ERASMUS UNIVERSITY ROTTERDAM

Erasmus School of Economics

Bachelor Thesis BSc2 Econometrics / Economics

Evaluation of Fair Boolean Rule Set Building for Binary Classification

July 4, 2021

Author
Tom Teurlings
482163

Supervisor
M.H. Akyuz
Second assessor
U. Karaca

Abstract

In this report, we evaluate the performance of the column generation (CG) framework in building disjunctive normal form rule sets for fair binary classification. Furthermore, we investigate whether the fair column generation framework offers a solution for building models in circumstances where false negatives have high (societal) costs. Experiments on six data sets are executed and the performance is measured upon the following metrics: accuracy, fairness, and predictive value. Thereby, the CG model is tested in imperfect circumstances, such as small data sets, and is benchmarked with the accuracy-tuned model with and without the sensitive variable. We find that models built by the CG framework subject to fairness constraints achieve superior fairness with comparable accuracy. The framework should be considered as a global discrimination reducer since it substantially reduces overall discrimination. It treats all discrimination similarly which causes reverse discrimination.

Contents

1 Introduction

Machine learning, an area of artificial intelligence, is one of the most disruptive technological innovations in the last century. Implications in labor-intensive tasks save us a lot of time and can be used for endlessly many possibilities. These highly efficient data-driven algorithms are often used for tasks such as classification, regression, and reinforcement learning in fields like advertising, education, recruitment, credit, and many others. Its applications range from feed optimisation on social media (Twitter) to navigation of self-driving cars (Tesla) and deep voicing (Baidu). Another application worth mentioning is IBM's Watson, which can accurately recommend treatments against cancer and makes shopping suggestions in the retail branch. All these newly invented tools are useful, though the consequences they have on our lives are unforeseen.

There is scepticism about the role automated decision-making should play in our lives as these algorithms are not perfect. The effects on society should always be considered, as people trust these algorithms and their results to a large extent. For instance, a Reuter investigation (Dastin, 2018) found that Amazon's AI curriculum selection tool exhibited a large gender bias. While the tool was built to simply select and hire the best candidates, it led Amazon to consider almost only male candidates. When considering fairness of such algorithms, it is crucial to understand there are numerous ways to define and measure fairness. We desire methods that do not discriminate and therefore do not favour certain groups. It is of special importance for applications in fields like health and criminal justice, given that decisions in these fields have life-changing impact. For patients who are hoping to get experimental treatment for their illness, crime victims, or people desperate for a second mortgage, it is vitally important that these procedures are conducted fairly. In this report, we will look at the cases where false-negatives have a large impact. As the urge for fair decision-making models is obvious, this research carries out a variety of experiments to investigate methods assuring fairness in binary classification.

The focus of this research lies on evaluating the approach proposed in Lawless and Günlük (2021) to attain fairness in machine learning algorithms. We will measure the improvement in fairness the *equality of opportunity constraints* have on the models build by the column generation framework by two evaluation methods; namely the maximum disparity between false-negative rates (Lawless and Günlük, 2021) and the detection

of explainable and illegal discrimination (Mehrabi et al., 2019). The main goal of this paper is to answer the following question: *'How does the column generation framework with equality of opportunity constraints perform in building binary decision rule sets based upon fairness, accuracy, and predictive value?'*. Consequently, the following sub-questions are considered:

- *How does the column generation framework perform based on the level of overall and illegal discrimination?*
- *How does the column generation framework perform in imperfect circumstances?*
- *To what extent do the findings hold for data on other domains?*

Based on the measures of fairness and discrimination introduced by Lawless and Günlük (2021) and Mehrabi et al. (2019), this paper tries to answer these questions. Where the former measure gives an indication of overall fairness. And the latter examines the explainable part of the discrimination. For example in the Amazon case, part of the discrimination can be explained by the fact that most technical people are male. Nonetheless, this should be handled by observing someone's technical skills and not by gender. On basis of these measures, we contribute to a more complete overview of the trade-off between fairness and accuracy for machine learning in binary classification in the following ways:

- A more detailed overview of the performance is given by looking into the explainable and unexplainable parts of the discrimination. We will quantify the effect which the added equality of opportunity constraints have on these measures and analyse possible underlying reasons for this. Thereby, it will be checked whether the proposed building method causes reverse discrimination. The fairness improvement of the column generation (CG) is evaluated with a naive approach - excluding the protected variable - as a benchmark.
- By carefully explaining how the constraints affect the classifier's fairness, we get insights into the implementation details of the proposed approach. We attempt to explain how the methods work with different kinds of data. More specifically, the effect of including the fairness constraints on the accuracy-fairness trade-off in samples with few observations and already fair data is tested. We determine

whether data is already fair or not based upon our fairness metric, the false-negative rate.

- Data sets on other domains are included to examine to what extent the results apply for classification on educational, marketing, and medical data. This is of great importance since there are a lot of real-life applications in these fields.

The rest of the paper is structured as follows. Section 2 discusses current solutions to remove unfairness, their limitations and motivate why the CG framework is chosen. Followed by an explanation, in Sect. 3, of the methods used in this research. Section 4 addresses the data and its added value, followed by the most important analyses of some experiments. Thereafter, Sect. 5 will give an in-depth discussion of the results and the fit of the solution for our intended goal. Section 6 wraps up with a concise summary of our main findings and provides some implications for future research.

2 Related Literature

In this section, we discuss relevant literature to our work. First, an overview of the research on fairness in machine learning is given and approaches to tackle these are categorized. Thereafter, we discuss solutions to remove unfairness, their limitations and motivate why the CG framework is chosen.

2.1 Fairness in the Field of Machine Learning

There has been a lot of research on fairness in machine learning (ML). The Human-computer interaction (HCI) community has researched the subject from political (Binns, 2018), social (Hamidi et al., 2018), and psychological (Woodruff et al., 2018) perspectives. Thereby, substantial effort has gone into determining what measures can be used as a stepping stone for mitigating unethical biases in such algorithms. While the field is maturing and algorithmic methods are becoming available, there is little real-world application yet (Holstein et al., 2019). Evidently, the hurdles are that fairness is context and application dependent, and that industries are not able to efficiently implement these solutions due to technical and organisational barriers.

Further, there are a lot of studies focusing on monitoring the fairness of ML algorithms. From Kamiran et al. (2010), it is known that there is an essential condition

for the use of fair ML techniques: the huge amount of unbiased data that is needed to train the algorithms. If this condition is not met, we should correct for it. Supervised learning methods are unbiased in the sense that they are objective with regard to data. If an algorithm is provided with biased data, it incorporates these biases which could result in biased decision-making. Therefore, it should always be questioned whether an automated decision-making process performs reasonably for everyone. Otherwise, resources and opportunities might be restricted for certain subgroups.

There are numerous ways to overcome unfairness, which can be categorized into three groups; pre-, in-, and post-processing (Bruha, 1999). Defined, respectively, as removing the underlying bias in the data, adjusting the optimization such that it is fair, and correcting the predictions of our model. Pre- and post-processing techniques try to correct for the bias and do not attempt to tackle the core problem. Existing data transformation approaches (pre-processing) try to find methods that transform data such that it becomes unbiased, leading to a not effective removal of unfairness (Agarwal et al., 2018). Adjusting the predictions of the classifier (post-processing) for the sensitive attribute is not always possible and often not the optimal solution. Therefore, our preference goes out to guaranteeing fairness by in-processing adjustments. This can be done by either imposing fairness constraints (Lawless and Günlük, 2021) or including fairness measures in the objective function. Fairness measure should be considered as a representation of the distance from the ideal fair situation. It is important not to confuse these measures with actual fairness. It follows that this paper tries to achieve a certain level of fairness (by imposing constraints) and does not attempt to optimise for it - as that would not mean optimising fairness, but a measure of it.

2.2 Limitations of Solutions

In Corbett-Davies and Goel (2018), three approaches of fairness are amplified for removing the unfairness in algorithms. The first approach is anti-classification which refers to excluding protected variables such as gender, race, marital status, or belief. Pedreshi et al. (2008) and Schwarcz (2020) show that this naive approach for removing the unfair differentiation does not work effectively as correlated explanatory variables, proxies, are used by the machine learning algorithms to discriminate - this is called red-lining. Red-lining poses a big problem for enforcing non-discrimination and data protection

laws. For instance, the proxies 'weight' and 'height' can often give away an individual's gender. Discrimination caused by these proxies is nonetheless unethical and unfair. Anti-classification is used as a benchmark, as we think a good method should at least outperform this simplistic solution. For the remainder of this paper, it is referred to as the naive approach.

The second approach is the requirement of equal predictive performance measures imposed by Dwork et al. (2012). There are a variety of these so-called classification parity requirements. For fair classification, Zafar et al. (2019) describe that a model should have no disparity treatment (Agarwal et al., 2018), impact, or mistreatment. In their paper, classification parity requirements are divided into these three groups. From these, we focus on disparity mistreatment, which means that misclassification rates should be equal. Conversely, the other two do not depend on the true label. Moreover, we use the equality of opportunity requirement (Hardt et al., 2016) as statistical measure for fairness, corresponding to the parity in false-negative rates among groups. Alternative disparity mistreatment requirements include measures such as equalized odds (Hardt et al., 2016), the balance of the negative class (Kleinberg et al., 2016), error-rate balance (Chouldechova, 2017), or overall accuracy equality (Berk et al., 2018). As the goal of this paper is to find a way to effectively guarantee equality for those cases where false-negatives have high costs, our preference goes out to equality of opportunity proposed by Hardt et al. (2016) as statistical definition for fairness.

Last, we could balance predictions such that treatment and opportunities are equal conditional on risk estimates - this condition is referred to as calibration. Unfortunately, Pleiss et al. (2017) show that this approach is only compatible with one fairness constraint and does not outperform randomizing a percentage of predictions for an existing classifier on a variety of data sets.

Where anti-classification and classification parity cause reverse discrimination for the groups they are designed to protect; calibration gives little guarantee of fair decision-making. This paper focuses on analysing the effect classification parity has on the fairness and accuracy of an algorithm as we consider it the most promising approach. We will consider the column generation framework for building rule sets as it is shown by Lawless and Günlük (2021) to be a good way to construct fair and interpretable models. Note that we only evaluate this approach with the equality of opportunity constraint, but other measures of fairness could be included as well.

3 Methodology

In this section, the rule set builder used for binary classification is explained. In a similar way as Lawless and Günlük (2021), Boolean decision rules in Disjunctive Normal Form (DNF) are considered to solve the linear relaxation of the Integer Problem (IP) for binary classification. When solving the IP with a column generation (CG) framework the equality of opportunity constraints will be included in order to correct for unfair differentiation. The process of building fair and interpretable rule sets can be decomposed into the following parts: Binarizing data; Integer program; Column generation; Evaluation methods.

3.1 Binarizing Data

First of all, the data is transformed into a binary-valued format. Which makes it simple to make DNF-rule sets, because classification based on a subset of features comes down to a specific combination of 0s and 1s. For categorical data, such as a person's occupation, one-hot encoding is applied. For numerical data, we make ten quantiles by determining threshold values. Note that this is not restrictive, but creates a lot of features. All these features could be selected to determine the optimal rule set. Because the features are binary-valued we have $(2^p - 2)$ feasible rule sets, where p denotes the number of features. Evaluating all of these possibilities would be very inefficient, hence CG techniques are used, similar to Dash et al. (2018).

3.2 The Integer Problem

Before introducing the IP let us introduce notations. Let $\mathcal{K}$ denote the set of all possible DNF rules and c_k the corresponding complexity of a rule. $\mathcal{K}_i \subset \mathcal{K}$ is the set of rules met by observation $i \in I$, where I is the set of all observations. The complexity is determined by the number of conditions in the rule plus one. Further, let us have two subsets that give the true value of the dependent variable, namely $\mathcal{P} = \{i \in I : y_i = 1\}$, and $\mathcal{Z} = \{i \in I : y_i = 0\}$. Data points are labeled by group $g_i \in G$ based on their characteristic feature. The formulation for the IP subject to fairness constraint is as follows:

$$z_{\text{mip}} = \min \sum_{i \in \mathcal{P}} \zeta_i + \sum_{i \in \mathcal{Z}} \sum_{k \in \mathcal{K}i} w_k \tag{1}$$

$$\text{s.t.} \quad \zeta_i + \sum_{k \in \mathcal{K}i} w_k \geq 1 \qquad i \in \mathcal{P} \tag{2}$$

$$C\zeta_i + \sum_{k \in \mathcal{K}i} 2w_k \leq C \qquad i \in \mathcal{P} \tag{3}$$

$$\sum_{k \in \mathcal{K}} c_k w_k \leq C \tag{4}$$

$$w \in \{0,1\}^{|\mathcal{K}|}, \zeta \in \{0,1\}^{|\mathcal{P}|} \tag{5}$$

$$\frac{1}{|\mathcal{P}_1|} \sum_{i \in \mathcal{P}_1} \zeta_i - \frac{1}{|\mathcal{P}_2|} \sum_{i \in \mathcal{P}_2} \zeta_i \leq \epsilon \tag{6}$$

$$\frac{1}{|\mathcal{P}_2|} \sum_{i \in \mathcal{P}_2} \zeta_i - \frac{1}{|\mathcal{P}_1|} \sum_{i \in \mathcal{P}_1} \zeta_i \leq \epsilon \tag{7}$$

Where (1) corresponds to the objective function, also known as the Hamming loss function (Dash et al., 2018). This function is used as a proxy for classification error. Let w_k and ζ_i denote whether rule k is selected in the rule set and whether observation i is misclassified, respectively. Thus the objective of the IP is to minimize the sum of the misclassified data points plus the number of rules selected. The second term will penalize complex rule sets and be in favour of rule sets that are easy to grasp. (2) and (3) force the correct specification of falsely predicted points. If no rule is satisfied, e.x. $w_k = 0 \ \forall \ k$, ζ_i must be 1. When any rule is satisfied ζ_i must be 0. (4) ensures the interpretability of the model by restricting the maximum complexity. (6) and (7) denote the fairness constraints of equality of opportunity as in Hardt et al. (2016).

When ϵ is set to 0, strict equality of opportunity is imposed. Models build under these restrictions are referred to as Strict Fair CG. Although, Strict Fair CG does not necessarily correspond to a completely fair model; $\epsilon = 0$ means that the average false-negative rate is equal among groups. Deviating from strict fairness does not necessarily mean obtaining an inferior model since this measure is merely a representation of fairness. In our analysis, tests for a variety of tolerance levels are provided. The level of tolerance that best improves fairness at the cost of a reasonable amount of accuracy loss and predictive value is chosen. Models build under the restriction $\epsilon = 1$ are referred to as accuracy-tuned models, as models automatically met these restrictions.

3.3 The Column Generation Framework

Due to the large number of feasible rule sets, an efficiently way to solve the problem is needed. The linear programming (LP) of the IP, displayed on the previous page, is solved using the CG framework. The linear relaxation makes the IP computationally less hard, by relaxing integrality constraint (5). The framework proposed in Dash et al. (2018) starts off with a small subset of all possible rules. This initial LP is solved and a rule set $S = \{k \in \mathcal{K} : w_k = 1\}$ is found. Next, a search for the rule with the most negative reduced cost is performed. The search for such a missing variable $j \in J$, the set of binary features, is called the *pricing problem* and can be done by solving the below IP. The *pricing problem* is considered until no other variables with negative reduced cost can be found.

The objective function (8) corresponds to the reduced costs. Where the first is the number of misclassified observations, the second term is the improvement in false-negative rate, and the third term the increased cost of the additional rule. Let $(\alpha_i, \mu_i, \lambda_i)$ denote the optimal solution for the initial restricted LP, where these variables are associated with constraints (2), (3), and (4), respectively. Further, z_j denotes whether the rule concerns feature j. The term $(1 + \sum_{j \in J} z_j)$ can thus be interpreted as the complexity of a rule. Constraints (9), (10) and (11) have similar purposes as (2), (3), and (4), namely enforcing correct specification and restricting the rule complexity. All rules in set I^- (rules with a negative coefficient in the objective function) are considered until no new rule is found. The maximal complexity of a rule D can be independently set of the complexity C in the master problem. We will consider the case where $D = C - 1$.

$$z_{cg} = \min \quad \sum_{i \in \mathcal{Z}} \delta_i + \sum_{i \in \mathcal{P}} (2\alpha_i - \mu_i)\, \delta_i + \lambda \left(1 + \sum_{j \in J} z_j\right) \tag{8}$$

$$\text{s.t.} \quad \delta_i + \sum_{j \in S_i} z_j \leq D \qquad i \in I^- \tag{9}$$

$$\delta_i + \sum_{j \in S_i} z_j \geq 1 \qquad i \in I^+ \tag{10}$$

$$\sum_{j \in J} z_j \leq D \tag{11}$$

$$z \in \{0,1\}^{|J|}, \delta \in \{0,1\}^{|\mathcal{P}|} \tag{12}$$

3.4 Evaluation Methods

After solving the master problem with the CG framework, the results should be evaluated accordingly. However, this is not as simple as it sounds. There is no straight way to quantify fairness. As mentioned, we concentrate on fairness defined as no mistreatment among groups. Which basically means equal false-negative rates between groups - as the focus is on undesirable outcomes in situations with high (societal) stakes. This section discusses the two evaluation methods used to evaluate the performance of the models built by the Fair CG framework.

3.4.1 Maximum Disparity

Lawless and Günlük (2021) noted that strict equality of the false-negative rate between groups is not a realistic assumption, as it is too restrictive. Hence, Lawless and Günlük (2021) took a more practical approach in evaluating the proposed framework and used the maximum disparity between groups in false-negative rate, displayed in (13). Where $\Delta(d)$, the maximum disparity, is bounded by ϵ in constraints (6) and (7). Note that the maximum disparity regards the whole effect as direct discrimination.

$$\Delta(d) = \max_{g,g' \in \mathcal{G}} \left| \mathbb{P}(d(X) = 0 \mid Y = 1, G = g) - \mathbb{P}\left(d(X) = 0 \mid Y = 1, G = g'\right) \right|. \tag{13}$$

3.4.2 Quantifying Explainable and Illegal Discrimination

Part of the discrimination might be explained by valid underlying reasons. We quantify the explainable and illegal part of the discrimination in a similar way as Mehrabi et al. (2019). The explainable effect of every binary variable of the Fair CG and accuracy-tuned model will be calculated. With the difference in characteristic group proportions and equality within variable groups, we are able to quantify the explainable part. These insights are used to examine to what extent the CG framework with fairness constraints mitigates illegal discrimination. It is desired to see that Fair CG will reduce the amount of illegal discrimination and will not create reverse discrimination.

To illustrate the quantification of the explainable and illegal part of the discrimination, here follows an example: We consider the COMPAS data set, which is a set commonly used for binary classification. Our task is to classify persons as risky or not. It has been found that this assessment is biased, where blacks are more often classified as risky. The explanatory variable *45 years or older* seems useful for determining whether

a person is risky or not. According to Ulmer and Steffensmeier (2014), younger people are more prone to committing crimes. If the proportion of younger persons is higher among blacks in the dataset this can explain part of the discrimination.

Table 1: Illegal discrimination quantification for accuracy tuned model on COMPAS data

	Age >= 45		Age < 45		Agregated	
	B	W	B	W	B	W
Observations	1475	2707	628	468	2103	1661
Actually positive (non-risky)	652	1469	170	192	822	1661
Predicted negative (risky)	322	442	129	105	451	547
False-negative rate	49.4%	30.1%	75.9%	54.7%	54.9%	32.9%
Combined false-negative rate	39.7%		65.3%		-	
Expected predicted negative	259.1	583.7	111.0	125.3	-	-

Table 1 shows the proportions for the binary variable *45 years or older* for the accuracy-tuned model, this model is not bound by fairness constraints and is therefore regard as start situation. The level of discrimination is defined as the difference between the false-negative rates between blacks (B) and whites (W), we find a 21.93% difference in the false negative rates. Further, the part of discrimination that can be explained by equating the false-negative rate between characteristic groups based on the explanatory variable group therein. In Table 1, the combined false-negative rates for the explanatory variable are calculated. Thereafter, expected number of negative predictions is given, and we are able to calculated the explainable discrimination. For the accuracy-tuned model the explainable discrimination is 2.33% and the illegal discrimination is 19.60%.

Table 2: Illegal discrimination quantification for Fair CG model on COMPAS data

	Age >= 45		Age < 45		Agregated	
	B	W	B	W	B	W
Observations	1475	2707	628	468	2103	1661
Actually positive (non-risky)	652	1469	170	192	822	1661
Predicted negative (risky)	285	668	130	125	415	793
False-negative rate	43.7%	45.5%	76.5%	65.1%	50.5%	47.7%
Combined false-negative rate	44.6%		70.8%		-	
Expected predicted negative	290.7	655.1	120.3	135.9	-	-

Now we consider a model build with the fairness constraints, the Fair CG. Table 2 shows that the overall, explainable, and illegal discrimination for the fairness tuned model are 2.74%, 2.39%, and 0.35% respectively. This shows that most of the discrimination has been removed. More specifically, almost all the discrimination removed is illegal discrimination. These computations will be done for all explanatory variables that

occur in the final models. Again, the naive approach (removing the sensitive variable) is used for comparison.

4 Experiments

As mentioned, the imposed equality of opportunity constraints improves fairness by enforcing equal false-negative rates among groups. This improvement is important but should not come at the expense of the performance of the model. Therefore, the performance of the model is measured upon the following standards: The accuracy-fairness trade-off, the effectiveness of reducing the false-negative rate gap among groups, and the predictive value. First, we explain what data is used for our experiments and why. Second, the results will be discussed per category. Finally, we will go into the illegal and overall discrimination reduction of the Fair CG.

4.1 Data

In this academic report, we aim to provide a more complete evaluation of how the method used in Lawless and Günlük (2021) performs. In fields like hiring and credit lending, we have seen that equality of opportunity requirements are successfully applied. We are curious if this is also the case for other areas such as marketing, health, and education. So six data sets, that are often used in machine learning, in different domains are included in our analysis. The Adult, Default, ILDP, Student Performance, and Bank Marketing data set can be found in the UCI machine learning data set repository Dua and Graff (2017). Further, the COMPAS and Student Performance data are obtained from Kaggle.

In Table 3 an overview of the data sets and their classification task is given. It can be seen that the sets differ in sample size. The ILDP and Student Performance sets will be considered as a separate category as we are wondering how the models perform on small sets. The other data sets can be separated into those that already fair (Default and Bank Marketing) and those who are not (Adult and COMPAS). Equality of opportunity focuses on the false-negative rate. That is why, it is much more relevant for applications where the consequences of false negatives outweigh those of false positives. Therefore, this is more interesting for the application on the COMPAS, Default, ILDP, and Student Performance data sets than on the Adult and Bank Marketing data sets. We

are interested in the potential difference between these groups. Note that the ILDP and Student Performance data sets are of fairly small sample size for machine learning. We would like to see how the rule set builder performs with not only the fairness constraints, but also by a limited amount of data to learn from. The Bank Marketing classification has *Marital status* as protected variable. We hypothesize that the data and protected variable used, do not significantly alter the results of the Fair CG.

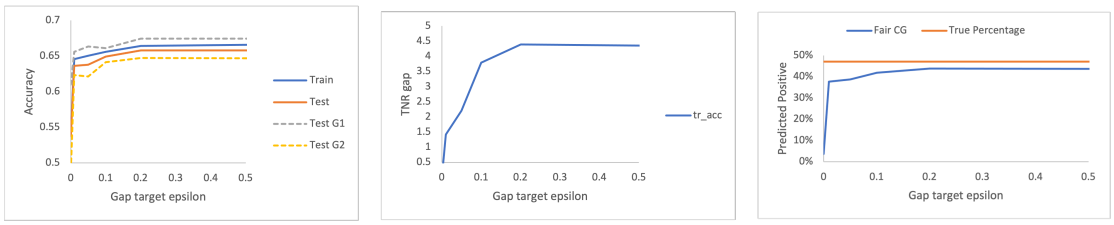
Table 3: Data description

Dataset	# observations	# features	Discrimination factor	Area	Goal classification
Adult	32562	14	Gender	Economic	Income greater than 50.000
COMPAS	5279	7	Race	Crime/justice	Risk assessment of felons
Default	30000	23	Gender	Finance	Default credit risk
ILDP	584	10	Gender	Health	Diagnoses of liver diseases
Bank Marketing	45211	17	Marital status	Marketing	Willingness for term deposit
Student Performance	480	16	Gender	Education	High grade (>7)

4.2 Maximum Disparity

As mentioned, the result can be categorized on its features and will be discussed accordingly. First, the COMPAS and Adult data sets are large unfair data samples. Second, the Default and Bank Marketing represent the fair data samples. Last, the ILDP and Student Performance data set are considered as small and unfair.

4.2.1 COMPAS and Adult



(a) **(b)** **(c)**

Figure 1: The 10-fold mean accuracy (a), total negative rate (b), and proportion of true predictions (c) of the fairness tuned model are plotted against target gap ϵ for the COMPAS data set

One can find the three measures for a variety of ϵ-values for the COMPAS data set in Figure 1. For Strict Fair CG, the model does not perform well. It has a mean accuracy of 53.8% and only predicts a *negative* label, meaning a person is considered non-risky, in 3% of the cases. The proportion of data points actually considered risky is 47.0% in the COMPAS data, which is displayed by the orange line in Figure 1c. Therefore, a

default prediction of *negative* would have an accuracy of 53.0%. As the number of false-negatives is minimized, it is logical that the percentage of predicted non-risky persons is much lower than the actual portion in the data. In practice, our model should at least predict a *positive* label in a substantial portion of the cases. For instance, when making a model for distinguishing whether a person must be considered a potential danger to society, as in the COMPAS data. After all, there is little use for a model if it does not split the risky from non-risky persons. Therefore, it is needed to evaluate a variety of larger values for ϵ.

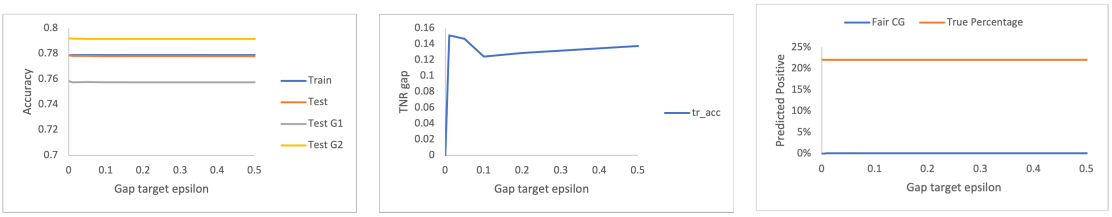
As the fairness constraints are relaxed both the accuracy of the model and the number of *positive* label predictions increases. It can be seen, in Figures 1a and 1b, that after slightly increasing ϵ, the model performs much better and is almost similar in mean accuracy as the accuracy-tuned model. As the constraints are relaxed more, the actual false-negative rate increases rapidly. The challenge is finding a value for ϵ for which all three performance measures are acceptable. A more gradual increase in fairness was expected, so that there would be flexibility in picking an accuracy-fairness level. The value of ϵ must always be chosen based on the accuracy, the fairness, the proportion of *positive* labels, and the classification task - and must be done with care. We found that the models for the Adult and COMPAS data set performed best for the values 0.05 and 0.01, respectively, of ϵ. The results for the Adult data set are in line with these results, hence these results and more implementation details are included in Appendix A. In addition, we have included an analysis for the Adult data set with maximum complexity 100 for comparison with our reference paper Lawless and Günlük (2021).

Table 4, reports the accuracies, fairness, and their associated standard deviation for four different models. It can be seen that a substantial improvement in the mean fairness only has a small penalty on the mean accuracy for the Fair CG. It can be seen that the Fair CG models outperform the naive approach (removing the sensitive variable) and Strict Fair CG on fairness.

Table 4: Mean accuracy and fairness for 10-fold cross-validation

	Adult		COMPAS	
	Accuracy	Fairness	Accuracy	Fairness
Accuracy-tuned model	81.1 (1.1)	4.6 (0.9)	65.7 (2.6)	19.8 (3.4)
Naive approach	80.1 (1.2)	3.9 (1.3)	66.0 (2.3)	19.7 (3.3)
Fair CG	78.6 (0.8)	1.5 (1.7)	63.6 (2.4)	4.4 (2.1)
Strict Fair CG	76.0 (0.4)	0.0 (0.1).	53.8 (3.0)	0.8 (2.6)

4.2.2 Default and Bank Marketing



(a) **(b)** **(c)**

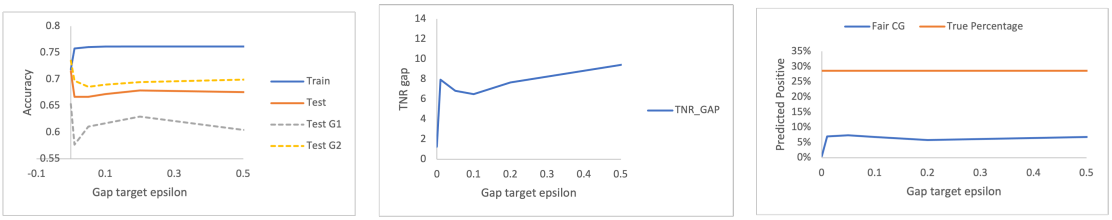
Figure 2: The 10-fold mean accuracy (a), total negative rate (b), and proportion of *positive* predictions (c) of the fairness tuned model are plotted against target gap ϵ for the Default data set

In Figure 2a, it can be seen that increasing ϵ does not alter the performance of the model as the total negative rate is close to zero for all values. This indicates that the equality of opportunity constraints, (6) and (7), are easily met and do not restrict the optimisation. In Table 5, one can observe that the default set contains only a very limited amount of unfairness. With the exception of the case of $\epsilon = 0$, results look the same for the Bank Marketing data set. For that case, a very low mean accuracy is found due to the strict fairness constraint. For all other values of ϵ, the classification models for the Default and Bank Marketing data set are not outperforming a default *negative* prediction. This is probably caused by the skewness in the data; the proportion of *positive* labeled data points is 22.2% and 11.7% for the Default and Bank Marketing data sets respectively. An unsurprising but important thing to note is that the imposed constraints do not worsen the results if already satisfied; so imposing them never hurts the performance of a model. As there is little unfairness in the data, additional investigation would not give new insights. Therefore, we leave it at this.

Table 5: Mean accuracy and fairness for 10-fold cross-validation

	Default		Bank Marketing	
	Accuracy	Fairness	Accuracy	Fairness
Accuracy-tuned model	77.8 (0.6)	0.1 (0.1)	89.3 (0.5)	0.5 (0.3)
Naive approach	77.9 (0.7)	0.1 (0.0)	89.3 (0.5)	0.5 (0.3)
Fair CG	77.9 (0.6)	0 (0)	89.4 (0.5)	0.5 (0.3)
Strict Fair CG	77.8 (0.6)	0.1 (0.1)	58.0 (39.6)	0.1 (0.1)

4.2.3 ILDP and Student Performance



(a) **(b)** **(c)**

Figure 3: The 10-fold mean accuracy (a), total negative rate (b), and proportion of *positive* predictions (c) of the fairness tuned model are plotted against target gap ϵ for the ILDP data set

The difference in the mean train and test accuracy in Figure 3a indicates that the model is overfitted, which leads to poor performance. When looking at the start of the course in 3b, a steep increase in the total negative rate is observed. A clear explanation for this is not found. The steep increase is followed by a much more slow increase compared to the larger data sets. This shows that the fairness constraints work, but in a different way than in the larger data sets. Further, it can be seen that for both data sets only around 10% of the predictions are *positive*, while in reality around 30% is labeled *positive*. While this seems low, it is not worse than the accuracy-tuned model. In practice, you can imagine that even such a small selection might be enough to be a cost-efficient solution for automated decision-making in large operations.

The results for the ILDP and Student Performance data sets in Table 6 are surprising as an improvement in accuracy is found by imposing the fairness constraints, which is counter-intuitive. It must be noted that the magnitude of the standard deviation is very large, which can be explained by the fact that the data sets only consist of 584 and 480 observations. In hope to find more accurate results, a 5-fold evaluation is performed. This analysis still had large standard deviations and did not give better results, in the sense that we could draw better conclusions from it. It is included in the Appendix.

Similar to the COMPAS and Adult data sets, the Strict Fair CG model performs approximately the same as a default *negative* prediction on its accuracy as seen in Table 6. Moreover, the 10-fold cross-validation results give a good indication that the Fair CG works properly. Unfortunately these should be taken with a grain of salt, because it is hard to draw conclusions with such high mean standard deviations.

Table 6: Mean accuracy and fairness for 10-fold cross-validation

	ILDP		Student Performance	
	Accuracy	Fairness	Accuracy	Fairness
Accuracy-tuned model	70.0 (7.0)	8.8 (7.0)	72.5 (5.8)	6.6 (6.4)
Naive approach	72.2 (8.9)	5.5 (5.0)	74.0 (5.3)	3.3 (3.4)
Fair CG	71.2 (8.0)	3.3 (2.4)	74.2 (6.5)	3.0 (2.7)
Strict Fair CG	71.3 (8.8)	0 (0)	71.0 (5.3)	0.8 (2.6)

4.3 Illegal and Reverse Discrimination

Next, we investigate to what extent Fair CG removes illegal discrimination for the COMPAS and Adult data. The maximum disparity experiments show that the fairness constraints reduce the overall level of discrimination. However, some of the discrimination might be justifiable. For example, in the Adult data set men on average earn more than women. This can be partly explained by the fact that men on average work more hours per week in the data set. A fair model should differentiate based on the number of hours worked and not on gender. Completely equating pay for gender would cause reverse discrimination. In this section, we will investigate how the proposed fairness constraints handle these cases.

In Figure 4, the level of overall and illegal discrimination are displayed for the model tuned for accuracy, Fair CG, and the naive approach. The explanatory variables are ordered based on the correlation with the sensitive variable and only the highest correlated binary variable of a feature is shown. In this case, the *score factor high* is the variable most correlated with the sensitive variable with a correlation of 0.341. As Kamiran et al. (2013) describe, these variables are called proxies and cause red-lining. It can be seen that the naive approach does not remove the discrimination effectively and also does not succeed in reducing the illegal portion of discrimination in the model. Conversely, the Fair CG seems to substantially reduce the amount of discrimination in the model from 21.93% to 2.74%. Furthermore, it looks like the line of illegal discrimination is

vertically shifted to the new level of overall discrimination. As the proposed Fair CG treats all explanatory variables as equally discriminated, it can be considered as a global discrimination reduction. With the consequence of causing reverse correlation for the variables correlated with the sensitive variable, because of this the blacks with a high score factor will actually be positively discriminated due to the fairness tuned model.

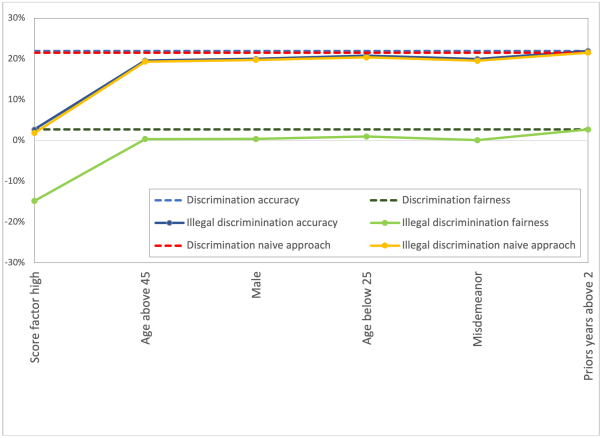


Figure 4: Overall and illegal discrimination for multiple models for the COMPAS data set

Table 7: Correlation between sensitive variable and explanatory variables in the COMPAS data set

Explanatory variable	Correlation
Score factor high	0.341
Age above 45	0.148
Male	0.137
Age below 25	0.129
Misdemeanor	0.108
Priors years above 2	0.086

The analysis for the Adult data set gives similar results. It can be seen that the Fair CG effectively reduces the amount of overall discrimination from 11.87% to 0.26%. The naive approach offers a small improvement and achieves a level of 10.63%. In line with previous results, it is seen that the Fair CG is really good at removing overall discrimination. Compared to the results for the COMPAS data set, Fair CG seems to create reverse discrimination to a lesser extent. Other than just shifting the line vertically, it looks like the model does handle some of the explainable discrimination. Given that it only treats a small portion of discrimination as illegal and causes reverse discrimination, we regard Fair CG more as a global than a local discrimination reducer. Further, it is surprising to see that the binary variable *Black* is correlated with the sensitive variable *Gender* in Table 8; as we know from biology that these variables are not correlated.

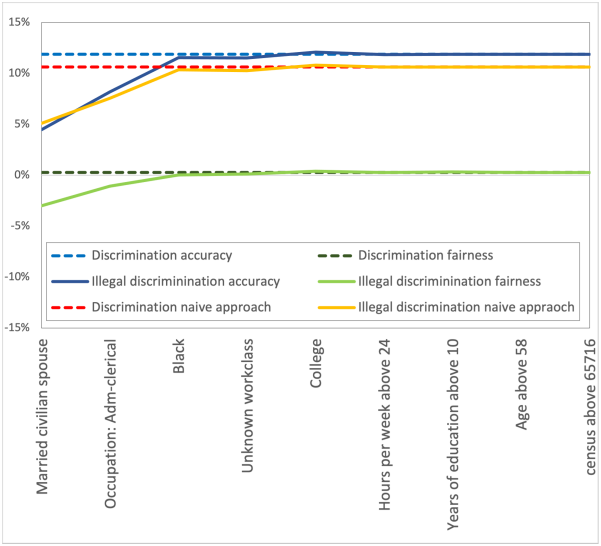


Figure 5: Overall and illegal discrimination for multiple models for the Adult data set

Table 8: Correlation between sensitive variable and explanatory variables in the Adult data set

Explanatory variable	Correlation
Married civilian spouse	0.447
Occupation: Adm-clerical	0.243
Black	0.132
Unknown workclass	0.100
College	0.069
Hours per week above 24	0.046
Years of education above 10	0.038
Age above 58	0.037
census above 65716	0.034

5 Discussion

All in all, the results from the previously stated experiments suggest that the Fair CG is a good solution to solve unfairness in automated decision-making. It can be seen that, for the right parameters, the models substantially improve on fairness with comparable accuracy and predictive value. It was expected to see a more gradual increase in unfairness with the target gap ϵ, such that there would be the flexibility to select the desired accuracy-fairness level. Instead, we found that the increase was abrupt, hence we emphasize the importance of picking a good level of target gap ϵ. Moreover, our experiments indicate that the implication of the equality of opportunity constraints never worsens the models' performance. Most importantly, we have seen the DNF sets built by the column generation framework with fairness constraints achieve better fairness and a comparative accuracy compared to the benchmark, the naive approach.

Furthermore, we have seen that the column generation framework with fairness constraints effectively removes overall discrimination, but causes reverse discrimination for proxies of the protected variable. So it does not solve conditional discrimination. Therefore, we suggest (for proper usage) to first remove these dependencies, before applying the column generation framework for building fair DNF rule sets. The Fair CG framework should be regarded as a global discrimination reducer. We consider local dis-

crimination reducers a promising area for future research. While computationally more intensive, these solutions might be able to deal with illegal discrimination better.

In our experiments, we found some instructions for the usage of the fair binary decision-rule model builder. The model must be trained on enough balanced data. We have seen that the model performs fine for a skewness up to 30%, more skewness causes the model to tend to a default *negative* prediction as in all similar machine learning cases. Further, we have seen that for smaller data sets the models have some mixed results. While the constraints seem to be working, they do so to a lesser extent. A phenomenon not strange to observe in machine learning, where the model is only as good as its data. Moreover, we have seen that the model causes reverse discrimination in cases where the explanatory variables are correlated with the protected variable. Hence, we suggest using the Fair CG framework when building a binary classification model on a large balanced data set with uncorrelated attributes.

It is suggested to use these kinds of models for informative and non-crucial decision-making in time-intensive tasks as they do not work perfectly and can not remove discrimination completely. The fair rule set building method is great for purposes where false-negatives are not acceptable to reduce the workload such as health, hiring, education, and justice. It can also be used in analysis or selection procedures in fields like marketing, and finance, however these decisions should not hugely influence people's lives. Another insight mention-worthy is that our experiments do not indicate the fairness improvement of the Fair CG framework to be domain-specific, as no substantial difference among data sets is observed. Because of the improvement in fairness of the automated decision-making models, we suggest the usage in applications such as medical treatments, loans, and justice which have a large impact on people's lives. For companies already depending on these kinds of systems, it is a huge improvement.

We can conclude that the implication of the equality of opportunity constraints in the column generation framework can effectively reduce the inequality of the false-negatives as a global discrimination reducer. Unfortunately, it is not possible to completely remove the unfairness in the model. For the usage of such models, it should also be noted that the proposed adjustment is conclusive in only a small part of the cases. This has the consequence that frequently *positive* observations are labeled *negative*. Because we try to minimize the false-negative predictions, the model is more careful in predicting *positive* values. Therefore, it is only recommended to use this model in cases where the

consequences of false-negatives are far more important than false-positives. It should be stressed that these models are important for detecting and information purposes and not able to make automated decisions yet.

6 Conclusion

The goal of this research was to find out whether the column generation framework with fairness constraints is a good method to build fair binary classification models. Where fairness is interpreted as no disparity mistreatment among groups. The research question of this paper is: *'How does the column generation framework with equality of opportunity constraints perform in building binary decision rule sets based upon fairness, accuracy, and predictive value?'*. The executed experiments show that the Fair CG is able to create superior models based upon its fairness, with comparable accuracy. The Fair CG framework is useful for building models in applications where false-negatives have high (societal) costs when used with the correct level of ϵ. The models are able to substantially reduce the level of overall discrimination. The models treat almost all discrimination as illegal discrimination, while some of the discrimination may be explainable. Therefore, Fair CG can be considered global discrimination reducers, which has the possible side-effect of causing reverse discrimination. As for all machine learning algorithms, we have seen that the building works properly on balanced and large data sets. When one of these criteria is not fulfilled, the performance of the Fair CG seems to deteriorate.

For future research, we would suggest investigating how the column generation framework could be used to minimise illegal discrimination instead of the overall discrimination. Furthermore, we are curious how the equality of opportunity constraints can be used in existing local discrimination reducers. Thereby, it would be interesting to see how the proposed method performs in a non-binary setting. Another reference for future research is to investigate how the column generation framework with the balance of the negative class requirements, as defined in Kleinberg et al. (2016), performs in binary classification tasks where false-positives are of interest such as online advertising and marketing.

References

Alekh Agarwal, Alina Beygelzimer, Miroslav Dudik, John Langford, and Hanna Wallach. A reductions approach to fair classification. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 60–69. PMLR, 10–15 Jul 2018. URL `http://proceedings.mlr.press/v80/agarwal18a.html`.

Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research*, page 0049124118782533, 2018.

Reuben Binns. Fairness in machine learning: Lessons from political philosophy. In *Conference on Fairness, Accountability and Transparency*, pages 149–159. PMLR, 2018.

Ivan Bruha. Pre-and post-processing in machine learning and data mining. In *Advanced Course on Artificial Intelligence*, pages 258–266. Springer, 1999.

Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 5(2):153–163, 2017.

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

Sanjeeb Dash, Oktay Günlük, and Dennis Wei. Boolean decision rules via column generation. *arXiv preprint arXiv:1805.09901*, 2018.

Jeffrey Dastin. Amazon scraps secret ai recruiting tool that showed bias against women, Oct 2018. URL `https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G`.

Dheeru Dua and Casey Graff. Uci machine learning repository. 2017.

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226, 2012.

Foad Hamidi, Morgan Klaus Scheuerman, and Stacy M Branham. Gender recognition or gender reductionism? the social implications of embedded gender recognition systems. In *Proceedings of the 2018 chi conference on human factors in computing systems*, pages 1–13, 2018.

Moritz Hardt, Eric Price, and Nathan Srebro. Equality of opportunity in supervised learning. *arXiv preprint arXiv:1610.02413*, 2016.

Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé III, Miro Dudik, and Hanna Wallach. Improving fairness in machine learning systems: What do industry practitioners need? In *Proceedings of the 2019 CHI conference on human factors in computing systems*, pages 1–16, 2019.

Faisal Kamiran, Toon Calders, and Mykola Pechenizkiy. Discrimination aware decision tree learning. In *2010 IEEE International Conference on Data Mining*, pages 869–874. IEEE, 2010.

Faisal Kamiran, Indrė Žliobaitė, and Toon Calders. Quantifying explainable discrimination and removing illegal discrimination in automated decision making. *Knowledge and information systems*, 35(3):613–644, 2013.

Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807*, 2016.

Connor Lawless and Oktay Günlük. Fair and interpretable decision rules for binary classification. 2021.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *arXiv preprint arXiv:1908.09635*, 2019.

Dino Pedreshi, Salvatore Ruggieri, and Franco Turini. Discrimination-aware data mining. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 560–568, 2008.

Geoff Pleiss, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q Weinberger. On fairness and calibration. *arXiv preprint arXiv:1709.02012*, 2017.

Daniel Schwarcz. Proxy discrimination in the age of artificial intelligence and big data. 2020.

Jeffrey Todd Ulmer and Darrell J Steffensmeier. The age and crime relationship: Social variation, social explanations. In *The nurture versus biosocial debate in criminology: On the origins of criminal behavior and criminality*, pages 377–396. SAGE Publications Inc., 2014.

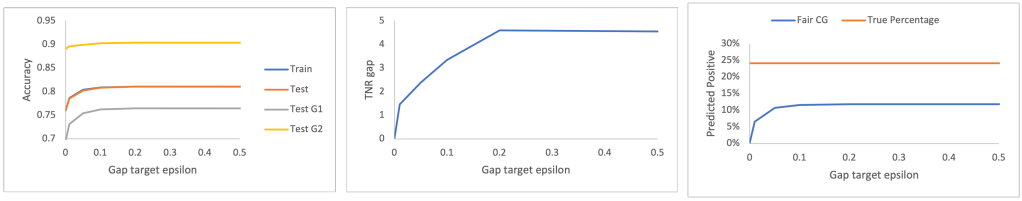
Allison Woodruff, Sarah E Fox, Steven Rousso-Schindler, and Jeffrey Warshaw. A qualitative exploration of perceptions of algorithmic fairness. In *Proceedings of the 2018 CHI conference on human factors in computing systems*, pages 1–14, 2018.

Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez-Rodriguez, and Krishna P Gummadi. Fairness constraints: A flexible approach for fair classification. *J. Mach. Learn. Res.*, 20(75):1–42, 2019.

A Appendix. Additional Results

This section gives additional figures and information which is supplementary and provides a supportive view to the finding of this paper. For all data sets not displayed in the main text, the accuracy, total negative rate, and proportion of true predictions are plotted against the false-negative rate gap target ϵ. Additionally, the most striking features will be highlighted. This section is broken up by data set in arbitrary order.

A.1 Adult



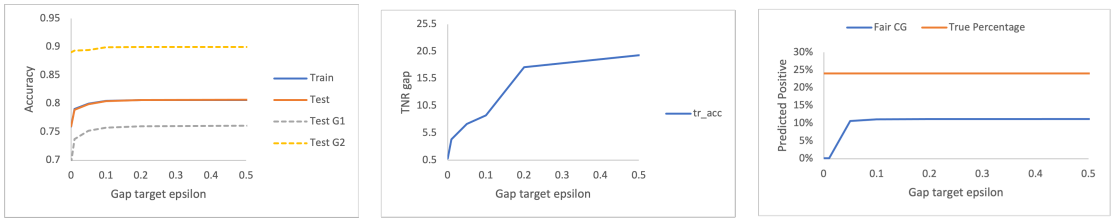
(a) (b) (c)

Figure 7: The 10-fold mean accuracy (a), total negative rate (b), and proportion of true predictions (c) of the fairness tuned model are plotted against target gap ϵ for the Adult data set for complexity 20

In our analysis, we have chosen to work with a maximal complexity of 20, as for all other data set, but also report results for a maximal complexity of 100 for comparison with Lawless and Günlük (2021). Contradicting to the results of the COMPAS data set we observe that with the strict equality of opportunity constraints, $\epsilon = 0$, the model actually manages to predict 20.10% with a *positive* label and reports a fairness of 0.0. When we look more closely at the 10 models, since we use 10-fold cross-validation, it can be seen that some models report an accuracy of around 25%. These do not make much sense and are therefore disregarded. Figure 7(b) shows that for other values of ϵ the model is only good for a minor reduction in fairness.

Figure 8 gives the results for a maximum complexity of 100, these look more promising. Like the fairness-tuned model for the COMPAS data, we see for strict fairness the model is similar to a default false prediction. When relaxing the constraints the model starts to also predict positive cases, the mean accuracy of the model increases along with the total negative rate gap. In line with our previous results, it can be seen that the column generation framework subject to fairness constraints performs reasonably well for small values of targeted ϵ. As the accuracy does not increase after a certain point, we take the value for which the model has a reasonable proportion of true predictions

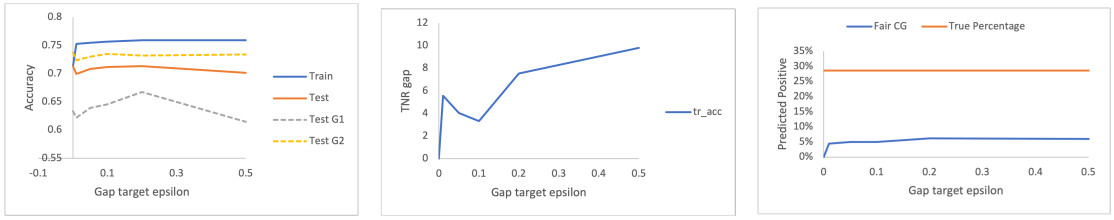
($\epsilon = 0.05$). Additional investigation for values of the target gap between 0.01 and 0.05 might yield an even better solution. Furthermore, we observe that the total negative rate gap gradually increases with the target gap. All in all, the column generation framework performs well in building a model that is superior in its fairness. The numbers in the paper are obtained with a maximum complexity of 100.



(a) **(b)** **(c)**

Figure 8: The 10-fold mean accuracy (a), total negative rate (b), and proportion of true predictions (c) of the fairness tuned model are plotted against target gap ϵ for the Adult data set for complexity 100

A.2 ILDP



(a) **(b)** **(c)**

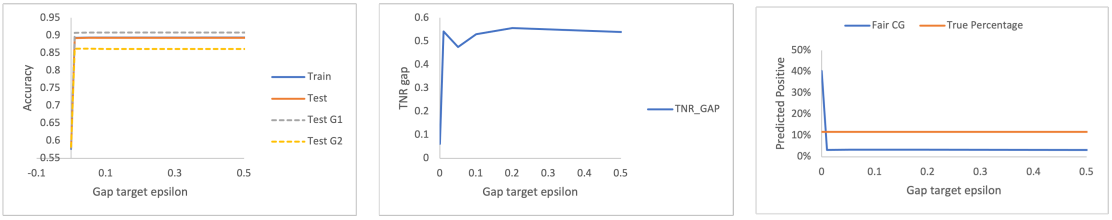
Figure 9: The 5-fold mean accuracy (a), total negative rate (b), and proportion of true predictions (c) of the fairness tuned model are plotted against target gap ϵ for the ILDP data set

The ILDP data set was one of the more useful analyses we did. As the small data set did exhibit unfairness, we were able to see how the fair CG performed in imperfect circumstances. In Figure 9, the results for the 5-fold cross-validation are displayed. We observe that the model is pretty overfitted as the mean accuracy is much higher for the train sample than the test sample. Due to the small sample size, the results should be taken with a grain of salt. The results are in line with those found in the COMPAS and adult models; Table 9 shows that the model has superior fairness with comparable accuracy. However, note that the model is outperformed by the naive approach. We see no clear indication for this.

Table 9: Mean accuracy and fairness for 5-fold cross-validation

	ILDP		Student Performance	
	Accuracy	Fairness	Accuracy	Fairness
Accuracy-tuned model	67.9 (4.0)	13.1 (6.4)	71.5 (5.5)	2.5 (2.6)
Naive approach	68.6 (3.6)	3.7 (3.4)	73.5 (4.2)	3.9 (2.5)
Fair CG	67.2 (3.8)	6.5 (3.7)	72.5 (6.4)	2.8 (2.1)
Strict Fair CG	71.5 (3.7)	1.3 (0.7)	70.2 (5.2)	0.7 (1.5)

A.3 Bank Marketing

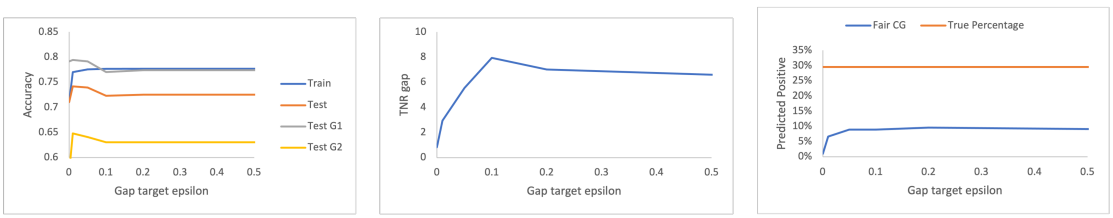


(a) **(b)** **(c)**

Figure 10: The 10-fold mean accuracy (a), total negative rate (b), and proportion of true predictions (c) of the fairness tuned model are plotted against target gap ϵ for the Bank Marketing data set

For the Bank Marketing data set, we see a very poorly performing model for the strict fairness restriction. While not exhibiting a large portion of unfairness, the model is definitely influenced by the equality of opportunity constraint. No underlying reason for this can be found. For further relaxations of the constraints, it can be seen that increasing ϵ does not alter the performance of the model. This indicates that the fairness constrictions are easily met. As stated in Table 5, it can be seen in the Default set contains only a very limited amount of unfairness. As addition investigation would not give new insights we leave it at this.

A.4 Student Performance



(a) **(b)** **(c)**

Figure 11: The 5-fold mean accuracy (a), total negative rate (b), and proportion of true predictions (c) of the fairness tuned model are plotted against target gap ϵ for the Student Performance data set

The results for the Student Performance model are surprising. Similar to other models, strict fairness is too restrictive for the model. However, we see a remarkable thing when relaxing the fairness constraints. At first, we see a steep increase, meaning the model does not perform well based on its fairness. After further increases we see the model improving, something we have not seen in other data sets. We see that for the values 0, 0.01, 0.05, 0.1 for the gap target the complexity of the rule set increases, this might explain the reduction in unfairness. In Table 6, it can be seen that the models. Similar to the 5-fold ILDP analysis, both shown in Table 9, We see some mixed results. For instance, it is very counter-intuitive to see that the restricted models achieve higher accuracy. As these results do not make much sense and the standard deviations are relatively high, we do not put too much weight on these results.

B Appendix. Programming Code

In the provided zip-file, the code for the Bachelor Thesis *'Evaluation of Fair Boolean Rule Set Builder for Binary Classification'* can be found. Thankfully, I could build upon the previous results of Connor Lawless. My starting point was his GitHub page: https://github.com/conlaw. My contributions can be found in the Jupyter Notebook file (ending with the .ipynb extension). Data and results have been excluded to keep this map as small as possible.

If you wish to reproduce my findings and run the experiments found in the paper, you should execute the code in the following order: Split into Cells, Fair CG Rule Generation, Fair CG Trials, Accuracy Experiment, Quantifying (Illegal) Discrimination.

Hopefully, this will be helpful. For questions, you can reach out to me.