = associate degree, 3 = bachelor's degree, 4 = master's/PhD). n is the number of observations. Avg Inc. is mean monthly income. Avg Job lvl is mean job level. OT is the percentage of overtime. Avg Env. Sat. is the average environment satisfaction.

Table 4: Intersectional group job characteristics for the training data.

3.6 Dataset Differences

The most notable difference between the two datasets is the class imbalance for the IBM dataset. This imbalance implies that comparisons need to be made carefully and that robust metrics for class imbalance need to be used, such as the F1 score. Furthermore, the differences in performance metrics and fairness metrics arise at least partially due to class imbalance. The datasets have different distributions for the variables monthly income and distance from home. The training dataset has a mean of $7,297$ and 50 for monthly income and distance from home, respectively, while the IBM dataset has a mean of $6,503$ and 9, respectively (see Table 24 in the Appendix for more details). A model trained on the training dataset and evaluated on the IBM dataset will systematically see lower values for distance from home compared to training. The same holds for monthly income. Given

these differences, results on the IBM dataset should be interpreted with care and solely as a robustness check.

3.7 Concluding Remarks on the Data

To conclude, the two datasets differ in several important aspects. The IBM dataset exhibits substantial class imbalance, while the training data has an approximately balanced target distribution. Moreover, the datasets show different disparities in the attrition rate across various sensitive variables, both in magnitude and direction. The differences with respect to the overall attrition rate may appear small for the sensitive attributes. However, the imbalances appear substantial when examining the intersectional level. Hence, a standard ML model is expected to learn and reproduce the disparities in the data, further justifying the need for fairness-aware ML discussed in the next chapter. Additionally, the disparities in the data justify the use of the selected sensitive variables. Moreover, in the training data, job related features show approximately equal distributions, reducing the risk of proxy discrimination. Furthermore, the type of bias in the data is difficult to determine, but this section shows disparities present in the data. Given that these findings are consistent with stereotypical biases, such as the gender pay gap [4], they suggest the presence of historical bias in the training data. However, the datasets are synthetically generated, and bias may therefore be a consequence of the data generation rather than historical bias. Nonetheless, the magnitude of the disparities motivates the use of fairness-aware ML regardless of their origin. Lastly, the differences between the primary dataset and the IBM dataset make the robustness check meaningful. A model that performs consistently across both datasets is robust to variations in the data.

Chapter 4

Methods

In this chapter, the methods and metrics to mitigate and measure fairness are discussed, additionally the baseline models are described. First Section 4.1 introduces notation. Section 4.2 discusses the choice and explanation of the formulas for the fairness metrics. After that, Section 4.3 discusses the fairness-aware ML models used, where the first three are existing methods, and the last Fair Ensemble model is a novel contribution proposed in this thesis. Thereafter, Section 4.4 provides the baseline prediction models, and lastly, Section 4.5 states the evaluation metrics used in this thesis.

4.1 Notation

Since there are numerous overlapping notations in this chapter, this section states the most common symbols to prevent confusion. Additional notation is already defined in Section 2.2.

- X: the dataset
- x_i: instance i in X
- y_i: the true label of instance i
- N: the total number of observations

4.2 Fairness Metrics

In this section, the fairness metrics chosen for evaluating model fairness are provided. This section will proceed by first determining which methods to use. Subsequently, the formulas and interpretations for the fairness metrics are discussed.

As mentioned in Section 2.2, Wachter *et al.* [54] propose a set of four questions that act as a guideline to choose the right fairness metrics for evaluation. Aligning with the framework, Figure 4 exhibits questions and their consequences.

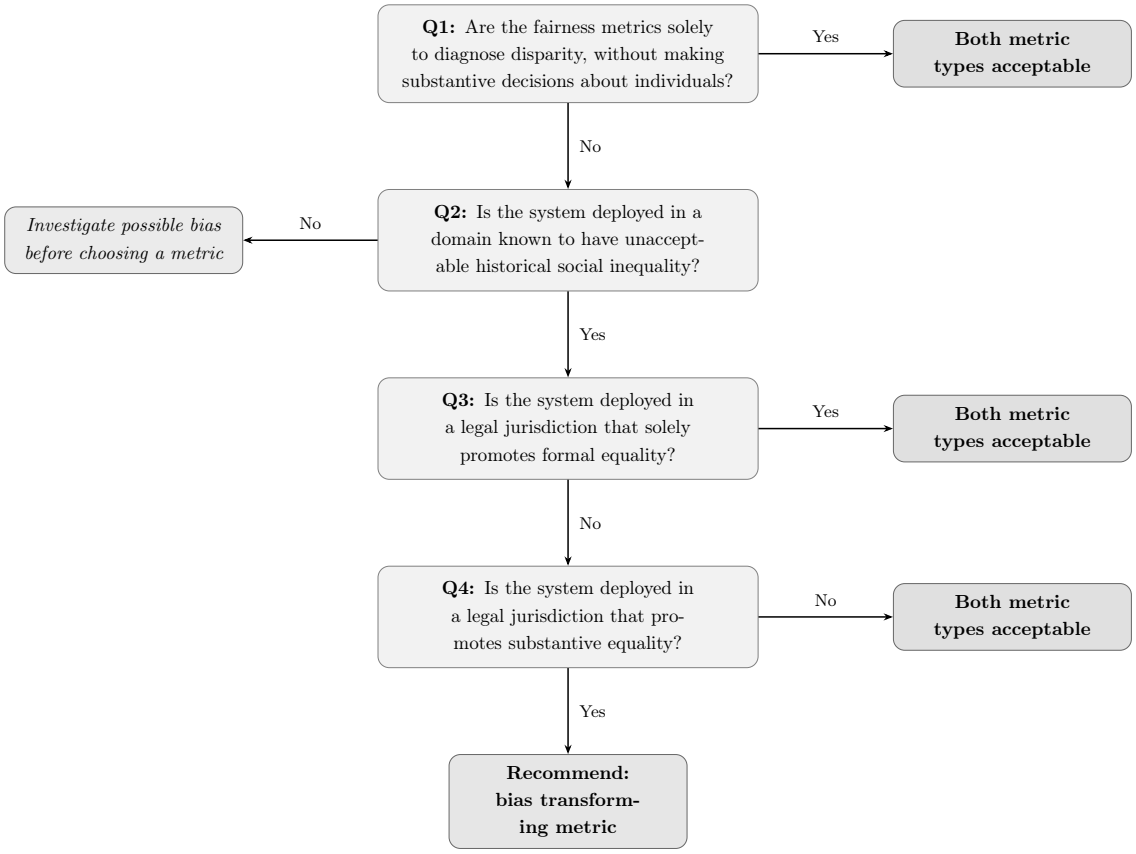


Figure 4: Decision framework for selecting fairness metrics, adapted from Wachter *et al.* [54].

In light of this thesis, these questions are answered accordingly. Question 1 is answered with 'No' since this thesis aims to provide insights for EY. Hence, EY may use the findings from this thesis. Therefore, the goal of this thesis is not to exclusively diagnose disparity. Consequently, question 2 is answered with 'Yes', HR data is known to consist of historical biases, such as the gender pay gap [4]. Accordingly, questions 3 and 4 are answered with 'No' and 'Yes', respectively. As the authors note, the European Union (EU) promotes substantive equality through instruments such as the EU non-discrimination law. Moreover, article 21 of the Charter of Fundamental Rights of the European Union explicitly mentions that no discrimination based on gender or age is allowed. Article 23 states, "Equality between women and men must be ensured in all areas, including employment, work, and pay."[10]. Additionally, according to the AI Act, a system that is designed to aid in decision-making on the termination of contracts or in monitoring and evaluating the performance of employees is regarded as high risk [49]. High risk systems have to adhere to extra requirements. This framework for selecting fairness metrics thus recommends the use of bias transforming fairness metrics. DP is chosen because it is the most common fairness metric used in the literature [24]. Moreover, DP and DF use the same quantities. However, DP takes the difference between probabilities, making it easier to interpret. Hence, DP is selected as a fairness metric. Additionally, SPSF is specifically designed for evaluating fairness in intersectional subgroups. SPSF has a similar interpretation to DP but weighs

subgroups by size. By using this weighing term, it ensures that the fairness metric is not dominated by values obtained for subgroups with a small number of observations, making it more stable in such cases. This is particularly useful for evaluating the IBM dataset. Moreover, unlike DP, which compares the probabilities of two subgroups, SPSF compares the probability given group membership to the overall probability. Since SPSF provides a complementary perspective, it is also selected to evaluate fairness.

4.2.1 Demographic Parity

DP is obtained when the probability of being assigned to class 1 is the same across all groups. Let $\hat{Y}$ be the prediction of a classifier; in the case of a binary sensitive attribute S, DP is defined as

$$P(\hat{Y} = 1|S = 0) = P(\hat{Y} = 1|S = 1).$$

Here, the measurement of DP is

$$P(\hat{Y} = 1|S = 0) - P(\hat{Y} = 1|S = 1). \tag{4.1}$$

For intersectional subgroups, let G consist of k subgroups, this expression then naturally extends to

$$P(\hat{Y} = 1|G = g_1) = P(\hat{Y} = 1|G = g_2) = ... = P(\hat{Y} = 1|G = g_k).$$

Here, $g_i \in \mathcal{G}\ \forall i \in \{1, 2, ..., k\}$. This definition of DP requires that the probabilities of being assigned to class 1 are the same for all k subgroups. The quantity to measure the extent to which this fairness metric is violated is

$$\max_{g_i, g_j \in \mathcal{G}} (|P(\hat{Y} = 1|G = g_i) - P(\hat{Y} = 1|G = g_j)|). \tag{4.2}$$

This extended version of DP (intersectional DP) loses its direction. Formula 4.1 has either a positive or a negative sign indicating the direction of unfairness. However, if S is multiclass, the maximum absolute difference in probabilities is taken. Hence, there is no direction of unfairness. The formula can be modified such that no absolute values are used; however, the direction then becomes difficult to interpret. To grasp this, let the maximum difference in probability be between the subgroup of females aged $18-30$ and females aged $31-45$. The sign of the metric would indicate which one of these two groups is more over-represented among instances with label 1; however, the metric does not tell us anything about the relation between females aged $18-30$ and males aged $18-30$. The metric would identify the most under-represented group with label 1, but the relationship to all other groups would be unclear. Therefore, a signed metric has limited added value in the context of this thesis, which is why the maximum with absolute value is used as the metric. In Equation 4.2, a high value indicates worse fairness levels for a classifier. Moreover, this metric measures the worst-case scenario, as every pair of subgroups other than the worst-performing one is ignored.

4.2.2 Differential Fairness

Although DF is not used to evaluate fairness, it is used to enforce fairness in the method by Foulds *et al.* [21] which is analyzed in this thesis. Therefore, DF is extensively discussed and rewritten in this subsection.

Differential fairness measures intersectional fairness and gives every subgroup equal weight, seeking fairness for all subgroups. Let Θ be the set of distributions that can generate the data. Foulds *et al.* [21] define a classifier to be ϵ-DF fair if

$$e^{-\epsilon} \leq \frac{P(\hat{Y} = y|G = g_i, \theta)}{P(\hat{Y} = y|G = g_j, \theta)} \leq e^{\epsilon}, \quad \theta \in \Theta, \quad \forall y \in \mathcal{Y},\ g_i, g_j \in \mathcal{G}.$$

However, since $P(\hat{Y} = y|G = g_j, \theta)$ is unknown in practice, the authors estimate DF by using estimates of this probability. This probability is estimated using the outcomes of a classifier on the data. That is, $P(\hat{Y} = y|G = g_i, \theta)$ is estimated by $\frac{N_{y,g_i}}{N_{g_i}}$. Here, N_{y,g_i} is the number of observations in subgroup g_i with predicted label y. Moreover, N_{g_i} is the number of observations for subgroup g_i. Subsequently, the definition is simplified to Empirical Differential Fairness (EDF), defined as

$$e^{-\epsilon} \leq \frac{\left(\frac{N_{y,g_i}}{N_{g_i}}\right)}{\left(\frac{N_{y,g_j}}{N_{g_j}}\right)} \leq e^{\epsilon}$$

$$= e^{-\epsilon} \leq \frac{N_{y,g_i}}{N_{g_i}} \frac{N_{g_j}}{N_{y,g_j}} \leq e^{\epsilon}.$$

Since the number of observations in intersectional subgroups can be small, empirical counts can produce unreliable results. The authors address this issue by assuming a symmetric Dirichlet prior for the probabilities per group. The posterior predictive then provides smoothed probability estimates. Given this symmetric Dirichlet concentration parameter α, the smoothed version becomes

$$e^{-\epsilon} \leq \frac{N_{y,g_i} + \alpha}{N_{g_i} + |\mathcal{Y}|\alpha} \frac{N_{g_j} + |\mathcal{Y}|\alpha}{N_{y,g_j} + \alpha} \leq e^{\epsilon},$$

and is called hard counts DF. Here, $\hat{P}(\hat{Y} = y|G = g_i, \theta) = \frac{N_{y,g_i}+\alpha}{N_{g_i}+|\mathcal{Y}|\alpha}$. Moreover, for imbalanced datasets and imbalanced predictions, the hard counts DF can become unstable. To address this issue, Foulds *et al.* propose using raw probabilities instead of hard counts. Let $G(x_l)$ be the group instance x_l belongs to. The soft counts DF is then written as:

$$e^{-\epsilon} \leq \frac{\sum_{l:\, G(x_l)=g_i} P(y|x_l) + \alpha}{N_{g_i} + |\mathcal{Y}|\alpha} \frac{N_{g_j} + |\mathcal{Y}|\alpha}{\sum_{l:\, G(x_l)=g_j} P(y|x_l) + \alpha} \leq e^{\epsilon}.$$

Here, $P(y|x_l)$ denotes the probability of being assigned to class y given the instance x_l. Thus, for soft counts $\hat{P}(\hat{Y} = y|G = g_i, \theta) = \frac{\sum_{l:\, G(x_l)=g_i} P(y|x_l)+\alpha}{N_{g_i}+|\mathcal{Y}|\alpha}$. Moreover, to measure unfairness, the maximum of the logarithm is taken, which results in a single ϵ value as follows:

$$\max_{g_i, g_j} (\log \left(\frac{\hat{P}(\hat{Y} = y|G = g_i, \theta)}{\hat{P}(\hat{Y} = y|G = g_j, \theta)} \right)), \quad \forall y \in \mathcal{Y}. \tag{4.3}$$

The maximum value again deals with the most adverse outcome. Additionally, the value of Formula 4.3 has an intuitive interpretation if $\epsilon = 0.69$, there is a subgroup that has a $e^{0.69} \approx 2$ times larger probability of receiving label 1 compared to another group.

Furthermore, Foulds *et al.* [21] state that DF is robust to Simpson's Paradox. That means a classifier that is ϵ-DF fair with respect to all subgroups is also ϵ-DF fair with respect to a subset of these subgroups.

4.2.3 Statistical Parity Subgroup Fairness

SPSF from Kearns *et al.* [31] measures the disparity between being assigned to class 1 overall and being assigned to class 1 given that the instance belongs to a certain subgroup. Moreover, this disparity is then weighted by the probability of belonging to subgroup g_i. A classifier is then defined to be γ-SPSF fair if the following holds.

$$P(G = g_i)(|P(\hat{Y} = 1) - P(\hat{Y} = 1|G = g_i)|) \leq \gamma, \qquad \forall g_i \in \mathcal{G}.$$

Here, $\gamma \in [0, 1]$. To be compatible with the fairness method of Foulds *et al.* [21], SPSF is also rewritten. For intersectional subgroup i:

$$\begin{aligned} &|P(\hat{Y} = 1|\theta) - P(\hat{Y} = 1|G = g_i, \theta)|P(G = g_i|\theta) \\ =&|P(\hat{Y} = 1|\theta)P(G = g_i|\theta) - P(\hat{Y} = 1|G = g_i, \theta)P(G = g_i|\theta)|. \end{aligned}$$

Additionally, for hard counts, the estimate for SPSF is

$$\begin{aligned} \hat{P}(G = g_i)(|\hat{P}(\hat{Y} = 1) - \hat{P}(\hat{Y} = 1|G = g_i)|) = \\ |\frac{N_1 + \alpha}{N + |\mathcal{Y}|\alpha}\frac{N_{g_i} + \alpha}{N + |\mathcal{Y}|\alpha} - \frac{N_{1,g_i} + \alpha}{N_{g_i} + |\mathcal{Y}|\alpha}\frac{N_{g_i} + \alpha}{N + |\mathcal{Y}|\alpha}|, \end{aligned}$$

Here, N is the total number of observations, N_{1,g_i} is the number of observations for the sensitive group g_i and predicted label 1, and N_1 is the number of observations with predicted label 1. Moreover, N_{g_i} is the number of observations with group membership g_i, and $|\mathcal{Y}|$ is the number of classes of the target variable. To obtain the soft counts estimates, the empirically obtained probabilities are plugged in

$$\begin{aligned} \hat{P}(G = g_i)(|\hat{P}(\hat{Y} = 1) - \hat{P}(\hat{Y} = 1|G = g_i)|) = \\ |\frac{N_{g_i} + \alpha}{N + |\mathcal{Y}|\alpha}\frac{\sum_l P(y = 1|x_l) + \alpha}{N + |\mathcal{Y}|\alpha} - \frac{\sum_{l:G(x_l)=g_i} P(y = 1|x_l) + \alpha}{N_{g_i} + |\mathcal{Y}|\alpha}\frac{N_{g_i} + \alpha}{N + |\mathcal{Y}|\alpha}|. \end{aligned}$$

The measure of unfairness then becomes

$$\max_{g_i \in \mathcal{G}}(\hat{P}(G = g_i)|\hat{P}(\hat{Y} = 1) - \hat{P}(\hat{Y} = 1|G = g_i)|).$$

Similar to intersectional DP and DF, SPSF takes the worst case scenario into account and ignores all other subgroups. Moreover, the weighting term of SPSF leads to scarce subgroups receiving a smaller weight and thus being regarded as less important. This may underestimate fairness with respect to certain groups, as the lack of data for subgroups often leads to bias [54] in the predictions. However, this weighting term is needed for Kearns *et al.* [31] to provide generalization guarantees. Moreover, this metric is still included since it complements the other fairness metrics. Intersectional DP and DF both take the worst case into account, whereas SPSF weights by group size. Analyzing both intersectional DP and SPSF gives a more complete picture of fairness.

4.3 Fairness Methods

Having established how fairness is measured, this section discusses four fairness-aware ML mechanisms. First, the pre-processing method Reweighing method is discussed; this method is well known in the fairness-aware ML literature. Reweighing provides a straightforward benchmark for fairness-aware ML. Subsequently, the in-processing technique DF method is discussed, which deals with intersectional subgroups directly. Additionally, the in-processing method Reductions Based Approach is discussed, which provides an elegant way of calculating fair predictions. Lastly, a novel in-processing technique is introduced: the Fair Ensemble. This Fair Ensemble provides an intuitive method that works with a broad range of fairness-aware ML methods.

4.3.1 Reweighing

Reweighing is a pre-processing approach introduced by Kamiran and Calders [28], in which a weight is given to each observation in the dataset. A classifier allowing for weighted observations is trained on the subsequent weighted dataset. Reweighing is particularly useful if bias in predictions arises from structural inequalities represented in the data, as it actively tries to achieve a balanced dataset. It starts by assuming that group membership and the label should be statistically independent; in mathematical terms, that is:

$$P_{exp}(Y = y, G = g_i) = P(Y = y)P(G = g_i) = \frac{N_{g_i}}{N}\frac{N_y}{N},$$

where N_y is the number of observations with label y. However, in a biased dataset, this equation often does not hold. The observed probability is defined as

$$P_{obs}(Y = y, G = g_i) = \frac{N_{y,g_i}}{N}.$$

Here N_{y,g_i} is the number of observations with true label y, and group membership g_i. Reweighing uses the expected probability and the observed probability to assign lower weights to instances that are overrepresented relative to the expected probability. That is, observations belonging to group-label combinations that have a higher probability than the expected probability are assigned a lower weight, while combinations that have a lower probability than the expected probability obtain a higher weight. By assigning weights, the classifier puts less emphasis on over-represented group-label combinations and more on under-represented group-label combinations. This leads to a more balanced dataset, which should, in turn, result in fairer predictions.

The weighting term is calculated by the ratio of the expected probability to the observed probability. The weighting term for instance i thus is

$$w_i = \frac{P_{exp}(Y = y, G = g_i)}{P_{obs}(Y = y, G = g_i)}.$$

After applying reweighing, the following holds for the resulting dataset:

$$P_{exp}(Y = y, G = g_i) = P_{obs}(Y = y, G = g_i) = P(Y = y)P(G = g_i), \quad \forall g_i \in \mathcal{G}.$$

This means that G and Y are statistically independent. To show this, let the weight of group (y, g_i) be denoted by $w(y, g_i)$. The weighted sum of group (y, g_i) is then written as

$$\begin{aligned}
&\sum_{j:Y_j=y,G_j=g_i} w(y, g_i) = N_{y,g_i} w(y, g_i) \\
&= N_{y,g_i} \frac{P_{exp}(Y = y, G = g_i)}{P_{obs}(Y = y, G = g_i)} \\
&= N_{y,g_i} \frac{\left(\frac{N_{g_i}}{N} \frac{N_y}{N}\right)}{\left(\frac{N_{y,g_i}}{N}\right)} \\
&= N_{y,g_i} \frac{N}{N_{y,g_i}} \left(\frac{N_{g_i}}{N} \frac{N_y}{N}\right) \\
&= N P_{exp}(Y = y, G = g_i) = N P_{obs}(Y = y) P_{obs}(G = g_i).
\end{aligned}$$

To determine the probability of belonging to group g_i with label y, this weighted count needs to be normalized. That is done as follows:

$$P_w(Y = y, G = g_i) = \frac{N P_{obs}(Y = y) P_{obs}(G = g_i)}{\sum_{g'} \sum_{y'} N P_{obs}(Y = y') P_{obs}(G = g')},$$

where $\sum_{y'} P_{obs}(Y = y') = 1$ and $\sum_{g'} P_{obs}(G = g') = 1$. Hence, the equation reduces to

$$\frac{N P_{obs}(Y = y) P_{obs}(G = g_i)}{\sum_{g'} \sum_{y'} N P_{obs}(Y = y') P_{obs}(G = g')} = P_{obs}(Y = y) P_{obs}(G = g_i) = P_{exp}(Y = y, G = g_i).$$

Therefore, in the weighted dataset, G is statistically independent of Y. With this weighted dataset, a classifier that can handle sample weights can be trained. Such classifiers include LR and XGBoost.

4.3.2 Differential Fairness Method

The paper by Foulds *et al.* [21] proposes a straightforward approach to mitigate unfairness accounting for intersectional subgroups. A classifier is trained where the objective function incorporates a penalty term. The Feed Forward Neural Network (FFNN) classifier is used with ReLu activation functions. The outputs are logits, and the binary cross entropy loss function with logits is used as the primary loss function. The additional penalty for unfairness is weighted with the hyperparameter λ, such that the binary cross entropy loss function becomes

$$-\frac{1}{N} \sum_{j=1}^{N} [y_j \log(p_j) + (1 - y_j) \log(1 - p_j)] + \lambda F(\epsilon_{\text{base}}). \tag{4.4}$$

Here, y_j is the true label of instance j, and p_j is the corresponding probability of being assigned to class 1. Moreover, $F(\epsilon_{base})$ is the fairness penalty, where ϵ_{base} determines the extent to which unfairness is deemed acceptable. Therefore, to ensure a nonnegative penalty term that does not result in a negative penalty, $F(\epsilon_{\text{base}}) := max(0, \text{DF} - \epsilon_{\text{base}})$, with DF being the largest disparity of DF between two subgroups.

Moreover, the authors propose a burn-in phase where the FFNN is trained for a predetermined number of epochs using the binary cross entropy loss function with logits. The burn-in phase is used so that the fair model already has accurate predictions before the fairness objective is introduced.

4.3.3 Reductions Based Approach

In contrast to the DF method, the reductions based approach proposed by Agarwal *et al.* [1] frames fairness as a constrained optimization problem. The objective is to minimize the error of a classifier f, subject to the fairness constraint. The authors note that DP can be formulated as a set of equality constraints in the following way:

$$\begin{aligned} \mu_{g_i}(f) - \mu_*(f) &\leq 0 \\ \mu_{g_i}(f) - \mu_*(f) &\geq 0. \end{aligned} \tag{4.5}$$

Here, $f(X)$ denotes the vector of predictions $\{f(x_j)\}_{j=1}^N$. Moreover, $\mu_{g_i}(f) = \mathbb{E}[f(X)|G = g_i]$, $\mu_*(f) = \mathbb{E}[f(X)]$, and X is the dataset. Let $\mathbb{K} = \mathcal{G} \times \{0, 1\}$, which gives the set two elements for each sensitive attribute, one per constraint. Furthermore, let $\mathbb{J} = \mathcal{G} \cup \{pop\}$, which adds the element population to the set of intersectional subgroups. With this notation, the matrix form of the constraints becomes:

$$\mathbf{M}\boldsymbol{\mu}(f) \leq \mathbf{c}.$$

In this factorized form, $\boldsymbol{\mu}(f) \in \mathbb{R}^{|\mathbb{J}|\times 1}$, and when multiplied with $\mathbf{M}$ captures both inequalities from 4.5. Additionally, $\mathbf{c} \in \mathbb{R}^{|\mathbb{K}|\times 1}$. Moreover, $\mathbf{M} \in \mathbb{R}^{|\mathbb{K}|\times|\mathbb{J}|}$ is a matrix containing entries in $\{-1, 0, 1\}$ such that, when multiplied with $\boldsymbol{\mu}$, the inequality constraints are obtained. The values of $\mathbf{M}$ are defined as $M_{(g_i,1),g'} = \mathbf{1}(g_i = g')$, $M_{(g_i,0),g'} = -\mathbf{1}(g_i = g')$, $M_{(g_i,1),pop} = -1$, and $M_{(g_i,0),pop} = 1$, with $\mathbf{1}$ being the indicator function.

Given these constraints, the problem of finding a fair classifier is defined by

$$\min_{f\in\mathcal{F}} Loss(f) \qquad s.t. \quad \mathbf{M}\boldsymbol{\mu}(f) \leq \mathbf{c}. \tag{4.6}$$

Here, $\mathcal{F}$ is the set of classifiers, and $Loss(f)$ is the loss of classifier f. Moreover, the authors propose the use of a randomized classifier. A randomized classifier is a probability distribution Q over $\mathcal{F}$. Intuitively, this randomized classifier can reach any convex combination of points outputted by the single classifiers, whereas single classifiers are restricted to a discrete set of loss-fairness combinations. Therefore, a randomized classifier can potentially reach trade-off points that a single classifier could not attain. The optimization problem 4.6 is then written as

$$\min_{Q\in\Delta} Loss(Q) \qquad s.t. \quad \mathbf{M}\boldsymbol{\mu}(Q) \leq \mathbf{c}.$$

Here, Δ is all possible probability distributions over $\mathcal{F}$. However, this problem is not tractable since the true distribution of the data is not known. Hence, the quantities in the minimization problem are estimated based on the available data. Moreover, a small nonnegative error term ϵ is added to the allowed unfairness $\mathbf{c}$, this results in $\hat{\mathbf{c}}$. The empirical optimization problem is denoted as

$$\min_{Q\in\Delta} \hat{Loss}(Q) \qquad s.t. \quad \mathbf{M}\hat{\boldsymbol{\mu}}(Q) \leq \hat{\mathbf{c}}. \tag{4.7}$$

This constrained optimization problem can be written in Lagrangian form. The Lagrangian multiplier has a single multiplier for each of the elements in the set $\mathbb{K}$. Additionally, an l_1 norm is imposed on the Lagrangian multiplier. Let $\mathcal{L}(Q, \boldsymbol{\lambda}) = \hat{Loss}(Q) + \boldsymbol{\lambda}^T(\mathbf{M}\hat{\mu}(Q) - \hat{\mathbf{c}})$. Equation 4.7 in Lagrangian form is then defined as:

$$\min_{Q \in \Delta} \max_{\lambda \in \mathbb{R}_+^{|\mathbb{K}|}, ||\boldsymbol{\lambda}||_1 \leq B} \mathcal{L}(Q, \boldsymbol{\lambda}).$$

With the corresponding dual

$$\max_{\boldsymbol{\lambda} \in \mathbb{R}_+^{|\mathbb{K}|}, ||\boldsymbol{\lambda}||_1 \leq B} \min_{Q \in \Delta} \mathcal{L}(Q, \boldsymbol{\lambda}).$$

Both have solutions where the objective value of the primal is equal to the objective value of the dual since $\mathcal{L}(Q, \boldsymbol{\lambda})$ is linear in both Q and $\boldsymbol{\lambda}$ (shown in Appendix A.2), with convex and compact domains among other conditions. This optimization problem is viewed as a saddle point problem and is solved by the corresponding Algorithm 1.

Algorithm 1 Reductions based approach from Agarwal *et al.* [1].

Require: Training data $\{(x_j, y_j, g_j)\}_{j=1}^N$, constraints $\mathbf{M}\hat{\boldsymbol{\mu}}(Q)$, bound B, learning rate η, tolerance ν

Initialize $\boldsymbol{\phi}^{(1)} = \mathbf{0} \in \mathbb{R}^{|\mathbb{K}|}$

for $t = 1, 2, \ldots T$ **do**

λ-player: exponentiated gradient update

$$\lambda_k^{(t)} = B \cdot \frac{e^{\phi_k^{(t)}}}{1 + \sum_{k' \in \mathbb{K}} e^{\phi_{k'}^{(t)}}} \quad \forall k \in \mathbb{K}$$

Q-player: best response via cost-sensitive classification

$f^{(t)} \leftarrow \textsc{BestClassifier}(\boldsymbol{\lambda}^{(t)})$

Update running averages

$\hat{Q}^{(t)} = \frac{1}{t}\sum_{t'=1}^{t} f^{(t')}, \quad \hat{\boldsymbol{\lambda}}^{(t)} = \frac{1}{t}\sum_{t'=1}^{t} \boldsymbol{\lambda}^{(t')}$

Check suboptimality of current average play

$\nu_t = \max\Big\{\mathcal{L}(\hat{Q}^{(t)}, \textsc{BestLambda}(\hat{Q}^{(t)})) - \mathcal{L}(\hat{Q}^{(t)}, \hat{\boldsymbol{\lambda}}^{(t)}), \mathcal{L}(\hat{Q}^{(t)}, \hat{\boldsymbol{\lambda}}^{(t)}) - \mathcal{L}(\textsc{BestClassifier}(\hat{\boldsymbol{\lambda}}^{(t)}), \hat{\boldsymbol{\lambda}}^{(t)})\Big\}$

if $\nu_t \leq \nu$ **then**

return $(\hat{Q}^{(t)}, \hat{\boldsymbol{\lambda}}^{(t)})$

end if

Gradient step on $\boldsymbol{\phi}$

$\boldsymbol{\phi}^{(t+1)} = \boldsymbol{\phi}^{(t)} + \eta\left(\mathbf{M}\hat{\boldsymbol{\mu}}(f^{(t)}) - \hat{\mathbf{c}}\right)$

end for

In this algorithm, $\hat{Q}^{(t)}$ is the empirical approximation of Q, which equally weights all classifiers up to point t. Additionally, $\boldsymbol{\phi}$ is updated using regular gradient updates, while $\hat{\boldsymbol{\lambda}}^{(t)}$ is updated with exponentiated gradient updates. The algorithm runs for a pre-defined

number of iterations T or until the difference between utility gains for the best $\hat{Q}^{(t)}$ and the best $\hat{\boldsymbol{\lambda}}^{(t)}$ is smaller than hyperparameter ν.

The function BestLambda($\hat{Q}^{(t)}$) is straightforward to obtain; either the constraints are satisfied, and hence the maximum of the Lagrangian is obtained for $\hat{\boldsymbol{\lambda}} = 0$. Or the entire weight B is placed on λ_k, where k is the most violated constraint.

Moreover, BestClassifier($\hat{\boldsymbol{\lambda}}^{(t)}$) is solved by choosing Q such that the Lagrangian, $\mathcal{L}(Q, \boldsymbol{\lambda})$, is minimized. Since $\mathcal{L}(Q, \boldsymbol{\lambda})$ is linear in Q, and Δ is a convex and compact set, the function is minimized at an extreme point of the set. That means BestClassifier($\hat{\boldsymbol{\lambda}}^{(t)}$) is obtained by choosing a single classifier f. Furthermore, the minimization problem can be solved by regarding it as a cost sensitive classifier. For a single classifier f, and DP the following expression is obtained:

$$\mathcal{L}(Q, \boldsymbol{\lambda}) = \hat{Loss}(f) + \boldsymbol{\lambda}^T(\mathbf{M}\hat{\boldsymbol{\mu}}(f) - \hat{\mathbf{c}})) = \hat{\mathbb{E}}[\mathbf{1}\{f(X) \neq Y\}] - \boldsymbol{\lambda}^T\hat{\mathbf{c}} + \sum_{k,g_i} \frac{M_{k,g_i}\lambda_k}{p_{g_i}} \hat{\mathbb{E}}[f(X)\mathbf{1}\{G = g_i\}] + \sum_k M_{k,pop}\lambda_k \hat{\mathbb{E}}[f(X)].$$

Here, p_{g_i} is the probability of belonging to group $g_i \in \mathcal{G}$. The classifier f that minimizes this corresponds to solving the cost sensitive classifier written as

$$\arg\min_{f \in \mathcal{F}} \sum_{j=1}^{N} f(x_j)C_j^1 + (1 - f(x_j))C_j^0. \tag{4.8}$$

Moreover,

$$C_j^0 = \mathbf{1}\{Y_j \neq 0\},$$

and

$$C_j^1 = \mathbf{1}\{Y_j \neq 1\} + \frac{\lambda_{G(x_j)}}{p_{G(x_j)}} - \sum_{g_l \in \mathcal{G}} \lambda_{g_l}.$$

Here, $\lambda_{g_i} = \lambda_{g_i,1}$ - $\lambda_{g_i,0}$, and p_{g_i} is the probability of belonging to group $g_i \in \mathcal{G}$. Moreover, the costs are intuitive. For both labels, it holds that predicting the label incurs a cost for misclassifications. Furthermore, since label 1 is assumed to be the desired outcome, no extra fairness cost is assigned for predicting label 0. While for predicting label 1, there is an additional penalty proportional to the price λ_{g_i}, divided by the probability of belonging to g_i. Additionally, an overall term, $\sum_{g_i \in \mathcal{G}} \lambda_{g_i}$, is subtracted to ensure balanced cost updates for all groups. The problem 4.8 can therefore be solved by regular cost sensitive classifiers.

Furthermore, the authors set most hyperparameters of Algorithm 1 during solving to fit the problem at hand. The upper bound of the l_1 norm B is set to $\frac{1}{\epsilon}$. The tolerance $\nu := 0.5\frac{\sigma(|f(X)-Y|)}{\sqrt{N}}$, where $\sigma(\cdot)$ denotes the standard deviation. The tolerance thus depends on the amount of errors weighted by the number of observations. Lastly, the learning rate η is initialized as $\frac{\eta_0}{B}$ at iteration 0 with a default value of 2. However, during training, at every multiplicatively increasing interval, if the best gap ν^* has not improved by 20% compared to the previous best gap ν^*_{prev}, η shrinks by 20%. This then means:

$$\nu^* > 0.8\,\nu^*_{\text{prev}} \quad \Longrightarrow \quad \eta \leftarrow 0.8\,\eta.$$

Wherever this condition is checked, it is delayed by a factor of `check_t` = 1.6. During early iterations, this is checked more often; whereas, as the iterations progress, this check

is executed less frequently. The algorithm assumes that the acquired learning rate becomes more stable as the algorithm progresses through its iterations.

Combining these results leads to a solvable algorithm where, for a given level of unfairness and a set of classifiers, this randomized classifier lies on the Pareto frontier.

4.3.4 Fair Ensemble

As Kearns *et al.* [31] noted, a classifier can seem fair with respect to a certain attribute; however, taking a closer look at more granular subgroups might reveal unfairness. The Fair Ensemble (FE) method proposed in this section aims to mitigate unfairness by training three separate fair classifiers, one for each of the sensitive attributes. The final prediction is made by obtaining an equally weighted sum of the probabilities. With the obtained probabilities, predictions are made for the FE. The method will be evaluated empirically to verify whether combining individual fairness-aware ML models can produce improvements at the intersectional level.

The FE thus uses classifiers that are trained to be fair at the group level; for instance, one for the attribute gender and one for the attribute education. By combining the probabilities of these classifiers, the FE aims to obtain a classifier that is fair regarding the intersectional subgroups. However, the FE never considers all intersectional subgroups, it only uses group-level fair classifiers, which is computationally cheap. Considering m binary sensitive attributes, the number of intersectional subgroups is then 2^m; for multiclass sensitive variables, the number of subgroups will be even larger. The reason for using FE is the theory that taking averages over group level fair classifiers could result in a classifier that is fair regarding the intersectional subgroups, which is a theory that is tested empirically. In conclusion, the FE seeks to produce fair classification at the intersectional subgroup level without considering all intersectional subgroups.

For the general setting with $|\mathcal{S}|$ number of attributes, let $f_1, f_2, ..., f_{|\mathcal{S}|}$ be the fair classifiers trained to be fair with respect to a different single protected variable. The FE probabilities are then defined as

$$\boldsymbol{\pi} = \sum_{i=1}^{|\mathcal{S}|} w_i P_i(X).$$

Here, w_i is the weight assigned to classifier i, e.g., $\frac{1}{|\mathcal{S}|}$ for equal weights. Moreover, $P_i(X)$ represents the probabilities that the i-th classifier assigns to the instances in the dataset X for having the label 1. For binary classification with a classification threshold 0.5, the final predictions are

$$\hat{Y} = \mathbf{1}\{\boldsymbol{\pi} \geq 0.5\}.$$

Here $\boldsymbol{\pi}$ is the probability vector of the dataset X. This thesis uses equal weights because no attribute is preferred over another. However, this choice can be adjusted. For example, if the data shows significantly more disparity with respect to a specific sensitive attribute, the weight corresponding to the classifier of this attribute can be increased. However, the sum of weights needs to be one, and the weights need to be non-negative in order to obtain a true probability.

Although the FE does not provide any fairness guarantees, it can serve as an intuitive approach to increasing fairness in ML models. Moreover, it can be paired with any fairness-

aware ML model as long as the probabilities of the instances can be accessed, and the framework can be extended to the setting of multi-class prediction.

4.4 Prediction Methods

The methods discussed in the previous section all try to mitigate unfairness; however, this section discusses the baseline models against which the fairness-aware ML models are evaluated. The models chosen as baselines are logistic regression and XGBoost. These are selected to be consistent with prior work on employee attrition prediction [19, 47, 48, 58]. Moreover, both methods are capable of handling weighted datasets, which is useful for the Reweighing and the Reductions Based Approach, allowing for clearer comparison.

4.4.1 Logistic Regression

Logistic regression is a model that can be used for binary classification, it models the probability of belonging to a certain class, applying a sigmoid function to a linear combination of the features. Following the notation of Hosmer *et al.* [25]:

$$\mathbb{E}[Y|x_i] = \pi(x_i) = \frac{1}{1 + e^{-(\beta_0 + \sum_{j=1}^{p} \beta_j x_j)}}.$$

Here, $Y \in \{0, 1\}$, x_i is a single observation of X, and the coefficients β are estimated using maximum likelihood estimation. The sigmoid function maps all linear combinations of the features to a $[0, 1]$ interval. Moreover, $\mathbb{E}[Y|x_i]$ can be used as a probability. An instance is predicted to have the label 1 if this probability is above a certain threshold; in this thesis, that threshold is 0.5. Moreover, the function $g(x_i)$ is defined to be the logit transformation:

$$\begin{aligned} g(x_i) &= \ln(\frac{\pi(x_i)}{1 - \pi(x_i)}) \\ &= \beta_0 + \sum_{j=1}^{p} \beta_j x_j. \end{aligned}$$

Logistic regression thus assumes a linear relationship between the features and the logits [25], which can limit the ability of logistic regression to capture nonlinear relationships.

4.4.2 XGBoost

XGBoost is an ensemble tree boosting method that builds trees sequentially. The model uses a differentiable and convex loss function with an added regularization term that penalizes model complexity to prevent overfitting. The objective function is defined as:

$$Loss(\hat{Y}, Y) + \sum_{k} \Omega(f_k),$$

where, f_k is the k-th tree and $\Omega(f_k)$ is a penalty term that increases as model complexity increases. At each step, a tree f_k is fitted by minimizing the second-order Taylor approximation of the objective function. Additionally, to further control overfitting, XGBoost uses shrinkage and column subsampling. For faster computations, finding splits for different features is parallelized within trees. Furthermore, XGBoost can handle sparse data and missing values by using sparsity-aware split finding [11].

4.5 Evaluation Metrics

Before defining the evaluation metrics, recall that the label 1 was determined to correspond to staying; therefore, class 0 represents the minority class, which consists of the employees who left. However, most evaluation metrics, such as the F1 score, assume the minority class is assigned the label 1. Standard formulas would therefore evaluate performance solely on the majority class instead of the minority class. Hence, the following notation is introduced:

- TP_0: employees who were correctly predicted to leave
- FP_0: employees who were incorrectly predicted to leave
- TN_0: employees who were correctly predicted to stay
- FN_0: employees who were incorrectly predicted to stay

This notation is necessary to define the evaluation metrics below.

The first metric discussed is accuracy, which is a straightforward metric that measures the percentage of correctly classified instances and is denoted by

$$\textbf{Accuracy} = \frac{TP_0 + TN_0}{N}.$$

However, since the IBM dataset is highly imbalanced, accuracy may yield misleadingly high values. A classifier that predicts the majority class can attain high values of accuracy on an imbalanced dataset. Therefore, the F1 score is used as an evaluation metric. The F1 score combines recall and precision, ignoring the true negatives, to produce an evaluation metric that is more robust to class imbalances.

Precision captures the fraction of predicted positives that are actually positive. Precision lies in the interval $[0, 1]$, where a higher value is desired. Precision is given by:

$$\textbf{Precision} = \frac{TP_0}{TP_0 + FP_0}.$$

Moreover, recall is the fraction of positives that are correctly captured. The value of recall is in $[0, 1]$. A value of 1 means no positive instances are missed, and a value of 0 means all positive instances are missed.

$$\textbf{Recall} = \frac{TP_0}{TP_0 + FN_0}.$$

The F1 score balances recall and precision and is written as

$$\textbf{F1} = 2 \times \frac{\textbf{Recall} \times \textbf{Precision}}{(\textbf{Recall} + \textbf{Precision})}.$$

The F1 score also lies in the interval $[0, 1]$, with a higher value indicating a better performing model.

Chapter 5

Experimental Setup

In this chapter, the details on model training and experimental setup are discussed. Section 5.1 discusses the overall model training, Section 5.2 additionally shows the obtained hyperparameters for each model. Lastly, Section 5.3 discusses the methods of testing the results.

5.1 General Model Training

The common training pipeline is as follows: the model is trained using the training set and evaluated on the test set. Certain methods require a development set, for instance, to determine which classifier to use. This development set is constructed by taking a 30% split from the training data using a stratified split on the intersectional subgroups. The resulting development set consists of $17,865$ observations. Additionally, all models are also evaluated on the IBM dataset. The Differential Fairness Method (DFM) normalizes the numerical features distance from home and monthly income, as normalizing numerical features often improves neural network performance [39].

The sensitive attributes are included as features during both training and inference, which is consistent with the requirements for the methods used in this thesis.

5.2 Method-Specific Model Training and Hyperparameter Tuning

All models use Optuna for hyperparameter tuning. Optuna is an optimization framework that implements effective searching and pruning techniques [2]. Most methods are tuned using stratified 5-fold cross validation, where the data is stratified by the number of subgroups, and each model receives 50 trials to obtain the best set of hyperparameters. These settings allow for a rigorous search in the hyperparameter space while remaining computationally accessible. Additionally, the fairness constraint tolerance ϵ of the Reductions Based Approach and the weighting term for the fairness penalty λ for the DF method are tuned on a withheld development set. Since both parameters regulate the utility-fairness trade-off, they should be optimized using a multi-objective function. Instead, a grid-search is performed to identify the values that yield the most favorable balance between predictive performance and the level of unfairness.

The fairness metric used to evaluate the utility-fairness trade-off depends on the metric that the model uses. For example, DFM uses DF during training. Hence, to select the value of λ, the performance is evaluated on DF (both soft and hard counts). All other methods use intersectional DP as a fairness metric; therefore, intersectional DP is used to select the best hyperparameter values. This framework is appropriate since each model is then evaluated based on the metric it aims to minimize.

Reweighing

The Reweighing method uses a classifier that can handle sample weights to make predic-

tions. In this thesis, logistic regression and XGBoost are considered since they can handle sample weights and are often employed in the literature on employee attrition prediction. The classifier that achieves the best utility-fairness trade-off on the development set is chosen. XGBoost yielded a higher F1 score and lower fairness metrics; hence, the classifier used is XGBoost. Table 5 shows all hyperparameters that were considered and the obtained values. All folds are evaluated based on the F1 score.

Hyperparameter	**Search Range**	**Selected Value**
Maximum tree depth (max_depth)	$\{3, 4, \ldots, 10\}$	3
Learning rate ($learning\_rate$)	$[10^{-4}, 0.25]$ (log-uniform)	0.043
Number of estimators ($n\_estimators$)	$\{100, 300, 500, 700, 1000\}$	500
Minimum child weight (min_child_weight)	$\{1, 2, \ldots, 10\}$	3
Minimum loss reduction (γ)	$[10^{-8}, 5.0]$ (log-uniform)	0.307
Subsample ratio ($subsample$)	$[0.5, 1.0]$	0.822
L1 regularization (reg_alpha)	$[10^{-8}, 5.0]$ (log-uniform)	3.628×10^{-5}
L2 regularization (reg_lambda)	$[1.0, 10.0]$ (log-uniform)	1.456

Table 5: Hyperparameter search space and selected values for Reweighing with XGBoost.

As mentioned earlier, XGBoost actively tries to reduce overfitting, where most of these hyperparameters heavily influence the models risk of overfitting.

Differential Fairness Method The differential fairness method makes use of a NN for predictions. Foulds *et al.* [21] use a simple 3 layer NN. This thesis adopts this framework of a 3-layer FFNN with ReLu activation functions and the Adam optimizer.

Moreover, DFM is trained using batched gradient descent. However, as Foulds *et al.* [21] note, small batch sizes potentially lead to small subgroups, which, in turn, could produce unreliable estimates of the fairness metric. Therefore, a large batch size of 5000 was chosen to obtain reliable estimates for the fairness metrics. Moreover, the burn-in number of epochs is 50, and the fairness-aware training epochs are set to 500, which aligns with the paper. Furthermore, the parameter ϵ_{base} is set to 0. The AI Act does not provide any level of unfairness that is allowed; therefore, this thesis assumes the goal of the fairness penalty is to penalize any magnitude of unfairness. However, setting ϵ_{base} to 0 does not imply that the model will have outputs that perfectly satisfy the fairness metric. The objective function of the DF method contains the parameter λ. This parameter controls the fairness-predictive performance trade-off and cannot be tuned simultaneously with the other hyperparameters because λ would be set close to 0 to minimize the objective function. Therefore, the development set is used, and λ is tuned via grid search using the best set of hyperparameters.

Table 6 shows the tuned hyperparameters and selected values. Each set of hyperparameters is trained for 10 burn-in epochs and 50 fair-learning epochs. Moreover, each trial is evaluated using Equation 4.4, where, during training $\lambda = 0.01$, in line with the paper. The model performance of λ is evaluated on the development set, where the model is trained for 50 burn-in epochs and 500 fairness-aware epochs. $\lambda = 0.5$ is selected, as it achieved the highest F1 score while improving DF fairness metrics compared to lower values of λ. Increasing λ further quickly degrades the F1 score and the soft DF metrics. Notably, as

λ approaches 1, the fairness metrics computed using hard counts actually increase, likely due to probabilities being close between subgroups but on opposite sides of the threshold.

Hyperparameter	Search Range	Selected Value
Learning rate	$[10^{-4}, 10^{-1}]$	0.0001
Dropout rate	$[0.0, 0.5]$	8.2×10^{-4}
Hidden units	$\{32, 64, 128, 256, 512\}$	512
λ	$\{0.05\} \cup \{0.1, 1.0\}$ in steps of 0.1	0.5

Table 6: Hyperparameter search space and selected values for the DFM.
The search range for λ consisted of 0.05 and values between 0.1 and 1.0 in steps of 0.1.

Here, the dropout rate is a regularization hyperparameter, the learning rate influences training stability, and the hidden units determine the model complexity.

Reduction Based Approach This method makes use of a base classifier that can handle a weighted classification problem. To align with the baseline models, the base classifiers used are logistic regression and XGBoost. Jointly tuning ϵ and the parameters of the base classifier would result in a circular dependency since the best hyperparameters of the base classifier depend on ϵ and vice versa. Given the value of ϵ, the hyperparameters of the weighted classifiers are tuned using Optuna with the objective to maximize the F1 score. This aligns with how Agarwal *et al.* [1] treat the base classifier; they regard it as a black-box algorithm that tries to maximize utility without having explicit knowledge about the definition of fairness.

The value of ϵ appeared to neither influence F1 nor intersectional DP substantially. This suggests that if the fairness constraints are activated, the algorithm is likely to produce classifiers that rapidly achieve low values of intersectional DP. Moreover, ϵ is not only a hyperparameter, it also indicates the allowed level of unfairness. The parameter, therefore, is also interpretable; e.g., a value of 0.01 indicates that an intersectional DP value of 1% is allowed. Consequently, ϵ is set to the default value of 0.01 from the paper.

To determine which base classifier to use, F1 and intersectional DP are evaluated for both logistic regression and XGBoost with $\epsilon = 0.01$. The classifier yielding the best trade-off in terms of F1 score, fairness, and standard deviation is then selected. XGBoost overall obtained better predictive performance while performing similarly in terms of intersectional DP; hence, XGBoost is selected as the base classifier.

Additionally, the number of iterations is kept at the default level of 50 since improving it did not yield a notable difference in predictive performance and fairness metrics.

Table 7 shows the selected base classifier, where the name denotes which subgroups were considered; for instance, Intersectional corresponds to the setting that considers all intersectional subgroups.

Model	Classifier
Intersectional	XGBoost
Gender	XGBoost
Age	XGBoost
Education	XGBoost

Table 7: Selected classifiers for each Reductions Based Approach model.

Hyperparameter	Search Range	Selected Value
Maximum tree depth (*max_depth*)	$\{3, 4, \ldots, 10\}$	3
Learning rate (*learning_rate*)	$[10^{-4}, 0.25]$ (log-uniform)	0.022
Number of estimators (*n_estimators*)	$\{100, 300, 500, 700, 1000\}$	700
Minimum child weight (*min_child_weight*)	$\{1, 2, \ldots, 10\}$	4
Minimum loss reduction (γ)	$[10^{-8}, 5.0]$ (log-uniform)	1.440×10^{-4}
Subsample ratio (*subsample*)	$[0.5, 1.0]$	0.740
L1 regularization (*reg_alpha*)	$[10^{-8}, 5.0]$ (log-uniform)	8.488×10^{-6}
L2 regularization (*reg_lambda*)	$[1.0, 10.0]$ (log-uniform)	1.813

Table 8: Hyperparameter search space and selected values for XGBoost in the Reductions Based Approach considering intersectional subgroups.

Ensemble The ensemble method makes use of existing fair classifiers. In this thesis, the fair classifiers that will be used are those obtained by using the Reduction Based Approach, since it has shown promising results for single multi-class sensitive attributes [1]. The hyperparameters for the base classifiers of the individual fairness-aware ML models are obtained in the same way as for the Reduction Based Approach; the hyperparameter values are denoted in the three tables below. Additionally, the selected classifiers are shown in Table 7.

Noteworthy is the fact that the education-based model and the gender-based model yield the exact same hyperparameters. A probable cause is that the seed provides the same search space, independent of the model, and it reveals that these hyperparameters provide the best performance for both models.

Hyperparameter	**Search Range**	**Selected Value**
Maximum tree depth (*max_depth*)	$\{3, 4, \ldots, 10\}$	3
Learning rate (*learning_rate*)	$[10^{-4}, 0.25]$ (log-uniform)	0.036
Number of estimators (*n_estimators*)	$\{100, 300, 500, 700, 1000\}$	300
Minimum child weight (*min_child_weight*)	$\{1, 2, \ldots, 10\}$	4
Minimum loss reduction (γ)	$[10^{-8}, 5.0]$ (log-uniform)	3.568
Subsample ratio (*subsample*)	$[0.5, 1.0]$	0.899
L1 regularization (*reg_alpha*)	$[10^{-8}, 5.0]$ (log-uniform)	5.012×10^{-7}
L2 regularization (*reg_lambda*)	$[1.0, 10.0]$ (log-uniform)	1.007

Table 9: Hyperparameter search space and selected values for XGBoost in the gender-based model.

Hyperparameter	**Search Range**	**Selected Value**
Maximum tree depth (*max_depth*)	$\{3, 4, \ldots, 10\}$	3
Learning rate (*learning_rate*)	$[10^{-4}, 0.25]$ (log-uniform)	0.042
Number of estimators (*n_estimators*)	$\{100, 300, 500, 700, 1000\}$	300
Minimum child weight (*min_child_weight*)	$\{1, 2, \ldots, 10\}$	2
Minimum loss reduction (γ)	$[10^{-8}, 5.0]$ (log-uniform)	0.149
Subsample ratio (*subsample*)	$[0.5, 1.0]$	0.904
L1 regularization (*reg_alpha*)	$[10^{-8}, 5.0]$ (log-uniform)	1.240×10^{-6}
L2 regularization (*reg_lambda*)	$[1.0, 10.0]$ (log-uniform)	1.764

Table 10: Hyperparameter search space and selected values for XGBoost in the age-based model.

Hyperparameter	Search Range	Selected Value
Maximum tree depth (*max_depth*)	$\{3, 4, \ldots, 10\}$	3
Learning rate (*learning_rate*)	$[10^{-4}, 0.25]$ (log-uniform)	0.036
Number of estimators (*n_estimators*)	$\{100, 300, 500, 700, 1000\}$	300
Minimum child weight (*min_child_weight*)	$\{1, 2, \ldots, 10\}$	4
Minimum loss reduction (γ)	$[10^{-8}, 5.0]$ (log-uniform)	3.568
Subsample ratio (*subsample*)	$[0.5, 1.0]$	0.899
L1 regularization (*reg_alpha*)	$[10^{-8}, 5.0]$ (log-uniform)	5.012×10^{-7}
L2 regularization (*reg_lambda*)	$[1.0, 10.0]$ (log-uniform)	1.007

Table 11: Hyperparameter search space and selected values for XGBoost in the education-based model.

Logistic Regression For Logistic Regression, certain parameters are important for convergence, such as the solver, tolerance, and maximum iterations. Others are important for regularization and, hence, generalizability, such as penalty and regularization strength. Each trial is evaluated using the F1 score. Moreover, the selected values for the hyperparameters can be found in Table 12.

Hyperparameter	Search Range	Selected Value
Solver	{`lbfgs`, `liblinear`, `saga`}	`saga`
Penalty	{`l1`, `l2`, `elasticnet`, `None`}	`None`
$l1$ ratio	`None` or $\{0, 1\}$	`None`
Tolerance (*tol*)	$[10^{-6}, 10^{-3}]$	7.69×10^{-4}
Regularization strength (C)	$[10^{-2}, 1]$	4.72×10^{-2}
Fit intercept	{`True`, `False`}	`True`
Maximum iterations	$\{100, 300, 500, 700, 1000\}$	700

Table 12: Hyperparameter search space and selected values with baseline Logistic Regression.

Note: Invalid solver–penalty combinations are remapped to valid combinations during optimization, and $l1$ ratio is only for elasticnet penalty.

XGBoost The search space of the XGBoost classifier is the exact same as the search space for the XGBoost classifier in the Reweighing method. The hyperparameter values obtained for the unconstrained XGBoost classifier maximizing the F1 score are shown in Table 13.

Hyperparameter	Search Range	Selected Value
Maximum tree depth (*max_depth*)	$\{3, 4, \ldots, 10\}$	3
Learning rate (*learning_rate*)	$[10^{-4}, 0.25]$ (log-uniform)	0.128
Number of estimators (*n_estimators*)	$\{100, 300, 500, 700, 1000\}$	100
Minimum child weight (*min_child_weight*)	$\{1, 2, \ldots, 10\}$	2
Minimum loss reduction (γ)	$[10^{-8}, 5.0]$ (log-uniform)	0.405
Subsample ratio (*subsample*)	$[0.5, 1.0]$	0.795
L1 regularization (*reg_alpha*)	$[10^{-8}, 5.0]$ (log-uniform)	0.004
L2 regularization (*reg_lambda*)	$[1.0, 10.0]$ (log-uniform)	4.888

Table 13: Hyperparameter search space and selected values for XGBoost.

5.3 Significance Testing

5.3.1 Significance Testing for Fairness Metrics

Point estimates for fairness metrics can be misleading. Hence, this thesis provides confidence intervals and empirical p-values for all the fairness metrics.

The confidence intervals are constructed using stratified bootstrapping samples with replacement. The samples are stratified based on intersectional group size to preserve reliable estimates for all subgroups. For each of these samples, all fairness metrics are computed, which is done for 500 iterations to obtain reliable estimates. Subsequently, the confidence intervals are equal to $[Q_{2.5}, Q_{97.5}]$, where $Q_{2.5}$ and $Q_{97.5}$ are the 2.5 and 97.5 percentiles, respectively. Moreover, using the bootstrap metrics, the most unfairly treated groups are tracked. This thesis considers the following null hypothesis:
H_0: the observed value of the fairness metric is consistent with what would be observed if the classifier's predictions would be randomly assigned to group membership.
That means this thesis tests how likely it is that this level of disparity is observed if the classifier's predictions were randomly paired with group membership, assuming there is no relationship between the sensitive variables and the predicted label. To test this hypothesis, the relationship between predicted labels and instances is removed by permuting the predicted labels and the probabilities that a classifier outputs simultaneously. This preserves the label-probability relationship while breaking the connection with group membership. Using these permutations, all fairness metrics are computed. In line with the permutation tests defined by Ojala & Garriga [42], the one-sided empirical p-value is then obtained by:

$$p = \frac{\sum_{i=1}^{500} \mathbf{1}\{fm_i^{perm} \geq fm^{obs}\} + 1}{\#\text{Iterations} + 1},$$

and the two-sided test is

$$p = \frac{\sum_{i=1}^{500} \mathbf{1}\{|fm_i^{perm}| \geq |fm^{obs}|\} + 1}{\#\text{Iterations} + 1}.$$

The one-sided test is used for fairness metrics where larger disparities indicate more unfairness, for instance, all intersectional fairness metrics. This is done because values as

large as the obtained value provide evidence against the null hypothesis. A two sided test is used for the fairness metric that includes signed values, which is only the DP difference for gender. For this fairness metric, values closer to zero are considered more fair, regardless of the direction. Both tests are bounded from below by the value $\frac{1}{501}$. Moreover, in these equations, fm_i^{perm} is the i-th permutation value of a certain fairness metric fm. Here, fm is either intersectional DP or SPSF. Additionally, fm^{obs} is the observed value for the fairness metric fm. The proportions result in an empirical permutation p-value that estimates the probability of observing a disparity this large under random permutations of group assignments. This proportion thus counts how often the fairness metric under permutations is as extreme as the observed fairness metric, indicating at least as extreme unfairness under the null. While the empirical p-value is not derived from an analytical distribution, it retains a direct probabilistic interpretation. That is, if the empirical p-values are below the chosen significance threshold, $\alpha = 0.05$, the null hypothesis is rejected, and it is assumed that this level of disparity is highly unlikely to be caused by chance.

For the interpretation of the confidence intervals and empirical p-values, it is important to note the sample size of each subgroup. For both the test set and the IBM dataset, the number of observations per subgroup is denoted in Tables 22 and 23, respectively. For specific subgroups, such as females aged 46+ with education level 1, the number of observations is small. Therefore, especially the empirical p-values and confidence intervals for the IBM data should be interpreted with care, as a low number of observations can lead to unreliable results.

5.3.2 Significance Testing Between Methods

Since point estimates with standard deviations can overlap between methods, this thesis tests whether the difference in a given metric between models is significant. The metrics considered are intersectional DP, SPSF, and F1 score. The framework of the proportions used in Section 5.3.1 has no natural equivalent permutation scheme to obtain the null for comparing two trained models, unlike the null of the previous section, which can be simulated by permutations. Since two models are different trained classifiers, simulating the null hypothesis that there is no difference for a given metric cannot be done by permutations. Hence, testing in this section is performed using confidence intervals. To test the difference in, for example, F1 score between models, 500 bootstrap samples are used. These bootstrap samples are again stratified on subgroup size. Consequently, for each method, F1, intersectional DP, and SPSF are calculated for all bootstrap samples. This is done without retraining the model, which ignores model variance but is computationally inexpensive. Subsequently, for every pair of models and every bootstrap sample, the difference in the metric value between two methods is computed. Consequently, with these differences, a 95% confidence interval is obtained. The null hypothesis states that the expected difference in the metric between two methods is zero. For the F1 score between model A and model B, that is:
$H_0 : \mathbb{E}[F1_A - F1_B] = 0$.
For intersectional DP and SPSF, the null is constructed similarly. A confidence interval that does not contain zero is equivalent to rejecting the two-sided null hypothesis at a 5% significance level, according to Theorem 9.2.2 from Casella & Berger [7, p.421].

Chapter 6

Results

This chapter will provide all results needed to answer the research question and sub-questions. Moreover, this chapter will use abbreviations in text, tables and figures. A list of abbreviations is provided below.

- LR = Logistic Regression
- XGB = XGBoost
- RW = Reweighing
- DFM = DF Method
- EG = Reductions Based Approach
- FE = Fair Ensemble

All results are provided for both the primary dataset and the IBM dataset. However, IBM results serve exclusively as a robustness check, and differences with the primary dataset can occur due to class imbalance and distributional differences in the IBM dataset. Moreover, as Section 3.5 identified, the differences in data distribution for attributes such as monthly income and distance from home, as well as the small number of observations in certain subgroups, could result in fairness metrics of different magnitudes for the IBM dataset.

6.1 Baseline Models

Table 14 shows the predictive performance metrics of both baseline models across both primary and IBM datasets, accompanied by the standard deviation. The standard deviations are obtained from 500 bootstrap samples. Both methods exhibit a small standard deviation for the primary dataset, suggesting low sampling variability for the metrics for this dataset. Additionally, XGB consistently outperforms LR on every metric for the primary dataset, where XGB attains a value of 0.695 for the F1 score, indicating a reasonable trade-off between precision and recall. Recall is particularly important as it measures the fraction of leavers that are captured correctly. The performance drops substantially on the IBM dataset, except the accuracy of LR. However, as noted in Chapter 3, the IBM dataset is imbalanced, and hence the accuracy can appear misleading. Moreover, LR outperforms XGB for all metrics except recall. On the IBM dataset, both models obtain low precision and therefore a low F1 score; this is due to the class imbalance of the IBM dataset. Recall that the evaluation metrics are computed with respect to class 0, the leavers, as defined in Section 4.5. The evaluation metric recall for the majority class, the stayers, likely remains high, while recall for the minority class, the leavers, likely drops substantially. The precision metric may be affected more, as a small false positive rate leads to a large absolute number of false positives, which heavily influences precision in the minority class, while

precision for the majority class is probably less affected, as the false positives are small compared to the size of the majority class. Therefore, accuracies remain broadly similar across datasets, but precision for the minority class likely drops considerably. Also, the standard deviation is higher for the IBM dataset, implying more uncertainty in the predictions. Overall, XGBoost performs better on the primary dataset, and LR seems to be marginally more robust to data coming from a different distribution.

Model	Accuracy	Precision	Recall	F1-score
		Primary Dataset		
LR	0.708 ± 0.004	0.698 ± 0.004	0.671 ± 0.006	0.684 ± 0.004
XGB	0.716 ± 0.004	0.704 ± 0.004	0.686 ± 0.005	0.695 ± 0.004
		IBM Dataset		
LR	0.711 ± 0.012	0.295 ± 0.015	0.571 ± 0.031	0.389 ± 0.019
XGB	0.683 ± 0.012	0.276 ± 0.014	0.592 ± 0.031	0.376 ± 0.018

Table 14: Baseline predictive performance on the primary and IBM datasets.

In Table 15 DP is shown with respect to the individual sensitive attributes, where the bold values indicate a value that is unlikely under the permutation null at the 5% significance level. Furthermore, the values in the squared brackets denote the confidence intervals. All values are shown to be significantly higher from those obtained from the permutation samples, meaning that the observed values of the fairness metrics are unlikely to arise if the predictions of attrition are assigned randomly independent of group membership (see Section 5.3.1 for the testing procedure). Moreover, a DP value of 0.179 for LR regarding DP Gender indicates that there is a 17.9 percentage point difference in the probability that a male is assigned to class 1 compared to a female. Additionally, for gender, DP is lower for the IBM dataset, while the sensitive variable age obtains a higher DP. This aligns with the findings in the EDA, which showed the IBM dataset having a smaller deviation in attrition rates for gender but a larger disparity between attrition rates for the attribute age. In addition, XGB and LR differ in the magnitude of violation; for example, LR attains a higher DP for gender on the primary dataset. This reflects that the models potentially learned a different decision boundary. Furthermore, XGB consistently exhibits a higher DP on the IBM dataset.

Model	DP Gender	DP Age	DP Education
	Primary Dataset		
LR	**0.179** [0.163, 0.193]	**0.082** [0.063, 0.101]	**0.055** [0.036, 0.081]
XGB	**0.170** [0.154, 0.185]	**0.083** [0.070, 0.104]	**0.080** [0.067, 0.106]
	IBM Dataset		
LR	**0.091** [0.044, 0.137]	**0.382** [0.326, 0.438]	**0.170** [0.093, 0.245]
XGB	**0.127** [0.082, 0.175]	**0.429** [0.368, 0.486]	**0.205** [0.125, 0.283]

Table 15: Baseline fairness metrics for a single sensitive attribute on the primary and IBM datasets. Bold values indicate an observed disparity that is unlikely under the permutation null ($\alpha = 0.05$), and the brackets denote the 95% confidence interval.

As the fairness metrics regarding a single sensitive attribute do not show the full picture, Table 16 exhibits the intersectional fairness metrics, where Int. DP is intersectional DP and SPSF is obtained using hard counts. SPSF measures the difference in probabilities between being assigned to class 1 given group membership and the overall probability of being assigned to class 1, weighted by the probability of belonging to a group. The value of SPSF has a similar interpretation to intersectional DP but is weighted by subgroup size. A low value of SPSF is desired. All metrics again show significant disparities. Moreover, the intersectional DP shows considerably higher values than the fairness metrics for a single sensitive attribute, indicating that there is a larger disparity between subgroups when the intersectional subgroups of all three sensitive variables are considered. However, this is also a combinatorial result, as it is expected that larger disparities will be observed when more subgroups with smaller sample sizes are considered. Subgroups with smaller sample sizes are more prone to sampling variability, potentially leading to larger disparities. Furthermore, LR outperforms XGB for both datasets.

Model	Int. DP	SPSF
	Primary Dataset	
LR	**0.339** [0.301, 0.392]	**0.007** [0.006, 0.009]
XGB	**0.340** [0.306, 0.392]	**0.009** [0.007, 0.010]
	IBM Dataset	
LR	**0.604** [0.566, 0.790]	**0.020** [0.015, 0.025]
XGB	**0.730** [0.650, 0.852]	**0.024** [0.019, 0.029]

Table 16: Baseline intersectional fairness metrics on the primary and IBM datasets. Bold values indicate an observed disparity that is unlikely under the permutation null ($\alpha = 0.05$), and the brackets denote the 95% confidence interval.

6.2 Fairness-Aware Models

The performance of the fairness-aware ML models with respect to single attribute DP is shown in Table 17. All methods considering intersectional subgroups (RW, DFM, and EG) show improvements with respect to all fairness metrics relative to the baselines for the primary dataset. For the primary dataset, DFM no longer exhibits significant disparities, indicating that the null hypothesis, which states that this value for the fairness metric is consistent with what would be observed under the permutation null, fails to be rejected. Moreover, for EG, DP Age and DP Education show no significant disparities. However, for DP Gender, the disparity remains significant. Additionally, EG Age and EG Education do not show significant disparities with respect to DP Age and DP Education, respectively. However, single attribute EG variants do not consistently decrease the fairness of the non-targeted sensitive attributes, occasionally increasing the fairness metrics compared to the baseline. This signals the need for fairness metrics considering intersectional subgroups. Notably, FE improves DP Gender and DP Age but increases the value for DP Education compared to the LR baseline.

For the IBM dataset, generally, all models that take the intersectional subgroups into account improve fairness for all three metrics, except RW for DP Education. Moreover, here RW, DFM, EG, and EG Gender no longer yield a significant difference for DP Gender. Additionally, DFM no longer shows a significant difference in DP Education. Remarkably, FE does not improve the fairness metrics with respect to the LR baseline, while it does improve the fairness metrics relative to XGB.

For both datasets, certain methods contain 0 in their confidence interval, e.g., DFM for DP Gender. This is because gender is a binary sensitive attribute, and hence, DP can be directional. A negative sign indicates that the selection rate of males is smaller than that of females. A confidence interval containing zero indicates that there is no evidence for the disparities being different from zero. Note that a confidence interval which contains zero and significance tests under the permutation null answer different questions. The confidence intervals show the sampling variability, while the permutations test assesses whether the observed disparity is likely under the permutation null, which does not imply that there is no unfairness in the classifier's predictions.

Method	DP Gender	DP Age	DP Education
	Primary Dataset		
LR (Baseline)	**0.179** [0.163, 0.193]	**0.082** [0.063, 0.101]	**0.055** [0.036, 0.081]
XGB (Baseline)	**0.170** [0.154, 0.185]	**0.083** [0.070, 0.104]	**0.080** [0.067, 0.106]
RW	**0.050** [0.034, 0.066]	**0.025** [0.010, 0.045]	**0.041** [0.019, 0.064]
DFM	0.003 [-0.013, 0.018]	0.004 [0.003, 0.027]	0.019 [0.008, 0.045]
EG	**0.018** [0.002, 0.033]	0.007 [0.002, 0.029]	0.015 [0.007, 0.040]
EG Gender	**0.024** [0.008, 0.040]	**0.079** [0.064, 0.099]	**0.079** [0.057, 0.103]
EG Age	**0.165** [0.149, 0.181]	0.015 [0.005, 0.035]	**0.083** [0.064, 0.105]
EG Education	**0.179** [0.162, 0.195]	**0.073** [0.060, 0.094]	0.018 [0.007, 0.044]
FE	**0.134** [0.118, 0.149]	**0.051** [0.036, 0.073]	**0.063** [0.043, 0.086]
	IBM Dataset		
LR (Baseline)	**0.091** [0.044, 0.137]	**0.382** [0.326, 0.438]	**0.170** [0.093, 0.245]
XGB (Baseline)	**0.127** [0.082, 0.175]	**0.429** [0.368, 0.486]	**0.205** [0.125, 0.283]
RW	0.041 [-0.005, 0.087]	**0.352** [0.297, 0.411]	**0.174** [0.094, 0.255]
DFM	-0.026 [-0.079, 0.021]	**0.193** [0.126, 0.260]	0.107 [0.062, 0.194]
EG	0.012 [-0.033, 0.062]	**0.311** [0.256, 0.375]	**0.136** [0.063, 0.222]
EG Gender	0.003 [-0.041, 0.048]	**0.412** [0.351, 0.478]	**0.221** [0.143, 0.300]
EG Age	**0.137** [0.087, 0.184]	**0.347** [0.291, 0.402]	**0.176** [0.110, 0.253]
EG Education	**0.145** [0.097, 0.196]	**0.390** [0.333, 0.453]	**0.137** [0.061, 0.224]
FE	**0.101** [0.054, 0.143]	**0.388** [0.330, 0.448]	**0.170** [0.097, 0.246]

Table 17: Fairness metrics for a single sensitive attribute method on the primary and IBM datasets. Bold values indicate an observed disparity that is unlikely under the permutation null ($\alpha = 0.05$), and the brackets denote the 95% confidence interval.

Table 18 shows that all fairness-aware ML models improve the fairness metrics, except for EG Education on SPSF, relative to the baselines for the primary dataset. DFM obtains the lowest fairness metrics and the fairness metrics do not significantly exceed the values obtained by the permutations on the primary dataset for both fairness metrics. It should be noted that no significance does not indicate fairness, but rather shows a lack of evidence that the disparity exceeds what would be obtained under the permutation null. Additionally, for the primary dataset, FE improves the fairness metrics compared to the baselines. However, the magnitude is smaller compared to fairness-aware ML models that consider intersectional subgroups. Furthermore, EG Gender reduces intersectional DP more than FE, signaling the weak performance of FE.

For the IBM dataset, DFM and EG show no significant disparity for intersectional DP. Hence, these methods improve fairness concerning all sensitive attributes compared to the baselines. Moreover, for the IBM dataset, all methods that take the subgroups into account improve both fairness metrics relative to the baselines, while the fairness-aware ML models

that focus on a single sensitive attribute often increase unfairness with respect to LR. Only the single sensitive attribute method EG Gender improves the fairness metrics compared to the baselines.

Additionally, all confidence intervals seem to be asymmetric around the point estimate. This is likely because the fairness metrics are bounded quantities and take the max-min difference. Hence, the bootstrap samples are expected to be right-skewed.

Method	Int. DP	SPSF
Primary Dataset		
LR (Baseline)	**0.339** [0.301, 0.392]	**0.007** [0.006, 0.009]
XGB (Baseline)	**0.340** [0.306, 0.392]	**0.009** [0.007, 0.010]
RW	**0.152** [0.128, 0.210]	**0.003** [0.003, 0.005]
DFM	0.086 [0.081, 0.157]	0.002 [0.002, 0.004]
EG	**0.111** [0.087, 0.169]	**0.003** [0.002, 0.004]
EG Gender	**0.188** [0.173, 0.254]	**0.005** [0.004, 0.006]
EG Age	**0.275** [0.266, 0.339]	**0.006** [0.006, 0.008]
EG Education	**0.290** [0.272, 0.348]	**0.007** [0.006, 0.009]
FE	**0.244** [0.226, 0.308]	**0.006** [0.005, 0.008]
IBM Dataset		
LR (Baseline)	**0.604** [0.566, 0.790]	**0.020** [0.015, 0.025]
XGB (Baseline)	**0.730** [0.650, 0.852]	**0.024** [0.019, 0.029]
RW	**0.528** [0.496, 0.721]	**0.016** [0.011, 0.021]
DFM	0.491 [0.411, 0.741]	**0.015** [0.011, 0.020]
EG	0.465 [0.442, 0.680]	**0.014** [0.010, 0.020]
EG Gender	**0.598** [0.552, 0.758]	**0.019** [0.015, 0.024]
EG Age	**0.669** [0.601, 0.814]	**0.023** [0.017, 0.027]
EG Education	**0.677** [0.616, 0.827]	**0.022** [0.017, 0.027]
FE	**0.669** [0.603, 0.815]	**0.022** [0.017, 0.027]

Table 18: Intersectional fairness metrics by method on the primary and IBM datasets. Bold values indicate an observed disparity that is unlikely under the permutation null ($\alpha = 0.05$), and the brackets denote the 95% confidence interval.

Table 19 shows the most disadvantaged group per model according to DP. The groups are encoded as Gender | Age | Education, where F | $18-30$ | Edu1 indicates a female aged $18-30$ with education level 1. This corresponds to the group yielding the lowest selection rate, that is, the probability of being assigned to the class of stayers given membership in a specific group. In the primary dataset, the baseline models showed the same group with respect to gender and age; however, they disagree on education. This shows that the most disadvantaged group is not necessarily a data property but differs per model. Nevertheless,

both models agree that females in the age range of $18-30$ are the most disadvantaged group. Several fairness-aware ML models show slightly different groups, this is because these models actively try to mitigate the differences in selection rates between groups. This might change the most disadvantaged group.

For the IBM dataset, the most disadvantaged group is the same for all methods, which indicates that this group is the worst off in the IBM dataset. Moreover, no fairness-aware ML model considered here changes that. Nonetheless, on the group level, both datasets mostly one young females aged between $(18-30)$, as the most disadvantaged group, indicating that fairness interventions reduce the magnitude of unfairness but do not remove the disadvantage with respect to females.

Method	Most Disadvantaged Group
Primary Dataset	
LR (Baseline)	F | $18-30$ | Edu1
XGB (Baseline)	F | $18-30$ | Edu3
RW	F | $18-30$ | Edu1
DFM	F | 46+ | Edu4
EG	F | $18-30$ | Edu3
EG Gender	F | $18-30$ | Edu1
EG Age	F| $31-45$ | Edu1
EG Education	F | $18-30$ | Edu3
FE	F | $18-30$ | Edu1
IBM Dataset	
LR (Baseline)	F | $18-30$ | Edu3
XGB (Baseline)	F | $18-30$ | Edu3
RW	F | $18-30$ | Edu3
DFM	F | $18-30$ | Edu3
EG	F | $18-30$ | Edu3
EG Gender	F | $18-30$ | Edu3
EG Age	F | $18-30$ | Edu3
EG Education	F | $18-30$ | Edu3
FE	F | $18-30$ | Edu3

Table 19: Most disadvantaged group as per DP for both datasets.

6.3 Predictive Performance of Fairness-Aware Models

The fairness-aware ML models have been shown to decrease the level of unfairness, as noted in the previous section. However, the predictive performance is of utmost importance. Hence, Table 20 shows the predictive performance of the models. For the primary dataset,

overall, the models exhibit little loss of predictive performance compared to the XGB baseline, and generally, the methods obtain better performance than the LR baseline. Notably, DFM obtains higher recall and lower precision than both baselines, where recall can be viewed as important since it identifies the number of leavers captured correctly.

For the IBM dataset, accuracy drops by at most 0.082. However, accuracy is not a reliable evaluation metric in the case of an imbalanced dataset, so more emphasis is placed on the other evaluation metrics. Importantly, precision generally drops compared to LR, but overall increases compared to XGB, despite most methods using XGBoost as a classifier. Nevertheless, these improvements are less than one standard deviation away from the baseline and are not consistent across methods. Therefore, these gains should not be interpreted as systematic gains. Additionally, DFM shows an increase in recall. Lastly, for F1, as it combines precision and recall, one can see the same patterns. Overall, the fairness-aware ML models do not exhibit significant drops in performance metrics, occasionally increasing. However, these changes mostly fall within a one standard deviation difference compared to the baselines, indicating no considerable change in predictive performance, except for DFM.

Method	Accuracy	Precision	Recall	F1-score
Primary Dataset				
LR (Baseline)	0.708 ± 0.004	0.698 ± 0.004	0.671 ± 0.006	0.684 ± 0.004
XGB (Baseline)	0.716 ± 0.004	0.704 ± 0.004	0.686 ± 0.005	0.695 ± 0.004
RW	0.713 ± 0.003	0.701 ± 0.004	0.681 ± 0.005	0.691 ± 0.004
DFM	0.701 ± 0.004	0.674 ± 0.004	0.708 ± 0.005	0.691 ± 0.004
EG	0.713 ± 0.003	0.703 ± 0.004	0.677 ± 0.005	0.690 ± 0.004
EG Gender	0.714 ± 0.003	0.704 ± 0.004	0.681 ± 0.005	0.692 ± 0.004
EG Age	0.713 ± 0.004	0.702 ± 0.004	0.680 ± 0.005	0.691 ± 0.004
EG Education	0.715 ± 0.003	0.703 ± 0.004	0.687 ± 0.005	0.695 ± 0.004
FE	0.716 ± 0.003	0.705 ± 0.004	0.684 ± 0.005	0.694 ± 0.004
IBM Dataset				
LR (Baseline)	0.711 ± 0.012	0.295 ± 0.015	0.571 ± 0.031	0.389 ± 0.019
XGB (Baseline)	0.683 ± 0.012	0.276 ± 0.014	0.592 ± 0.031	0.376 ± 0.018
RW	0.702 ± 0.012	0.287 ± 0.015	0.570 ± 0.030	0.382 ± 0.019
DFM	0.629 ± 0.013	0.245 ± 0.011	0.624 ± 0.029	0.352 ± 0.015
EG	0.720 ± 0.011	0.301 ± 0.016	0.558 ± 0.030	0.391 ± 0.020
EG Gender	0.704 ± 0.012	0.291 ± 0.015	0.579 ± 0.031	0.387 ± 0.019
EG Age	0.703 ± 0.012	0.283 ± 0.016	0.549 ± 0.031	0.374 ± 0.020
EG Education	0.696 ± 0.012	0.282 ± 0.015	0.575 ± 0.031	0.379 ± 0.019
FE	0.704 ± 0.012	0.288 ± 0.015	0.567 ± 0.031	0.382 ± 0.019

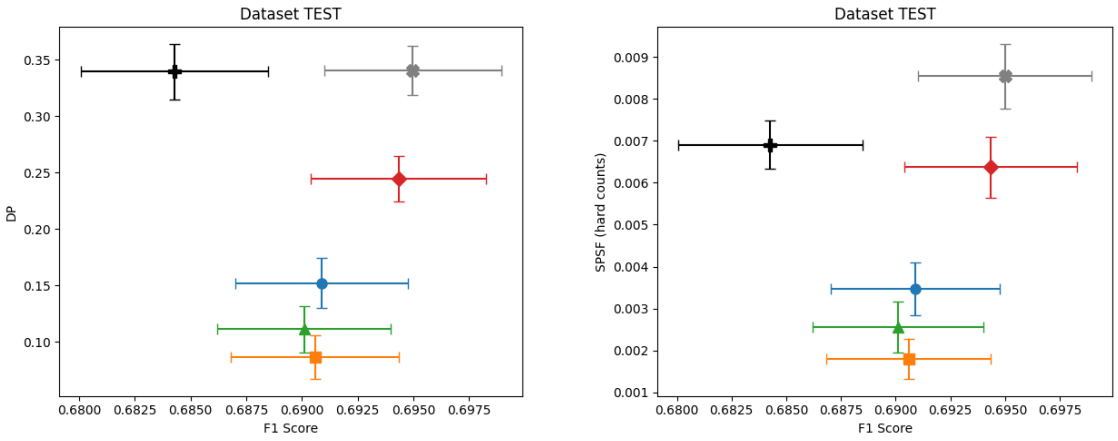
Table 20: Predictive performance by method for both datasets.

6.4 Fairness-Performance Trade-Off

In order to directly compare the different models, Figure 5 shows all methods with the F1 score on the x-axis and a fairness metric on the y-axis. Moreover, the error bars denote one standard deviation. The SPSF fairness metric is computed with hard counts. Models in the lower right are preferred, as they exhibit the highest F1 score and the lowest fairness metrics. The single sensitive attribute models are omitted, as they do not consider intersectional subgroups which is what this thesis researches.

The figure shows that, for the primary dataset, RW, DFM, and EG are placed close together, all dominating the LR baseline and substantially improving fairness at minimal predictive performance cost. For example, DFM reduces intersectional DP from 0.340 to 0.086, while the F1 score decreases only by 0.004 compared to the XGB baseline. A similar result is noted for EG. Moreover, FE is placed between the fairness-aware ML methods and the baseline methods in terms of the fairness metric, which further indicates its inferior performance. In the figure, no model dominates all other models. Additionally, the error bars for DFM and EG overlap on both axes, and the error bars for RW and EG overlap on both axes as well. The figure for SPSF shows the same pattern. This result indicates that there is no method that performs universally best on the primary dataset.

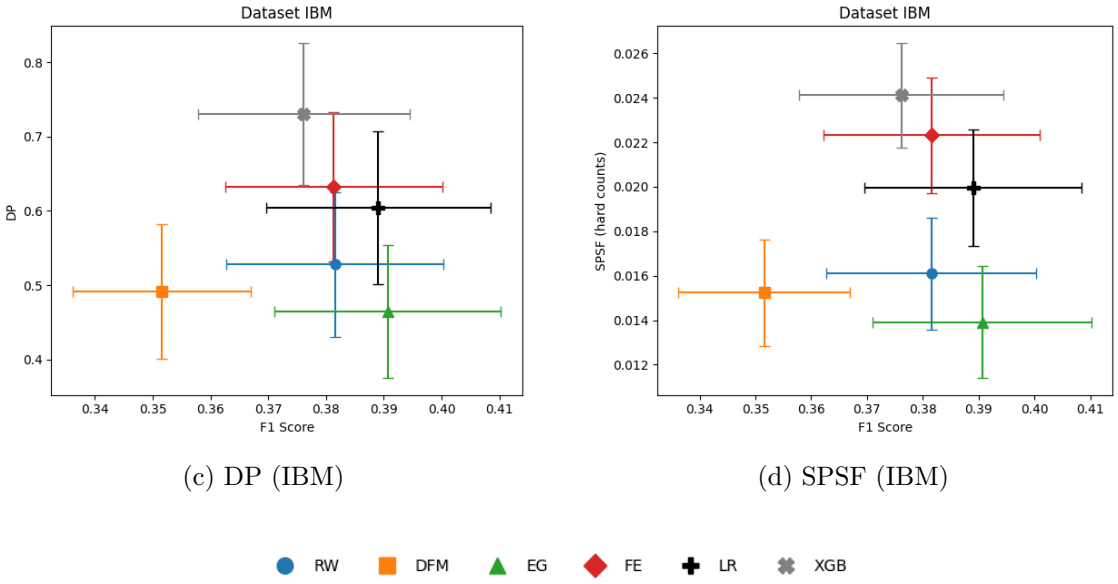
The results on the IBM dataset show a different pattern. The models are more clustered together, and DFM yields the lowest F1 score. This is potentially because the models are trained on a different dataset, and the IBM dataset shows signs of a different data distribution. Moreover, EG dominates all points, yielding the lowest values for the fairness metrics while not decreasing the predictive performance with respect to the baseline, indicating that it performs the best given a different data distribution. Additionally, FE increases the magnitude of the fairness metrics relative to LR, which indicates poor performance on different data distributions.



(a) DP (Test)

(b) SPSF (Test)

Figure 5: Fairness–performance trade-off results.



(c) DP (IBM)

(d) SPSF (IBM)

Figure 5: Fairness–performance trade-off results (continued).

Figure 5 identifies the most promising methods for each dataset. For the primary dataset, the methods are RW, DFM, and EG, as they show the lowest fairness metric at a low cost for predictive performance. For the IBM dataset, the methods are RW, EG, and the baseline LR. Although EG is the dominating model, it shows overlapping error bars with RW and LR. The Tables below assess whether there are significant differences in F1 and the fairness metric according to the procedure discussed in Section 5.3.2. If zero is not included in the confidence interval, H_0 is rejected. On the other hand, if the confidence interval includes zero, the null hypothesis is not rejected, but this does not imply that the true difference is zero. Moreover, no correction was applied for conducting multiple tests. As multiple tests are performed simultaneously, certain results could be significant by chance alone.

On the primary dataset, DFM and EG show significant differences compared to RW with respect to intersectional DP, as the confidence intervals of the differences do not include zero. However, DFM and EG do not exhibit significant differences with respect to each other. Moreover, for SPSF, the confidence interval of the difference between DFM and RW includes zero, so the null hypothesis cannot be rejected at the 5% significance level. This can happen even though in the plot they do not appear to have overlapping error bars concerning SPSF, as the plot and the test are not equivalent. Additionally, all confidence intervals concerning the F1 score include zero. Although this does not prove that the difference in the F1 score is zero, it implies comparable predictive performance, which is in line with the overlapping error bars with respect to the F1 score in Figure 5.

For the IBM dataset, the confidence intervals of the differences between EG and RW include zero for the fairness metrics, and hence H_0 is not rejected. Therefore, there is no evidence that the fairness metrics obtained by EG and RW are different. However, the difference between EG and LR, does not include zero for both fairness metrics, and hence the null hypothesis is rejected. The same holds for the confidence interval of the differences

for RW and LR for SPSF. Additionally, similar to the results for the primary dataset, zero is included in the confidence intervals on the differences in F1 scores across all model comparisons. This again aligns with the overlapping error bars in Figure 5. Note that for the IBM dataset, certain subgroups have a small number of observations. While SPSF is robust to this, the stratified bootstrap samples may still be unstable with the predicted probabilities for these subgroups. This should be kept in mind when interpreting the results of Table 21.

Metric	Comparison	95% CI for Difference
Primary Dataset		
Int. DP	EG – RW	**[−0.073, −0.014]**
Int. DP	EG – DFM	[−0.036, 0.051]
Int. DP	RW – DFM	**[0.006, 0.094]**
SPSF	EG – RW	**[−0.00147, −0.00009]**
SPSF	EG – DFM	[−0.00080, 0.00171]
SPSF	RW – DFM	[−0.00004, 0.00255]
F1-score	EG – RW	[−0.00413, 0.00236]
F1-score	EG – DFM	[−0.00644, 0.00547]
F1-score	RW – DFM	[−0.00570, 0.00558]
IBM Dataset		
Int. DP	EG – RW	[−0.140, 0.028]
Int. DP	EG – LR	**[−0.259, −0.006]**
Int. DP	RW – LR	[−0.189, 0.011]
SPSF	EG – RW	[−0.00462, 0.00049]
SPSF	EG – LR	**[−0.00942, −0.00215]**
SPSF	RW – LR	**[−0.00731, −0.00049]**
F1-score	EG – RW	[−0.00649, 0.02643]
F1-score	EG – LR	[−0.01876, 0.02327]
F1-score	RW – LR	[−0.02502, 0.01177]

Table 21: Pairwise bootstrap confidence intervals for differences between methods. Confidence intervals are computed for the difference in the reported metric between the compared methods. Bold confidence intervals indicate confidence intervals that do not include zero, rejecting H_0 at a 5% significance level.

Chapter 7

Discussion

The main results show that methods considering intersectional subgroups significantly reduce disparities with respect to the baselines on the primary dataset, where the cost in F1 score is negligible. A second finding is that there is not enough evidence to conclude that the disparities obtained by DFM are significantly different from those obtained by EG on the primary dataset, and no method dominates on both the primary and the IBM dataset.

This chapter discusses these observed results and patterns, where exceptions are noted and discussed as well. Lastly, this chapter discusses the limitations of this thesis and proposes directions for future work.

7.1 Baseline Disparities

The results for the baseline models established that the standard ML models show a significant disparity for all single attribute DP metrics. However, the magnitude of DP becomes larger when all intersectional subgroups are considered. For the primary dataset, DP Gender yields a value of 0.179 for the LR baseline, while intersectional DP attains a value of 0.339 for the LR baseline. This is an example of aggregation bias discussed in Section 2.1. Hence, the baselines not only show that there is substantial unfairness but also clearly indicate the need for intersectional analysis. Furthermore, the difference in the most disadvantaged subgroup identified between baselines shows that this property is at least partly a consequence of the choice of model in this dataset.

7.2 Fairness-Aware Methods on the Primary Dataset

The fairness-aware ML models decrease the level of unfairness for all intersectional fairness metrics relative to the baselines, except for EG Education in SPSF. Notably, DFM reduces the fairness metric the most and yields values that do not significantly exceed those obtained by permutations. It should be noted that a non-significant result indicates insufficient evidence that the obtained fairness metric exceeds the fairness metric expected under the permutation null. Nevertheless, the result indicates strong improvement with respect to the fairness metrics. A possible explanation for why DFM obtains non-significance is that DFM does not use intersectional DP directly as a fairness metric to minimize, but uses DF with soft counts. It actively tries to make the probabilities of being assigned to class 1 closer to each other, instead of the predictions themselves, which is done by EG. Hence, the choice of DF as a fairness criterion in the objective function for DFM is a possible reason for the model performing well in terms of minimizing intersectional DP and SPSF. However, the success of DFM cannot be entirely credited to the metric DF, as DFM also uses a NN as a predictor, which may provide additional advantages over, for example, XGBoost.

The second highest performing model in terms of intersectional fairness is EG, which is another in-processing model like DFM. However, the improvement in the fairness met-

rics is of smaller magnitudes, which is potentially due to the use of the fairness metric intersectional DP, as opposed to DF with soft counts.

Similar to DFM and EG, RW reduces intersectional DP compared to the baselines, but it attains higher intersectional DP than DFM and EG. A probable cause for this is that RW does not explicitly optimize over intersectional subgroups. However, the weights assigned according to the group-label pairs improve the fairness metric by a substantial amount relative to the baselines. RW is computationally cheap, but it comes with the cost of more unfairness compared to other methods considering intersectional subgroups.

Regarding SPSF, the models perform similarly compared to each other, with LR now performing marginally worse than FE in terms of SPSF. Lastly, all these methods improve the single attribute fairness metrics compared to the baselines.

The results of the test for detecting differences in fairness metrics from Section 5.3.2 show that the ranking of the methods is not solely based on the point estimates for the fairness metrics. The disparities obtained for DFM are shown not to be significantly different from those obtained by EG for the fairness metrics. Moreover, DFM and RW do differ significantly with respect to intersectional DP but not for SPSF. Additionally, EG and RW significantly differ for both fairness metrics. This result further implies that neither DFM nor EG outperforms the other.

The need for intersectional subgroup aware methods is shown by the results for the single attribute EG variants. These methods generally decrease single sensitive attribute unfairness and the intersectional fairness metrics. EG Gender reduces intersectional DP considerably, which is likely caused by the dominant disparity within the sensitive variable gender, as highlighted in Section 3.5. The other two single attribute EG variants also decrease intersectional DP relative to the baseline XGB. Nevertheless, all these methods exhibit aggregation bias, as they appear to have fairly low DP values for their targeted sensitive variable, while the intersectional DP values exhibit larger values. This result is partially expected, as considering subgroups at a more granular level means each subgroup contains fewer observations. Consequently, the disparities estimated from these smaller subgroups are likely larger due to sampling variability.

The novel FE shows interesting results, which indicate that methods blind to intersectional subgroups perform worse than those that explicitly consider intersectional subgroups. Although FE reduces both intersectional fairness metrics compared to the baseline, EG Gender reduces these metrics further than FE. This shows that the assumption that an average over fairness-aware ML models with respect to a single sensitive attribute should be fair with respect to the intersectional subgroups does not hold in practice. Every model only enforces fairness upon the targeted sensitive attribute, and the average of the three models will carry over some of these improvements. However, the fact that the intersectional subgroups are never considered is likely a bottleneck in this approach.

7.3 Evaluation of the Utility-Fairness Trade-Off

All the fairness metric improvements come with a negligible predictive performance cost on the primary dataset. The models yielding the lowest fairness metrics (RW, EG, and DFM) show only a maximum decrease in F1 of 0.005. Moreover, F1 scores do not differ significantly for these models. Therefore, there is hardly any trade-off with respect to the F1 score in this setting. The best model, then, is the one that obtains the lowest fairness metric, which is either DFM or EG, as they do not show significant differences

between the fairness metrics. However, when the evaluation metrics of precision or recall are considered individually, there is a trade-off. For instance, DFM decreases precision but increases recall compared to other methods, while EG does not exhibit this behavior. As noted earlier, recall is important in the context of employee attrition prediction, as it captures the model's ability to classify the leavers correctly. Moreover, DFM was shown to exhibit fairness metrics that did not significantly exceed the fairness metrics obtained from the permutations for the primary dataset, which could be a reason to select DFM. Lastly, the fact that the cost in predictive performance is low is potentially because the primary dataset is balanced with a 48% attrition rate, which may not appear in real-world datasets as well.

7.4 Generalization to the IBM Dataset

IBM results do not measure how well fairness interventions generalize, instead, they measure how models perform under a distribution shift. The differences in fairness metrics and performance metrics are due to an interaction between the IBM data distribution and the models. The result that there is no universal winner is supported by the fact that EG dominates all other models on the IBM dataset, as can be seen in Figure 5. However, it does not dominate all models on the primary dataset. Moreover, EG and DFM exhibit intersectional DP that do not significantly exceed the intersectional DP obtained under permutations. Again, this result indicates a lack of evidence that the fairness metric obtained by the model exceeds the value of the fairness metric expected under the permutation null. A plausible explanation for EG's dominance is that EG uses a randomized classifier, which relies less on noisy dataset features. As a consequence, EG relies on the features differently and, therefore, potentially obtains lower values for the fairness metrics in this setting.

Notably, DFM attains the lowest F1 score, while it yields the highest F1 score out of the intersectional-aware methods on the primary dataset. However, this is possibly due in part to how the model is trained. The NN was trained in batches with a balanced attrition rate. The model predicts attrition more often than what is common in the IBM dataset. A small false positive rate leads to a large number of false positives, which heavily influences precision and, consequently, also the F1 score. Nevertheless, this is likely not a fundamental property of using DF as a regularizer.

Moreover, FE shows a large change in performance when applied to the IBM dataset. FE improves all fairness metrics compared to the baseline XGB but not with respect to LR. Furthermore, there is an additional explanation beyond the primary bottleneck of FE. For the IBM dataset, the sensitive attribute age is the most dominant one, as DP Age is considerably higher than DP Gender and DP Education. Since FE assigns equal weights, each sensitive attribute is considered equally important. However, for the IBM dataset, FE possibly assigns too large weights for gender and education relative to the disparities observed in the dataset, which are analyzed in Section 3.5. This is an untested hypothesis and is further addressed in Section 7.7. This potentially further explains the poor performance of FE on the IBM dataset. Therefore, FE offers a framework that is cheap to compute given the single sensitive attribute but should not be used as a replacement for methods that consider intersectional subgroup fairness in this setting.

7.5 Disadvantaged Subgroups

All methods show females as the most disadvantaged group. This is in line with historical biases against females, which are highlighted in Section 3.2. However, the fairness-aware ML models show that in certain cases, the most disadvantaged group can shift. In such cases, the maximum differences per group shift, and another group then becomes the most disadvantaged. For the IBM dataset, no fairness-aware ML model shifts the most disadvantaged group, while most methods decrease the level of unfairness relative to the baselines. This is explained by how the fairness metrics work. All fairness metrics used in the models take the worst case. Therefore, only the disparity itself is minimized, and this does not necessarily change the most disadvantaged subgroup.

7.6 Limitations

The results obtained in this thesis stem from models trained on synthetic data. Bias in the training data might not be real historical bias but rather may be due to the data generating process. Therefore, the results obtained in this thesis may not transfer to HR datasets from real organizations.

Furthermore, to test for significance between metrics obtained by two models, bootstrap samples are used. However, the models are not retrained using these bootstrap samples. Therefore, the test only captures uncertainty from the data and not from the model training and hyperparameter optimization. Additionally, 500 permutations and bootstrap samples were used, increasing this value further would yield more stable p-values and confidence intervals.

All fairness metrics are computed based on the 0.5 classifier threshold. Changing this threshold is a well known post-processing technique [23] that was deliberately excluded. Nevertheless, altering the threshold changes the precision-recall balance for every method, which might change the ranking of the models based on the fairness metrics.

Lastly, the equal weights in FE are a design choice in this research; the model might benefit from adjusted weighting. Varying the weights per method, where attributes with a higher disparity receive a higher weight, could improve FE.

7.7 Future Work

The results and limitations of this research provide pathways for new research. A natural direction is to validate the results obtained in this thesis on real-world HR datasets. This includes validating whether the method rankings observed in this thesis are preserved and verifying similar trade-offs. Additionally, the FE could be investigated using various methods for weighting the different classifiers, as this could possibly result in better fairness performance. Lastly, this thesis proposes research on the evaluation of other fairness metrics. This thesis follows the framework of Wachter *et al.* [54], which results in considering only bias transforming fairness metrics. However, this research fails to detect models where the TPR and FPR differ between subgroups. Hence, EO would be an interesting metric to evaluate in future research. Additionally, the bias transforming metrics such as FTA and counterfactual fairness were not evaluated due to their reliance on appropriate distance metrics and a SCM, respectively. However, as they take a different perspective on how they regard fairness, evaluating these fairness metrics would yield interesting research.

Chapter 8

Conclusion

This thesis investigated the trade-offs between predictive performance and fairness that pre- and in-processing fairness mitigation methods provide. This chapter concludes the answers to the research question and sub-questions.

To answer the first sub-question, *"How do baseline employee attrition prediction models perform in terms of predictive performance and intersectional fairness?"*, the baselines are assessed here. XGB outperforms LR on the primary dataset in terms of predictive performance, and both baselines exhibit values for the intersectional fairness metrics that were found to be larger compared to the values obtained by permutations. The single attribute fairness metrics appear modest in comparison to intersectional DP, as intersectional DP yields almost double the largest single sensitive attribute fairness metric. Therefore, there is a difference when considering intersectional subgroups, which implies evidence of Simpson's Paradox and the need for intersectional fairness analysis. On the IBM dataset, the pattern changes; DP Gender shrinks while DP Age increases substantially. This is due to an interaction of the methods and the differences between the IBM and the primary dataset discussed in Chapter 3.

The second sub-question *"How do pre- and in-processing fairness mitigation methods reduce unfairness for various sensitive attributes and their intersectional subgroups in employee attrition prediction?"*, is addressed by discussing the performance of these methods. All intersectional-aware methods reduce every fairness metric compared to the baseline on the primary dataset, which includes both single sensitive attribute and intersectional fairness metrics. DFM shows the strongest decrease relative to the baselines for the primary dataset. DFM yields values for which the null hypothesis cannot be rejected, indicating insufficient evidence that the observed fairness metric exceed the fairness metrics expected under the permutation null. EG also shows promising results, while RW yields the weakest result among the intersectional-aware methods on the primary dataset but is inexpensive to compute. Furthermore, the single-attribute EG methods improve fairness concerning their targeted sensitive attribute compared to the baselines, but they often leave intersectional DP high. The novel FE method shows mixed results. On the primary dataset, it improves fairness compared to the baselines but underperforms compared to EG Gender. Additionally, for the IBM dataset, it occasionally worsens intersectional fairness compared to the baselines. Hence, the results do not support the assumption that averaging single-attribute fair classifiers yields intersectional fairness.

The final sub-question, *"What trade-offs between predictive performance and intersectional fairness can be identified when applying the fairness mitigation methods compared to the baseline employee attrition prediction models?"* can be answered by interpreting the fairness-aware ML models. The increase in fairness comes with a negligible cost in F1 score for the primary dataset. The maximum drop in F1 score is 0.005 for the intersectional-aware methods. Additionally, there is a lack of a detected difference between the F1 scores of these methods. DFM increases recall, which is a relevant performance metric in the context of employee attrition prediction. The results on the IBM datasets show different trade-offs, as the methods do not demonstrate the same performance compared to one

another. EG dominates all other models as it increases the F1 score and shows the lowest values for both intersectional fairness metrics.

This thesis ultimately aims to answer the research question: *"What trade-offs do pre- and in-processing fairness mitigation methods provide between predictive performance and intersectional fairness in employee attrition prediction?"* To answer this, the answers to the sub-questions are utilized. These findings show that pre- and in-processing fairness mitigation methods that target intersectional subgroups can substantially improve fairness relative to the baselines. The cost in F1 score is negligible for the primary dataset and generally small for the IBM dataset. DFM is the exception, which yields a 0.037 decrease in F1 score compared to LR on the IBM dataset. To conclude, in the context of this thesis, there is a minimal trade-off with respect to F1 score, as fairness can be improved without significantly reducing the F1 score.

A second conclusion is that no method yields universally the best performance, which can be seen by comparing method performance and considering the different datasets. The intersectional-aware methods exhibit similar performance on the primary dataset. Especially, DFM and EG achieve similar intersectional DP and SPSF, while both yield comparable F1 scores. The test that assesses the difference in intersectional DP and SPSF indicates that there is not enough evidence to prove that there is a difference between those two metrics for the methods DFM and EG. Additionally, for the IBM dataset, EG dominates all methods, while DFM yields the lowest F1 score. Moreover, single-attribute EG models perform worse compared to intersectional-aware methods. Combining such methods, as done by the FE, does not consistently improve intersectional fairness relative to the baselines. Moreover, for the methods themselves, there is a trade-off, as DFM attains higher recall compared to the baselines, but its precision is the lowest among all other methods for both datasets, particularly in the IBM dataset. Hence, no model is universally the best method in this context, and the preferred method depends on the preferences of the decision maker. For instance, a preference for high recall might lead a decision maker to choose the DFM method.

Finally, this thesis provides an empirical comparison of pre- and in-processing fairness-aware ML models for intersectional fairness in the employee attrition setting. Additionally, a novel FE method is proposed and assessed, which provides clear directions, such as different weights, for future research that could aid in improving the method's performance. Despite its potential classification as high-risk under the AI Act [49], this setting has not been previously evaluated on intersectional subgroups. This research demonstrates the risk of relying solely on single sensitive attributes when aggregation bias is present. Moreover, the findings from this thesis are relevant both to the fairness-aware ML literature and to organizations such as EY, as it evaluates models for employee attrition prediction and measures fairness.

References

[1] Agarwal, A., Beygelzimer, A., Dudík, M., Langford, J., & Wallach, H. (2018). A reductions approach to fair classification. *Proceedings of the 35 th International Conference on Machine Learning*. https://doi.org/10.48550/arXiv.1803.02453

[2] Akiba, T., Sano, S., Yanase, T., Ohta, T., & Koyama, M. (2019). Optuna: A next-generation hyperparameter optimization framework. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2623–2631. https://doi.org/10.1145/3292500.3330701

[3] Barocas, S., Hardt, M., & Narayanan, A. (2023). *Fairness and machine learning: Limitations and opportunities*. MIT Press. https://fairmlbook.org

[4] Blau, F. D., & Kahn, L. M. (2017). The gender wage gap: Extent, trends, and explanations. *Journal of Economic Literature*, *55*(3), 789–865. https://doi.org/10.1257/jel.20160995

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

[6] Calmon, F. P., Wei, D., Vinzamuri, B., Ramamurthy, K. N., & Varshney, K. R. (2017). Optimized pre-processing for discrimination prevention. *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 3995–4004. https://api.semanticscholar.org/CorpusID:3801798

[7] Casella, G., & Berger, R. L. (2002). *Statistical inference* (2nd). Duxbury Press. https://doi.org/https://doi.org/10.1201/9781003456285

[8] Castelnovo, A., Crupi, R., Greco, G., Regoli, D., Penco, I. G., & Cosentini, A. C. (2022). A Clarification of the Nuances in the Fairness Metrics Landscape. *Scientific Reports*, *12*(1), 4209. https://doi.org/10.1038/s41598-022-07939-1

[9] Caton, S., & Haas, C. (2024). Fairness in Machine Learning: A Survey. *ACM Computing Surveys*, *56*(7), 1–38. https://doi.org/10.1145/3616865

[10] Charter of Fundamental Rights of the European Union (2012). https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:C2012/326/02

[11] Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. https://doi.org/10.1145/2939672.2939785

[12] Chen, Z., Zhang, J., Sarro, F., & Harman, M. (2024). Fairness improvement with multiple protected attributes: How far are we? *2024 IEEE/ACM 46th International Conference on Software Engineering (ICSE '24)*, 1–13. https://doi.org/10.1145/3597503.3639083

[13] Chouldechova, A. (2016). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data*, *5*, 153–163. https://api.semanticscholar.org/CorpusID:1443041

[14] Council Directive 2000/78/EC of 27 November 2000 establishing a general framework for equal treatment in employment and occupation (2000). https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32000L0078

[15] Creager, E., Madras, D., Jacobsen, J.-H., Weis, M. A., Swersky, K., Pitassi, T., & Zemel, R. S. (2019). Flexibly fair representation learning by disentanglement. *International Conference on Machine Learning*. https://api.semanticscholar.org/CorpusID:174800294

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

[17] England, P., & Buldig, M. J. (2001). The wage penalty for motherhood. *American Sociological Review*, *66*, 204–225. https://doi.org/10.2307/2657415

[18] Directive 2006/54/EC of the European Parliament and of the Council of 5 July 2006 on the implementation of the principle of equal opportunities and equal treatment of men and women in matters of employment and occupation (recast) (2006). https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32006L0054

[19] Fallucchi, F., Coladangelo, M., Giuliano, R., & Luca, E. W. D. (2020). Predicting Employee Attrition Using Machine Learning Techniques. *Computers*, *9*(4). https://doi.org/10.3390/computers9040086

[20] Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., & Venkatasubramanian, S. (2015). Certifying and Removing Disparate Impact. *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 259–268. https://doi.org/10.1145/2783258.2783311

[21] Foulds, J. R., Islam, R., Keya, K. N., & Pan, S. (2020). An intersectional definition of fairness. *2020 IEEE 36th International Conference on Data Engineering (ICDE)*, 1918–1921. https://doi.org/10.1109/ICDE48307.2020.00203

[22] García, S., Ramírez-Gallego, S., Luengo, J., Benítez, J. M., & Herrera, F. (2016). Big data preprocessing: Methods and prospects. *Big Data Analytics*, *1*, 1–22. https://api.semanticscholar.org/CorpusID:16978267

[23] Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. *Proceedings of the 30th International Conference on Neural Information Processing Systems*, 3323–3331. https://doi.org/10.48550/arXiv.1610.02413

[24] Hort, M., Chen, Z., Zhang, J. M., Harman, M., & Sarro, F. (2024). Bias Mitigation for Machine Learning Classifiers: A Comprehensive Survey. *ACM J. Responsib. Comput.*, *1*(2), 11:1–11:52. https://doi.org/10.1145/3631326

[25] Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression* (3rd). John Wiley & Sons. https://doi.org/10.1002/9781118548387

[26] Jain, R., & Nayyar, A. (2018). Predicting employee attrition using xgboost machine learning approach. *Proc. Int. Conf. Syst. Model. Adv. Res. Trends, SMART*, 113–120. https://doi.org/10.1109/SYSMART.2018.8746940

[27] Kamiran, F., & Calders, T. (2009). Classifying without discriminating. *2009 2nd International Conference on Computer, Control and Communication*, 1–6. https://api.semanticscholar.org/CorpusID:1102398

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

[29] Kamiran, F., Calders, T., & Pechenizkiy, M. (2010). Discrimination Aware Decision Tree Learning. *Proceedings of the 2010 IEEE International Conference on Data Mining*, 869–874. https://doi.org/10.1109/ICDM.2010.50

[30] Kamishima, T., Akaho, S., Asoh, H., & Sakuma, J. (2012). Fairness-Aware Classifier with Prejudice Remover Regularizer. In P. A. Flach, T. De Bie, & N. Cristianini (Eds.), *Machine Learning and Knowledge Discovery in Databases* (pp. 35–50). Springer. https://doi.org/10.1007/978-3-642-33486-3_3

[31] Kearns, M., Neel, S., Roth, A., & Wu, Z. S. (2018, October). Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In J. Dy & A. Krause (Eds.), *Proceedings of the 35th international conference on machine learning* (pp. 2564–2572, Vol. 80). PMLR. https://proceedings.mlr.press/v80/kearns18a.html

[32] Kusner, M., Loftus, J., Russell, C., & Silva, R. (2017). Counterfactual fairness. *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 4069–4079. https://dl.acm.org/doi/10.5555/3294996.3295162

[33] Le Quy, T., Roy, A., Iosifidis, V., Zhang, W., & Ntoutsi, E. (2022). A survey on datasets for fairness-aware machine learning. *WIREs Data Mining and Knowledge Discovery*, *12*(3), e1452. https://doi.org/https://doi.org/10.1002/widm.1452

[34] Louizos, C., Swersky, K., Li, Y., Welling, M., & Zemel, R. (2016). The variational fair autoencoder. *International Conference on Learning Representations (ICLR)*. https://doi.org/10.48550/arXiv.1511.00830

[35] Madras, D., Creager, E., Pitassi, T., & Zemel, R. (2018, October). Learning adversarially fair and transferable representations. In J. Dy & A. Krause (Eds.), *Proceedings of the 35th international conference on machine learning* (pp. 3384–3393, Vol. 80). PMLR. https://proceedings.mlr.press/v80/madras18a.html

[36] Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2022). A Survey on Bias and Fairness in Machine Learning. *ACM Computing Surveys*, *54*(6). https://doi.org/10.1145/3457607

[37] Mitchell, S., Potash, E., Barocas, S., D'Amour, A., & Lum, K. (2021). Algorithmic Fairness: Choices, Assumptions, and Definitions. *Annual Review of Statistics and Its Application*, *8*, 141–163. https://doi.org/10.1146/annurev-statistics-042720-125902

[38] Neumark, D., & Johnson, R. (1997). Age discrimination, job separations, and employment status of older workers: Evidence from self-reports. *Journal of Human Resources*, *32*, 779–811. https://doi.org/10.2307/146428

[39] Niño-Adan, I., Portillo, E., Landa-Torres, I., & Manjarres, D. (2021). Normalization influence on ann-based models performance: A new proposal for features' contribution analysis. *IEEE Access*, *9*, 125462–125477. https://doi.org/10.1109/ACCESS.2021.3110647

[40] OECD. (2025). *Education at a glance 2025: Oecd indicators*. https://www.oecd.org/en/publications/education-at-a-glance-2025_1c0d9c79-en/full-report.html

[41] Oh, C., Won, H., So, J., Kim, T., Kim, Y., Choi, H., & Song, K. (2022). Learning Fair Representation via Distributional Contrastive Disentanglement. *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 1295–1305. https://doi.org/10.1145/3534678.3539232

[42] Ojala, M., & Garriga, G. C. (2010). Permutation tests for studying classifier performance. *J. Mach. Learn. Res.*, *11*, 1833–1863. https://doi.org/10.1109/ICDM.2009.108

[43] Pagano, T. P., Loureiro, R. B., Lisboa, F. V. N., Peixoto, R. M., Guimarães, G. A. S., Cruz, G. O. R., Araujo, M. M., Santos, L. L., Cruz, M. A. S., Oliveira, E. L. S., Winkler, I., & Nascimento, E. G. S. (2023). Bias and Unfairness in Machine Learning Models: A Systematic Review on Datasets, Tools, Fairness Metrics, and Identification and Mitigation Methods. *Big Data and Cognitive Computing*, *7*(1). https://doi.org/10.3390/bdcc7010015

[44] Pessach, D., & Shmueli, E. (2022). A review on fairness in machine learning. *ACM Comput. Surv.*, *55*(3). https://doi.org/10.1145/3494672

[45] Phelps, E. (1972). The statistical theory of racism and sexism. *American Economic Review*, *62*, 659–61. https://api.semanticscholar.org/CorpusID:38504493

[46] Prince, A. E., & Schwarcz, D. (2019). Proxy discrimination in the age of artificial intelligence and big data. *Iowa Law Review*, *105*(3), 1257–1318. https://www.scopus.com/inward/record.uri?eid=2-s2.0-85090516500&partnerID=40&md5=844baa0b7365eb9c7c863df7eefaaead

[47] Qutub, A., Al-Hssan, M., Aljohani, R., & Alghamdi, H. S. (2021). Prediction of Employee Attrition Using Machine Learning and Ensemble Methods. *International Journal of Machine Learning and Computing*, *11*(2), 110–114. https://doi.org/10.18178/ijmlc.2021.11.2.1022

[48] Raza, A., Munir, K., Almutairi, M., Younas, F., & Fareed, M. M. S. (2022). Predicting Employee Attrition Using Machine Learning Approaches. *Applied Sciences*, *12*(13), 6424. https://doi.org/10.3390/app12136424

[49] Regulation (EU) 2024/1689 of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) (2024). https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=OJ:L_202401689

[50] Society for Human Resource Management (SHRM). (2025, October). *Shrm releases 2025 benchmarking reports: How does your organization compare?* [Accessed: 2026-04-02]. https://www.shrm.org/about/press-room/shrm-releases-2025-benchmarking-reports--how-does-your-organizat

[51] Suresh, H., & Guttag, J. V. (2021). A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle. *Equity and Access in Algorithms, Mechanisms, and Optimization*, 1–9. https://doi.org/10.1145/3465416.3483305

[52] van Giffen, B., Herhausen, D., & Fahse, T. (2022). Overcoming the pitfalls and perils of algorithms: A classification of machine learning biases and mitigation methods. *Journal of Business Research*, *144*, 93–106. https://doi.org/10.1016/j.jbusres.2022.01.076

[53] Verma, S., & Rubin, J. (2018). Fairness definitions explained. *Proceedings of the International Workshop on Software Fairness*, 1–7. https://doi.org/10.1145/3194770.3194776

[54] Wachter, S., Mittelstadt, B., & Russell, C. (2021). Bias Preservation in Machine Learning: The Legality of Fairness Metrics Under EU Non-Discrimination Law. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3792772

[55] Zafar, M. B., Valera, I., Rogriguez, M. G., & Gummadi, K. P. (2017). Fairness Constraints: Mechanisms for Fair Classification. *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, 962–970. https://proceedings.mlr.press/v54/zafar17a.html

[56] Zemel, R., Wu, Y., Swersky, K., Pitassi, T., & Dwork, C. (2013). Learning Fair Representations. *Proceedings of the 30th International Conference on Machine Learning*, 325–333. https://proceedings.mlr.press/v28/zemel13.html

[57] Zhang, B. H., Lemoine, B., & Mitchell, M. (2018). Mitigating unwanted biases with adversarial learning. *AIES '18: Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 335–340. https://doi.org/10.1145/3278721.3278779

[58] Zhao, Y., Hryniewicki, M. K., Cheng, F., Fu, B., & Zhu, X. (2019). Employee Turnover Prediction with Machine Learning: A Reliable Approach. In K. Arai, S. Kapoor, & R. Bhatia (Eds.), *Intelligent Systems and Applications* (pp. 737–758, Vol. 869). Springer International Publishing. https://doi.org/10.1007/978-3-030-01057-7_56

Chapter A

Appendix

A.1 Additional EDA

Figure 6 exhibits the distribution of the target variable in the test data.

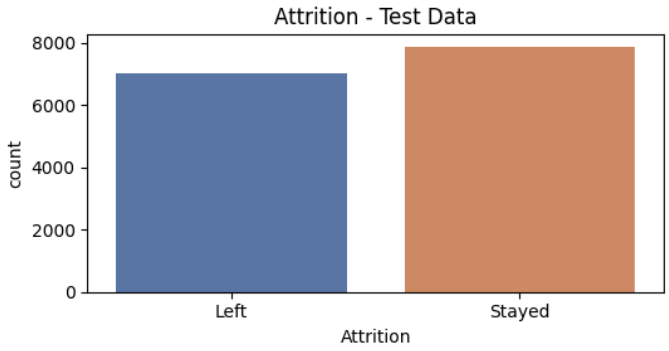


Figure 6: Attrition distribution for test data.

Figure 7 shows the attrition rate for intersectional subgroups in the test data.

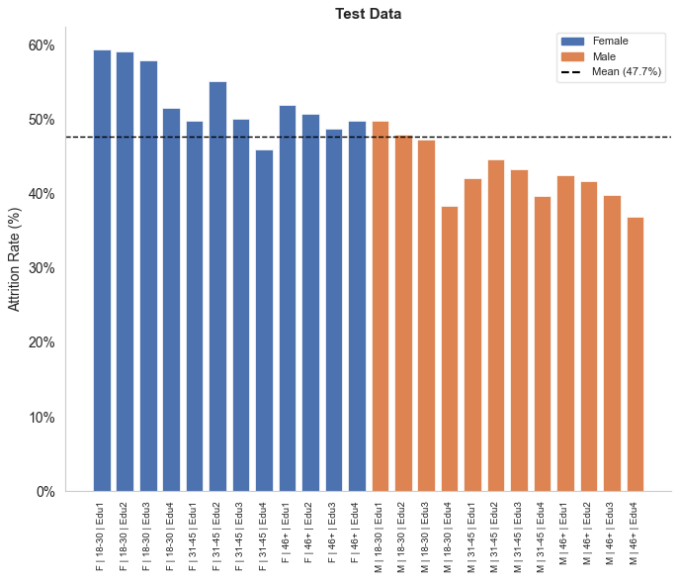
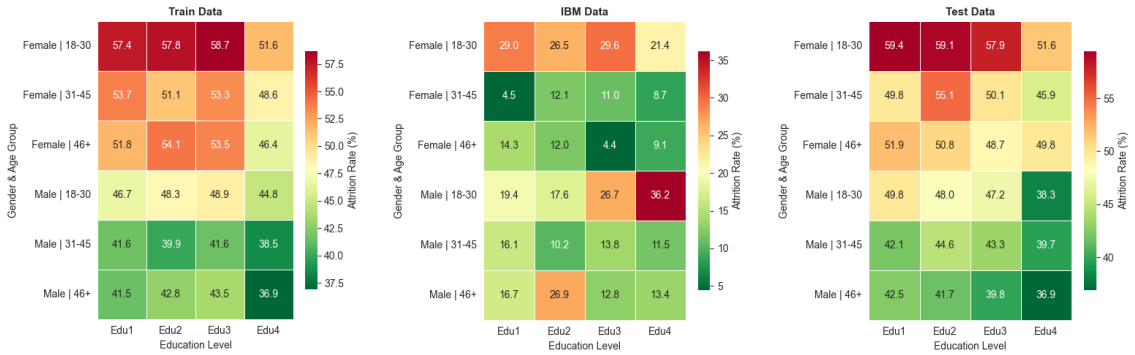


Figure 7: Intersectional Subgroups Attrition Rate for test Data.

Figure 8 shows the attrition rates for each subgroup in a heatmap for all datasets. The

y-axis denotes gender and age group, where the education level is shown on the x-axis, where Edu1 means education level 1.



(a) Training dataset (b) IBM dataset (c) Test dataset

Figure 8: Intersectional subgroups attrition rate (heatmap).

Tables 22 and 23 show the job characteristics for all the intersectional groups in the test data and the IBM data , respectively. Table 23 exhibits that the job characteristics change per subgroup. Moreover, n is significantly smaller than for the training data and test data.

Gender	Age	Edu.	n	Avg Inc. ($)	Avg Job Lvl	OT (%)	Avg Env. Sat.
F	18–30	1	451	7,200	1.79	29.7	2.56
F	18–30	2	563	7,342	1.82	33.0	2.58
F	18–30	3	686	7,320	1.79	31.9	2.61
F	18–30	4	576	7,461	1.79	31.8	2.64
F	31–45	1	416	7,096	1.79	37.3	2.63
F	31–45	2	572	7,302	1.83	33.7	2.64
F	31–45	3	693	7,272	1.85	30.9	2.54
F	31–45	4	601	7,285	1.82	31.6	2.59
F	46+	1	441	7,318	1.83	34.7	2.63
F	46+	2	583	7,279	1.76	30.7	2.54
F	46+	3	671	7,333	1.81	30.4	2.57
F	46+	4	560	7,083	1.78	32.9	2.59
M	18–30	1	524	7,223	1.76	35.3	2.66
M	18–30	2	669	7,315	1.77	31.2	2.62
M	18–30	3	813	7,264	1.77	29.5	2.59
M	18–30	4	684	7,249	1.81	33.6	2.66
M	31–45	1	570	7,324	1.82	37.5	2.56
M	31–45	2	695	7,279	1.78	35.7	2.59
M	31–45	3	879	7,379	1.87	30.9	2.63
M	31–45	4	663	7,182	1.81	33.3	2.60
M	46+	1	530	7,435	1.86	34.2	2.63
M	46+	2	652	7,256	1.78	35.6	2.61
M	46+	3	763	7,423	1.84	31.1	2.60
M	46+	4	645	7,160	1.84	35.3	2.59

Note: Edu. denotes education level (1 = high school, 2 = associate degree, 3 = bachelor's degree, 4 = master's/PhD). n is the number of observations. Avg Inc. is mean monthly income. Avg Job lvl is mean job level. OT is the percentage of overtime. Avg Env. Sat. is the average environment satisfaction.

Table 22: Intersectional group job characteristics for test data.

Gender	Age	Edu.	n	Avg Inc. ($)	Avg Job Lvl	OT (%)	Avg Env. Sat.
F	18–30	1	31	4,006	1.52	45.2	2.81
F	18–30	2	34	4,544	1.68	50.0	2.62
F	18–30	3	81	3,940	1.38	23.5	2.74
F	18–30	4	28	5,634	1.89	25.0	2.64
F	31–45	1	22	6,366	1.73	27.3	2.73
F	31–45	2	58	5,730	1.84	29.3	2.83
F	31–45	3	109	7,097	2.13	27.5	2.71
F	31–45	4	104	6,109	1.92	32.7	2.53
F	46+	1	7	11,808	2.71	14.3	2.00
F	46+	2	25	9,397	2.32	36.0	2.88
F	46+	3	45	11,285	2.56	42.2	3.00
F	46+	4	44	10,669	2.45	15.9	2.84
M	18–30	1	67	3,812	1.34	23.9	2.81
M	18–30	2	51	4,054	1.49	19.6	2.73
M	18–30	3	116	3,875	1.40	24.1	2.75
M	18–30	4	47	4,397	1.49	31.9	2.40
M	31–45	1	31	7,501	2.03	45.2	2.74
M	31–45	2	88	6,820	1.99	27.3	2.70
M	31–45	3	174	6,550	1.95	21.8	2.78
M	31–45	4	156	5,961	1.88	28.2	2.71
M	46+	1	12	10,339	2.33	25.0	2.75
M	46+	2	26	8,741	2.31	34.6	2.46
M	46+	3	47	11,452	2.60	25.5	2.83
M	46+	4	67	10,710	2.45	34.3	2.73

Note: Edu. denotes education level (1 = high school, 2 = associate degree, 3 = bachelor's degree, 4 = master's/PhD). n is the number of observations. Avg Inc. is mean monthly income. Avg Job lvl is mean job level. OT is the percentage of overtime. Avg Env. Sat. is the average environment satisfaction.

Table 23: Intersectional group job characteristics for IBM data.

Table 24 shows the differences in the distribution of the features monthly income and distance from home for the different datasets after removing outliers from the training data.

	Train Set		IBM Set		Test Set	
Variable	**Mean**	**Std**	**Mean**	**Std**	**Mean**	**Std**
Monthly Income	7,297	2,143	6,503	4,708	7,287	2,157
Distance From Home	50	28	9	8	50	29

Table 24: Monthly Income and Distance From Home Statistics Per Dataset.

A.2 Linearity of $\mathcal{L}(Q, \boldsymbol{\lambda})$ in Q and $\boldsymbol{\lambda}$.

Below it is shown that $\mathcal{L}(Q, \boldsymbol{\lambda})$ is linear in Q and $\boldsymbol{\lambda}$.

$$\begin{aligned}
\mathcal{L}(Q, \boldsymbol{\lambda}) &= \hat{Loss}(Q) + \boldsymbol{\lambda}^T(\mathbf{M}\hat{\boldsymbol{\mu}}(Q) - \hat{\mathbf{c}}) \\
&= \sum_{f \in \mathcal{F}} Q(f)\hat{Loss}(f) + \boldsymbol{\lambda}^T(\mathbf{M}\sum_{f \in \mathcal{F}} \hat{\boldsymbol{\mu}}(f)Q(f) - \hat{\boldsymbol{c}}) \\
&= \sum_{f \in \mathcal{F}} Q(f)(\hat{Loss}(f) + \boldsymbol{\lambda}^T\mathbf{M}\hat{\boldsymbol{\mu}}(f)) - \boldsymbol{\lambda}^T\hat{\boldsymbol{c}}.
\end{aligned}$$

This is clearly linear in Q and $\boldsymbol{\lambda}$. Moreover, the minimum of $\mathcal{L}(Q, \boldsymbol{\lambda})$ is chosen over the convex set Δ. Therefore, the minimum is attained at one of the extreme points in the set.

A.3 SPSF Using Soft Counts

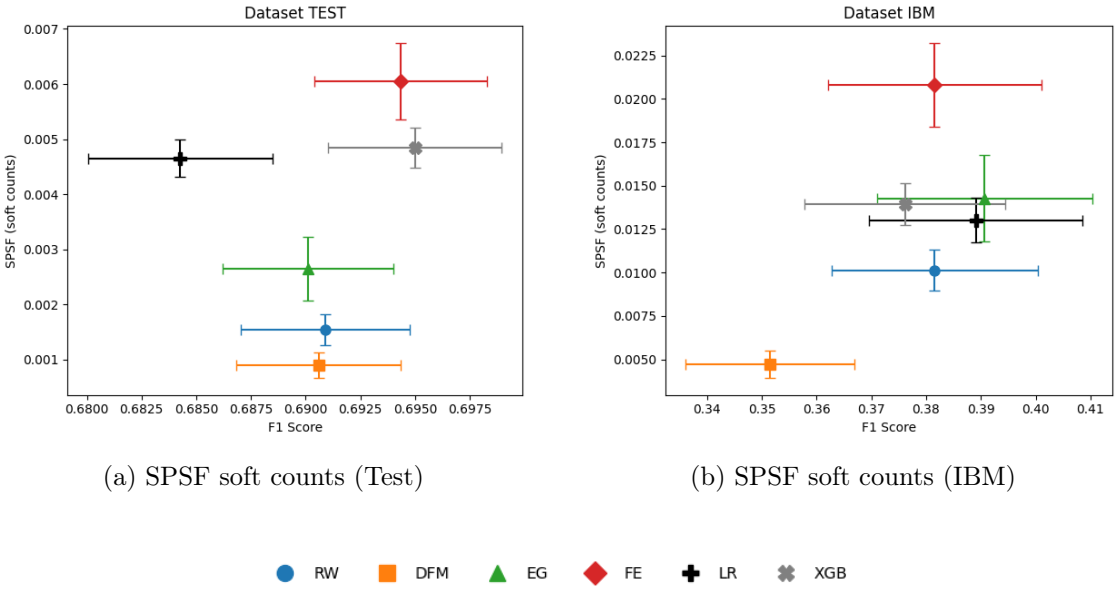
Table 25 shows the performance of the methods with respect to SPSF soft counts.

Method	SPSF (soft counts)
Primary Dataset	
LR	**0.005** [0.004, 0.005]
XGB	**0.005** [0.004, 0.006]
RW	**0.002** [0.001, 0.002]
DFM	0.001 [0.001, 0.002]
EG	**0.003** [0.002, 0.004]
EG Gender	**0.005** [0.004, 0.006]
EG Age	**0.006** [0.006, 0.008]
EG Education	**0.007** [0.007, 0.009]
FE	**0.006** [0.005, 0.008]
IBM Dataset	
LR	**0.013** [0.010, 0.015]
XGB	**0.014** [0.012, 0.016]
RW	**0.010** [0.008, 0.013]
DFM	**0.005** [0.004, 0.007]
EG	**0.014** [0.010, 0.020]
EG Gender	**0.018** [0.014, 0.023]
EG Age	**0.022** [0.016, 0.027]
EG Education	**0.022** [0.017, 0.028]
FE	**0.021** [0.015, 0.026]

Table 25: SPSF (soft counts) by method on the primary and IBM datasets. Bold values indicate an observed disparity that is unlikely under the permutation null ($\alpha = 0.05$), and the brackets denote the 95% confidence interval.

A.4 Additional Fairness-Performance Trade-Off Plots

Figure 9 shows the trade-off plot where SPSF is measured using soft counts.



(a) SPSF soft counts (Test)

(b) SPSF soft counts (IBM)

Figure 9: Fairness-performance trade-off results on the test and IBM datasets with SPSF soft counts.