Erasmus University Rotterdam

Erasmus School of Economics

Bachelor Thesis Econometrics and Operations Research

Fair and Interpretable Methods for Binary Classification

Author:

Jeske van de Sandt

Supervisor:

dr. M.H. Akyuz

Student ID number:

509415

Second assessor:

Msc. U. Karaca

July 4, 2021

The views stated in this thesis are those of the author and not necessarily those of the supervisor, second assessor, Erasmus School of Economics or Erasmus University Rotterdam.

Abstract

Automated decision making systems are more and more used to support or even replace human decision making. Hence, the used methods should not only be interpretable to the user, but also fair. A method is considerd fair if an individual receives a fair treatment, independent of their membership of a certain category, determined by for example their gender, race or age. In this research, two different approaches to fair and interpretable binary classification are considered: Fair Column Generation and pre-processing training data by performing Local Massaging and Local Preferential Sampling to remove illegal discrimination. CART is applied to the pre-processed data after which the performance of the methods is compared regarding fairness and accuracy, where Equality of Opportunity is used as fairness metric. Both methods obtain accuracy results comparable to results from simple, interpretable classification methods. For the standard machine learning datasets used in this research, FairCG slightly outperforms data pre-processing in combination with CART. The performances differ mainly considering fairness.

Contents

1 Introduction

Classification is an indispensable technique in a wide variety of fields; biology classifies organisms, marketing classifies customers and criminal justice classifies convicts. Binary classification is applicable if the objective is to divide a sample in two groups, based on classification rules. Common applications include labeling an email as spam or no spam and predicting possible conversion of a customer. In some cases, classification models complement human decisions on sensitive topics like loan approval and criminal justice. For these applications, classification methods are required to be interpretable to the people involved. Furthermore, they should be to be fair and free from discrimination. Pedreshi et al. (2008) describe discrimination in decision making as the unfair treatment of individuals based on their membership of a certain category, determined by for instance their gender, race or age. Discrimination is an ongoing issue which now also affects artificial decision making, according to Osoba and Welser IV (2017). Due to the rising awareness of the issue of unfairness in decision making, research has been shifting its focus from creating methods that are strictly optimized to achieve the highest accuracy to also taking into account important social matters like discrimination. In binary classification, as stated by Zafar et al. (2017), biased classification could occur if the value of a sensitive attribute influences decision making and the decision disproportionately hurts or favors people with a certain sensitive attribute value. Removing this attribute from the data is generally not sufficient to obtain fair classification since classifiers are trained on historical data in which prejudice and discrimination might already be present.

Recent studies suggest alternative methods for fair and interpretable binary classification. One of these methods is Fair Column Generation (FairCG), introduced by Lawless and Günlük (2020). FairCG makes use of integer programming to generate a set of decision rules to classify a dataset. Fairness constraints are added to the integer program to ensure the fairness of the classification. The results obtained by Lawless and Günlük (2020) show their method achieves similar results regarding accuracy compared to other simple and interpretable classification methods, while outperforming them regarding fairness. Because of these promising results, part of this paper is dedicated to review the FairCG method and reproducing the results from FairCG where Equality of Opportunity is used as the fairness metric. Similar to Lawless and Günlük (2020), the standard machine learning datasets Adult, Compas and Default are used to test FairCG. As an extension, another dataset is used, namely the German dataset from the UCI Machine Learning Repository.

Lawless and Günlük (2020) compare the performance of FairCG to the performance of Classification and Regression Trees (CART), a commonly used technique for binary classification. The results show that CART outperforms FairCG regarding accuracy, however FairCG performs better for fairness. Therefore in this research an approach is examined to make CART more fair. Kamiran et al. (2010) suggest using Local Massaging or Local Preferential Sampling to remove illegal discrimination from a dataset before a classification tree is trained on the data. In this research Local Massaging and Local Preferential Sampling are performed on the standard machine learning datasets also used for FairCG, after which CART will be applied on the pre-processed data. Finally, FairCG is also applied to the pre-processed data to

determine if FairCG or CART performs better in combination with data pre-processing.
This paper aims to provide a clear review of both the FairCG method and the pre-processing of data in combination with CART, as two different methods of fair and interpretable classification. Thereafter the performance of the methods can be compared and the main research question of this paper is given by: *How does FairCG perform compared to CART, where CART is applied to pre-processed data, regarding accuracy and fairness?'*

The application of both methods to the standard machine learning datasets yields FairCG obtaining better fairness and accuracy results than CART, when FairCG is trained on the original data and CART on pre-processed data. If both methods are trained on pre-processed data, they obtain more similar results. Here, Equality of Opportunity is used as fairness metric. The differences in accuracy and fairness are not remarkable; which method is preferable depends on the concerns of the user and the context in which the method will be applied. Data pre-processing results in the removal of a large part of the illegal discrimination in the datasets and therefore might be favorable if this is the specific intention of the user.
The remainder of this paper is structured in the following manner: in Section 2, the literature on fair and interpretable classification is discussed. Section 3 contains an extensive explanation of the methods and how they could be implemented. Sections 4 contains a brief explanation of the used datasets and the results from the application of the methods to the standard machine learning datasets. Finally, Section 5 gives a short summary and conclusion.

2 Literature

Fairness in Machine Learning has been an elaborately discussed topic over the past years. Some say using algorithmic techniques eliminate human biases in the process of decision making, but an algorithm can only be as good as the data it is trained on. Machine Learning fairness focuses on ensuring that inaccuracies in models and biases in data do not lead to models which treat certain individuals unfavorably, based on for example their sex or race (Oneto and Chiappa (2020)).
According to Chouldechova and Roth (2018) unfairness in machine learning can, among other things, be caused by biased data or by the fact that majority populations have a higher influence on the overall classification error and therefore they might be classified more accurate if the objective is to minimize this error. Barocas and Selbst (2016) explain how biases in the training data affect the classification method in two ways: first, if the method treats the cases in which prejudice has played a role as valid examples, the method will simply reproduce the prejudice. Second, if the method bases its decisions on a biased sample of the population, every decision which rests on these inferences could disadvantage under-represented groups in the dataset systematically.
To be able to determine how fair a classification method actually is, one needs to define fairness. A fairness definition can either be statistical or individual. This paper focuses on statistical fairness, which fixes a small number of groups, based on a sensitive attribute such as gender and race, and aims for parity of some statistical measures across these groups. Some examples of these measures are false negative

rate, positive prediction value and statistical parity, which Dwork et al. (2012) define as the property that the demographics of those receiving a certain classification are identical to the demographics of the entire population. Hardt et al. (2016) suggest using Equality Of Opportunity as a measure of fairness, which requires the false negative rates across the groups to be equal. This means all individuals involved should get a 'fair shot' to obtain a positive classification, and the group to which they belong should not influence this probability.

In addition to fairness, classification methods should be interpretable to be useful in a practical context. According to Carvalho et al. (2019) machine learning systems are becoming increasingly ubiquitous, which causes algorithmically informed decisions to have a large social impact. However, most of these systems remain black boxes, which means their internal logic is hidden to the users and often even experts can not fully understand the rationale behind their prediction. Since verifiability of decisions is often mandatory in practice, the demand for understandable machine learning methods increases. Several methods have been proposed for interpretable classification, such as the Supersparse Linear Integer Model (Ustun et al. (2013), Zeng et al. (2015)) and the Interpretable Classification Rule Mining Algorithm (Cano et al. (2013)).

The literature presents many different approaches to binary classification. Lawless and Günlük (2020) make use of boolean rules in Disjunctive Normal Form. One of the main advantages of using decision rules is that they are easy to interpret. This is because they have a general if-then structure; if a certain number of criteria is met, then an instance gets a positive classification. This structure resembles natural language and the way humans think. Furthermore, experiments done by Lakkaraju et al. (2016) show that decision sets can reach similair classification accuracy as state-of-the-art machine learning techniques. Dembczyński et al. (2010) make use of decision rules not only because of their human-interpretable form, but also because they are an aggragation model able to represent complex relations between attributes. Integer programming can be used to find a set of decision rules which is suited for performing the classification. The objective function of the integer program can be formulated in such a way that accuracy is maximized. Su et al. (2015) propose an objective function which include Hamming Loss and sparsity for the trade-off between accuracy and interpretability. The objective function of Lawless and Günlük (2020) minimizes the Hamming loss, which counts the number of rules that need to be changed in order to specify an instance correctly. In addition, constraints can be added to the integer program to make sure the rule set is appropriate for the specific usage it is designed for. To achieve an interpretable set of decision rules, Lawless and Günlük (2020) added complexity constraints, which control for the maximum allowed complexity of the chosen rules. They also added fairness constraints which enables them to bound the unfairness of the method.

Another method which is often used for binary classification is CART (Breiman et al. (1984), Song and Ying (2015), Pallara (1992)). Due to their tree structure, decision trees are interpretable and simple to understand for the users. In addition, they require little data preparation and can be applied to both categorical and numerical data. The results of the paper from Lawless and Günlük (2020) show that CART outperforms FairCG when only considering accuracy. However, it performs worse considering fairness.

In general, when training a decision tree, fairness metrics are not taken into account. Since fairness is one of the main goals in this research, different methods are considered to make the classifications done by CART more fair. Kamiran et al. (2010) adjust the splitting criteria of the decision tree by adding a non-discrimination approach. Aghaei et al. (2019) propose an mixed-integer optimization framework for learning optimal and fair decision trees. Kamiran et al. (2013) impose data pre-processing to prevent illegal discrimination from being present in the data. Since their methods show the ability of removing a large part of illegal discrimination from a dataset, this method will be implemented in this research. After this is done, CART will be applied to classify the data. The techniques will allow any machine learning classification method to result in fair classification when they are used in combination with the pre-processed data (Dunkelau and Leuschel (2019)).

3 Methodology

This paper reviews FairCG and data pre-processing techniques Local Massaging and Local Preferential Sampling as methods for fair and interpretable binary classification. This section provides and extensive description of both methods, as well as an explanation of how the methods will be implemented to test and compare their performance.

3.1 Fair Column Generation

FairCG uses integer programming to generate a set of decision rules in Disjunctive Normal Form. First, the data is binarized to transform categorical and numerical features into binary features. Thereafter a ruleset will be generated by iteratively solving the Master Program and Pricing Problem. An instance is classified positively if it meets at least one of the decision rules in the rule set. For the description of the FairCG method the notation of Lawless and Günlük (2020) is adopted.

3.1.1 Data preparation

To be able to classify the instances, a ruleset is generated. The decision rules in this ruleset are in DNF and consist of a number of binary features. An instance meets a rule when it has a value of 1 for all the features in the rule. Instances can be represented as a string of 0s and 1s, where they have value 0 if they do not have a feature, and 1 if they do. To achieve this structure, numerical and categorical features in the datasets are changed into binary features. For this purpose the approach of Dash et al. (2018) is used. An elaborate explanation of this approach is given by Günlük (2020). In the following section, his notation is used for demonstration.

To binarize the categorical features, one-hot encoding is used. This means a binary feature and a complement of this binary feature are created for every value a categorical feature can take. For example, categorical feature $w \in \{v_1, ..., v_5\}$ is replaced with 10 binary features, $[x_1, ..., x_5, (1 - x_1), ..., (1 - x_5)]$ where $x_i = 1$ if $w = v_i$ and 0 otherwise.

Numerical features are binarized using a sequence of increasing thresholds. In this research, the numerical

values are divided into 10 categories by using quantiles. For example, numerical feature $w \in \mathcal{R}$ is replaced with 20 binary features, $[x_1, ..., x_{10}, (1-x_1), ...(1-x_{10})]$ where x_i equals 1 if $w \leq t_i$ and $t_1 < t_2 < ... < t_{10}$.

3.1.2 Fairness

The aim of the FairCG method is to generate a set of decision rules $d : \{0,1\}^p \rightarrow \{0,1\}$ which minimizes the expected error $P(d(X) \neq Y)$ for samples (X_i, y_i) with labels $y_i \in \{0,1\}$ and features $X_i \in \{0,1\}^p$. Consider the case where each data point has an associated group which is determined by a sensitive feature $g_i \in \mathcal{G}$. This could be for example race or gender.

A number of different fairness metrics could be used to determine how fair a classification method is. Similar to Lawless and Günlük (2020), this paper considers Equality of Opportunity. This criterion has a large number of real life applications since it puts a large cost on false negative predictions, which is relevant considering for example loan approvals. The Equality of Oppurtunity requires the false negative rate to be equal across groups, so the following equation needs to hold:

$$P(d(X) = 0|Y = 1, G = g) = P(d(X) = 0|Y = 1) \quad \forall g \in \mathcal{G} \tag{1}$$

In practice it is often not realistic to demand for this equality to hold, so the focus is on putting a boundary on the unfairness for a classifier d:

$$\Delta(d) = max_{g,g' \in \mathcal{G}} |P(d(X) = 0|Y = 1, G = g) - P(d(X) = 0|Y = 1, G = g')| \tag{2}$$

When training classifier d, Equation (2) could be implemented in the objective function or added as a constraint in the form $\Delta(d) \leq \epsilon$ to control for the maximum allowed unfairness of the classifier.

3.1.3 Master Program

Because the input data is binary-valued, a DNF rule set consists of rules that check if an observation has a certain combination of 0s and 1s. A large integer problem can be used to find the set of decision rules which minimizes classification error subject to fairness constraints. To solve this integer problem, a LP relaxation can be solved using column generation. Therefore, first the Master Integer Problem is formulated.

Let $\mathcal{K}$ be the set of all possible rules made with the available features and $\mathcal{K}_i$ the subset of these rules which are met by instance $i \in \mathcal{I}$, where $\mathcal{I}$ is the set of all instances in the dataset. c_k is the cost of rule $k \in \mathcal{K}$, which equals the total number of features in the rule plus one. Based on their labels, the instances can be partitioned in $\mathcal{P} = \{i \in I : y_i = 1\}$ (positive classified instances) and $\mathcal{Z} = \{i \in I : y_i = 0\}$ (negative classified instances). Furthermore, the instances belong to a group $g \in \mathcal{G}$, based on the value of the sensitive attribute. For simplicity it is assumed there are two groups, $\mathcal{G}_1$ and $\mathcal{G}_2$. Let $\mathcal{P}_g = \mathcal{P} \cap \mathcal{G}_g$ and $\mathcal{Z}_g = \mathcal{Z} \cap \mathcal{G}_g$ with $g = \{1, 2\}$.

The binary variable w_k equals 1 if rule k is selected and 0 otherwise. Binary variable ζ_i equals 1 if data point i is classified incorrectly, where $i \in \mathcal{P}$. Parameter C represents the maximum total complexity of

all chosen rules and ϵ the maximum allowed unfairness. The formulation of the Master Integer Program becomes:

$$z_{mip} = \min \sum_{i \in \mathcal{P}} \zeta_i + \sum_{i \in \mathcal{Z}} \sum_{k \in \mathcal{K}_i} w_k \tag{3}$$

$$\text{s.t.} \quad \zeta_i + \sum_{k \in \mathcal{K}_i} w_k \geq 1 \quad i \in \mathcal{P} \tag{4}$$

$$C\zeta_i + \sum_{k \in \mathcal{K}_i} 2w_k \leq C \quad i \in \mathcal{P} \tag{5}$$

$$\sum_{k \in \mathcal{K}} c_k w_k \leq C \tag{6}$$

$$w \in \{0,1\}^{|\mathcal{K}|}, \zeta \in \{0,1\}^{|\mathcal{P}|} \tag{7}$$

$$\frac{1}{|\mathcal{P}_1|} \sum_{i \in \mathcal{P}_1} \zeta_i - \frac{1}{|\mathcal{P}_2|} \sum_{i \in \mathcal{P}_2} \zeta_i \leq \epsilon \tag{8}$$

$$\frac{1}{|\mathcal{P}_2|} \sum_{i \in \mathcal{P}_2} \zeta_i - \frac{1}{|\mathcal{P}_1|} \sum_{i \in \mathcal{P}_1} \zeta_i \leq \epsilon \tag{9}$$

The first term of the objective function (3) minimizes the number of false negative classifications, the second term minimizes the number of chosen rules which are met by instances with a negative classification. Constraint (4) ensures ζ_i takes value 1 if instance i does not meet any chosen decision rules but has a true positive classification. On the other hand, constraint (5) ensures that ζ_i only takes value 1 if instance i is met by none of the chosen decision rules. Constraint (6) puts a boundary on the complexity of the chosen rules. Constraints (8) and (9) bound the maximum unfairness, where Equality of Opportunity is used as fairness metric. The LP relaxation of the Master Program can be obtained by dropping constraint (7).

3.1.4 Pricing Problem

To solve the LP relaxation, the Column Generation Framework is used as is done by Dash et al. (2018). Here, the relaxation of the Master Program is first solved with a small subset of all possible rules, $\hat{\mathcal{K}} \subset \mathcal{K}$. Let $(\mu, \alpha, \lambda, \gamma_1, \gamma_2)$ be the dual solution from solving the LP relaxation of the Master Program with possible rule set $\hat{\mathcal{K}}$, where the dual variables are respectively associated with the constraints (4), (5), (6), (8) and (9). The dual solution will be used to formulate the pricing problem. The objective of this problem is to find rules with a negative reduced cost that are currently not in $\hat{\mathcal{K}}$.

A decision rule consists of a subset of all features $\mathcal{J}$ and an instance is met by this rule if the instance has a value 1 for all features contained in the rule. Let binary variable z_j equal 1 if a rule contains feature $j \in \mathcal{J}$ and 0 otherwise. Binary variable δ_i equals 1 if a rule misclassifies instance i. Let $\mathcal{S}_i$ correspond to the zero-valued features of instance i. The full pricing problem is given by:

$$z_{cg} = \min \sum_{i \in \mathcal{Z}} \delta_i + \sum_{i \in \mathcal{P}} (2\alpha_i - \mu_i)\delta_i + \lambda(1 + \sum_{j \in \mathcal{J}} z_j) \tag{10}$$

$$\text{s.t.} \quad D\delta_i + \sum_{j \in \mathcal{S}_i} z_j \leq D \quad i \in I^- \tag{11}$$

$$\delta_i + \sum_{j \in \mathcal{S}_i} z_j \geq 1 \quad i \in I^+ \tag{12}$$

$$\sum_{j \in \mathcal{J}} z_j \leq D \tag{13}$$

$$z \in \{0,1\}^{|\mathcal{J}|}, \delta \in \{0,1\}^{|\mathcal{P}|} \tag{14}$$

The first and second part of the objective function (10) minimize the number of misclassifications when a rule is chosen, the third parts minimizes the increase in complexity when a rule is chosen. The objective function represents the reduced cost for a generated rule. Constraints (11) and (12) make sure that variable δ_i accurately reflects whether or not the rule classifies data point i with a positive label. Constraint (13) puts a maxium bound on the complexity of the rule. The value of D is set equal to C - 1, where C is the maximum complexity from the Master Program.
Solving this problem yields the rule with the most negative reduced cost. This rule is added to $\hat{\mathcal{K}}$. Then the LP relaxation of the Master Program is solved again and the process is repeated, until no rule with a negative reduced cost can be found.

3.2 Data pre-processing to remove discrimination

As an extension to the paper of Lawless and Günlük (2020), another approach to fair classification is considered. This is done by pre-processing the data before the classification method is applied. Kamiran et al. (2013) suggest using Local Massaging or Local Preferential Sampling to remove illegal discrimination from the data. Their notation is adopted to demonstrate the methods.

3.2.1 Discrimination in data

Consider again the case where a dataset can be divided into one or more groups based on a sensitive attribute. Regularly it is determined by law that it is illegal to base decisions on the sensitive attribute. Whether or not it is legal to use an attribute in decision making depends on law and anti-discrimination policies. For the purpose of illustration is assumed that the data contains sensitive attribute gender, which could take the value m (male) or f (female). Here, males form the favored group, which means they have a higher probability of receiving positive classification. In this case, positive classification ($y = 1$) represents acceptance, for example to a college. Discrimination is present in the dataset if the acceptance probability is larger for males than for females and can be defined as:

$$D_{all} = P(y = 1|m) - P(y = 1|f) \tag{15}$$

Not all discrimination should be considered illegal. In fact, illegal discrimination can be defined as:

$$D_{illegal} = D_{all} - D_{expl} \tag{16}$$

where D_{expl} is equal to the explainable discrimination. Part of discrimination might be explainable if there is an explanatory attribute e which is correlated with the sensitive attribute, but also provides

objective information on the label. To remove illegal discrimination, males and females should have an equal probability of being accepted when they have the same value for explanatory attribute e. Therefore, given the value of e, the 'correct' acceptance probability needs to be determined, which is equal for males and females. This 'correct' acceptance probability is defined as the average of the acceptance probabilities of the males and females:

$$P^*(y = 1|e) := \frac{P(y = 1|e, m) + P(y = 1|e, f)}{2} \tag{17}$$

Here is assumed every instance that meets a certain threshold can receive a positive or negative classification and there is no predetermined maximum number of positive or negative classifications. Therefore, another definition of correct probability acceptance needs to be used if for example the number of available spots at a university is limited.

Using the correct acceptance probability, the explainable discrimination for attribute e can be defined as:

$$D_{expl} = \sum_{i=1}^{n} P(e = e_i|m)P^*(y = 1|e) - \sum_{i=1}^{n} P(e = e_i|f)P^*(y = 1|e) \tag{18}$$

where n is the number of different values the explainable attribute e can take. The explainable discrimination equals the difference in probability of acceptance between males and females, if every individual with the same value for attribute e has the same probability of being accepted, independent of their gender.

3.2.2 Local Massaging and Local Preferential Sampling

For fair classification, classifiers should be trained on data that contains as little as possible illegal discrimination. To remove illegal discrimination from a dataset, the relationship between the sensitive and the outcome attributes should be removed. To accomplish this, simply removing the sensitive attribute is not sufficient, since some other attributes could be strongly related with the sensitive attribute. When decisions are then made based on these attributes, they are still likely to be discriminatory. On the other hand, removing all attributes that are correlated with the sensitive attribute will help to remove illegal discrimination. However, this could highly decrease the classification accuracy since these attributes might also contain objective information on the attribute. Therefore this paper considers Local Massaging and Local Preferential Sampling to remove illegal discrimination from the dataset.

The relation between the explanatory attribute e and the sensitive attribute s and label y is are measured as the information gain about the sensitive attribute given e and about the label given e. The information gains are defined as:

$$G(y, e) = H(y) - H(y|e) \quad \text{and} \quad G(s, e) = H(s) - H(s|e)$$

where $H(.)$ denotes entropy.

With Local Massaging, the dataset first gets partitioned based on the possible values the explanatory

variable can take. For both the favored and the unfavored group, the probability of acceptance given this value is computed and these values are used to compute the 'correct' probability of acceptance for both groups. Using this correct probability, the number of instances is calculated for which the label needs to be changed in order to make the acceptance probability equal to the correct acceptance probability. For this number of instances the label in changed to false for the favored group and to true for the unfavored group. The instances for which the labels are changed are the instances closed to the decision boundary. Algorithm 1 contains the algorithm for Local Massaging.

Similar to Local Massaging, Local Preferential Sampling starts with the computation of the correct probability of acceptance based on the value of the explanatory attribute and the number of instances that need to be changed. In this case however, the labels of the instances are not changed, but for the favored group, the instances with a positive classification get deleted and instances with a negative classification get duplicated. For the unfavored group, it is the other way around. Again, the removed and duplicated instances are the instances closed to the decision boundary. The algorithm for Local Preferential Sampling is given by Algorithm 2.

Algorithm 1: Local Massaging

input : dataset $(\mathbf{X}, \mathbf{s}, \mathbf{e}, \mathbf{y})$
output: modified labels $\hat{\mathbf{y}}$

`PARTITION` $(\mathbf{X}, \mathbf{e})$:
find all unique values of $e : \{e_1, e_2, ..., e_k\}$;
for $i = 1$ **to** n **do**
  make a group $X^{(i)} = \{X : e = e_i\}$;
end
for *each partition* $X^{(i)}$ **do**
  learn a ranker $p(y = 1|X^{(i)}, e_i) = \mathcal{H}_i(X^{(i)})^*$;
  rank `males` using $\mathcal{H}_i$ according to $p(y = 1|X^{(i)}, e_i)$;
  `relabel DELTA(male) males` that are closest to the decision boundary from 1 to 0 (Algorithm 3);
  rank `females` using $\mathcal{H}_i$ according to $p(y = 1|X^{(i)}, e_i)$;
  `relabel DELTA(female) females` that are closest to the decision boundary from 0 to 1;
end
$^*\mathcal{H}_i(X^{(i)})$ *is a ranker which ranks all instances for which* $e = e_i$ *using all attibutes* X*, according to their probability of receiving a positive classification*

After Local Massaging and Local Preferential Sampling are performed on the dataset, any classifier can be applied on the processed data. In this paper, a decision tree is trained on the processed data and then tested on original test data. Decision Tree is chosen as classification method since CART shows best performance considering accuracy in the paper of Lawless and Günlük (2020). For the tree, possible maximum depths 2 and 5 are used, since these will limit the number of branches and therefore keep the trees interpretable. Similar to FairCG, 10 fold cross validation is performed to give a mean an standard deviation for the obtained accuracy and fairness.

Algorithm 2: Local Preferential Sampling

input : dataset $(\mathbf{X}, \mathbf{s}, \mathbf{e}, \mathbf{y})$
output: modified labels $\hat{\mathbf{y}}$

`PARTITION` $(\mathbf{X}, \mathbf{e})$:
find all unique values of $e : \{e_1, e_2, ..., e_k\}$;
for $i = 1$ **to** n **do**
  make a group $X^{(i)} = \{X : e = e_i\}$;
end
for *each partition* $X^{(i)}$ **do**
  learn a ranker $p(y = 1|X^{(i)}, e_i) = \mathcal{H}_i(X^{(i)})$*;*
  rank `males` using $\mathcal{H}_i$ according to $p(y = 1|X^{(i)}, e_i)$*;*
  `delete` $\frac{1}{2}$`DELTA(male) males 1` that are the closest to the decision boundary (Algorithm 3)*;*
  `duplicate` $\frac{1}{2}$`DELTA(male) males 0` that are the closest to the decision boundary*;*
  rank `females` using $\mathcal{H}_i$ according to $p(y = 1|X^{(i)}, e_i)$*;*
  `delete` $\frac{1}{2}$`DELTA(female) females 0` that are the closest to the decision boundary*;*
  `duplicate` $\frac{1}{2}$`DELTA(female) females 1` that are the closest to the decision boundary*;*
end

Algorithm 3: subroutine DELTA(gender)

return $G_i * |p(y = 1|e_i, \texttt{gender}) - p^*(y = 1|e_i)|$,
where $p^*(y = 1|e_i)$ comes from Equation 17,
G_i is the number of `gender` people in $X^{(i)}$;

4 Experiments

This section provides the results from the application of FairCG and data pre-processing on the standard machine learning datasets Adult, Compas, Default and German.

4.1 Data

To implement FairCG, Lawless and Günlük (2020) used the Adult and Default datasets coming from the UCI machine learning repository. They also made use of the machine learning cleaned version of the Compas dataset from ProPublica. In this paper the equivalent datasets are used, as well as the German dataset from the UCI machine learning repository.

The Adult dataset contains 32561 observations for which the data contains information about among other things their age, occupation and the number of hours they work per week. The observations can be classified in two groups based on their income, where one group earns more than 50K per year and the other group earns less. The variable sex will be considered as the sensitive attribute.

The Default dataset contains 30000 observations of default payments in Taiwan. The outcome variable indicates whether or not an observation has a default payment. Some features are the amount of given credit, the education and the amount of previous payments. Again the gender variable is considered as the sensitive attribute.

The Compas dataset contains 5278 observations. The outcome variable tells whether or not the defendants commited crime within the two years after the commercial algorithm COMPAS made a prediction about

the defendants likelihood of reoffending. Race is considered as the sensitive attribute, and similar to the study of Lawless and Günlük (2020), only African American and Caucasian respondents are considered. The last dataset, German, contains 1000 observations and classifies people having either good or bad credit risks. The available variables contain for example the present employment, credit history and credit amount. Age is considered as the sensitive attribute, were a distinction is made between adults (older than 25) and non-adults (25 or younger). This threshold is chosen because the analysis of Kamiran and Calders (2009) shows this provides the most discrimination possibilities.

4.2 Fair Column Generation

The Fair Column generation method is used for generating rules in DNF to classify the data points. This section compares the results obtained in this research to the result from Lawless and Günlük (2020).

4.2.1 Implementation

When working with large datasets, solving the Master Program and the pricing problem becomes computationally intensive. Therefore a timelimit of 5 minutes for the overall training and 45 seconds for the pricing problem is implemented. When the optimal rule set is found for the relaxation of the MIP, the 1000 rules with the lowest reduced cost are selected and used to solve the Master Program.
To speed up the solving of the pricing problem, a random sample of the data, consisting of 2000 rows and 100000 non-zeros, is used instead of the entire dataset. Also, a greedy heuristic is used in addition to solving the pricing problem, to generate rules that contain up to five features. The greedy heuristic is used for a few CG cycles after which is switched to solving the pricing problem. If a large number of rules is generated during an iteration, the 100 rules with the lowest reduced costs are returned.
To generate the optimal rule set, a two-step approach is used. The first phase has the purpose of generating a rule set. This phase starts with an empty rule set. For the maximum allowed unfairness ϵ the values 0.2 and 1 are used. For the rule complexity parameters C the following values are used: Adult: $\{60, 80, 110\}$, Compas: $\{5, 15, 30\}$ and Default: $\{5, 15, 30\}$.
The rules generated in the first phase are used as the intitial rule set for solving the Master Program in the second phase. Here the following values for C are used: Adult: $\{80, 90, 100\}$, Compas: $\{10, 15, 20\}$ and Default: $\{10, 15, 20\}$. For the maximum allowed unfairness, the values 0.0, 0.01 and 1.0 will be used for all datasets.

4.2.2 Results Adult, Compas and Default datasets

Table 1 contains the results from FairCG for the Adult, Data and Compas dataset. The first row contains the result from optimizing for accuracy, the second row contains the results from optimizing for fairness. The accuracy equals the percentage of correctly classified data points. The fairness equals the difference in false negative rates between the two groups present in the data. Therefore it is prefered to obtain a high value for $BestAcc$ and a small value for $BestFair$. The accuracy and fairness are computed for the test data. In the table the mean accuracy and fairness from the 10 folds are given. The standard deviations are given in the parenthesis.

Table 1: Mean Accuracy and Fairness results from FairCG

	Adult		Compas		Default	
	Accuracy	Fairness	Accuracy	Fairness	Accuracy	Fairness
Best Acc	82.5 (1.0)	1.2 (7.5)	68.0 (2.2)	23.6 (6.9)	82.0 (0.6)	0.9 (4.2)
Best Fair	81.0 (2.2)	0.3 (4.8)	64.8 (1.4)	0.2 (2.4)	81.9 (0.8)	0.1 (4.5)

For the Adult dataset, the best accuracy is obtained when $C = 100$ and $\epsilon = 0.01$ and the best fairness when $C = 100$ and $\epsilon = 0.0$. The obtained results differ quite significantly from the results from Lawless and Günlük (2020). When tuned for best accuracy, this research obtains a slightly smaller accuracy, but also a smaller fairness value and thus a higher fairness. When considering best fairness, this research obtains a higher accuracy and a similar value for fairness, but with a larger standard deviation. It is not obvious what causes these differences. In their paper, Lawless and Günlük (2020) do not state at which complexity and maximum unfairness levels they reach their best accuracy and fairness levels. Possibly, similar results would be obtained with more details from their experiment. Table 5 in the Appendix contains the results for different values of C and ϵ and also contains the results from computing the fairness and accuracy for the training data.

When implementing $\epsilon = 0.0$ for the Compas dataset, not all folds obtained valid results. This means the total complexity of the rules used for the classification was 0 and thus no rules were actually used. All data points simply receive the same, most present classification. For $C = 10$, no folds got valid results, for $C = 15$ and $C = 20$ only some of the folds. The results given in Table 1 require all folds to have valid results and therefore $\epsilon = 0.01$ is used to obtain best fairness. Best accuracy is obtained when $C = 20$ and $\epsilon = 1.0$. Table 6 in the Appendix contains all results for the Compas dataset. A ' - ' indicates no folds had valid results, a ' * ' indicates only some of the folds had valid results.

Similar tot the Compas dataset, for the Default dataset the implementation of $\epsilon = 0.0$ gives no valid results for any fold when $C = 10$ and $C = 15$ and only for some folds for $C = 20$. The best accuracy is obtained for $C = 10$ and $\epsilon = 1.0$ and the best fairness is obtained for $\epsilon = 0.01$ and $C = 20$. Again, it needs to hold that all folds have valid results. All results for the Default dataset are given in Table 7 in the Appendix.

For the Compas and Default dataset, the results are more similar compared to the results obtained by Lawless and Günlük (2020). They may have not required all folds to have valid results which leads to higher fairness. In addition, it is likely Lawless and Günlük (2020) presented the training results in their paper since it does not seem possible to obtain a value of 0.0 for fairness with a 0.0 standard deviation for the test data. Other small differences could for example be caused by different folds, since they are created randomly. In addition, a random sample from the data is used to solve the pricing problem.

4.2.3 Results German dataset

As an extension on the paper of Lawless and Günlük (2020), FairCG is also applied on the German dataset. For the maximum allowed unfairness ϵ the values 0.0, 0.01 and 1.0 are used, similar to the other datasets. For the maximum complexity C, the values 20, 30, 40 were chosen after testing some other

values and checking which values lead to highest accuracy. The results for the German dataset are given in Table 2.

Similar to the Compas and Adult dataset, choosing $\epsilon = 0.0$ did not yield valid results for all 10 folds. In fact, depending on the maximum complexity, only one or two folds had valid results. Unlike the other datasets, for this dataset can be noted that higher allowed complexity and unfairness do not necessarily yield higher accuracy and fairness when considering the test data. This is likely due to the relatively small size of the dataset. For every fold, the test data contains of only 100 observations, which causes the results to be highly depending on the fold.

When considering the results for the training data, it is observable that higher complexity leads to higher accuracy. Relaxing the fairness constraint from $\epsilon = 0.01$ to $\epsilon = 1.0$ does not yield higher accuracy while it does leads to more unfairness. Therefore, choosing $\epsilon = 0.01$ seems most appropriate for this dataset.

Table 2: Results FairCG for German dataset

		$\epsilon = 0.0$		$\epsilon = 0.01$		$\epsilon = 1.0$	
	Complexity	Accuracy	Fairness	Accuracy	Fairness	Accuracy	Fairness
Test	20	74.5 (4.9)*	4.0 (5.6)*	71.1 (5.0)	0.9 (11.1)	71.2 (6.9)	0.4 (10.0)
	30	67.0**	0.9**	70.4 (5.1)	6.9 (15.9)	69.3 (4.2)	5.6 (12.3)
	40	71.0 (1.4)*	4.2 (2.7)*	71.4 (5.4)	3.6 (18.2)	69.4 (5.2)	6.0 (16.5)
Train	20	73.9 (3.4)*	0.0 (0.0)*	77.2 (0.7)	0.1 (0.6)	76.5 (0.9)	1.5 (2.7)
	30	79.2**	0.0**	79.5 (0.7)	0.1 (0.6)	79.1 (1.0)	3.5 (4.2)
	40	78.9 (4.6)*	0.0 (0.0)*	81.1 (0.7)	0.1 (0.5)	81.1 (0.7)	3.5 (5.8)

*Only folds with valid results were used to obtain these values

**No standard deviation available since there is only one fold with valid results

4.3 Data pre-processing

In addition to the FaiCG method, another method for fair and interpretable classification is considered. Here, the data is pre-processed using Local Massaging and Local Preferential Sampling before the classifier is trained on the data.

4.3.1 Local Massaging and Local Preferential Sampling

Discrimination in a dataset consists of explainable and illegal discrimination. The size of the explainable discrimination depends on the choice of explanatory attribute. This attribute is required to be associated with both the sensitive attribute and the label available in the dataset. The relationship between the explanatory attribute and the sensitive attribute and label is measured as the information gain about the sensitive attribute given the explanatory attribute and about the label given the explanatory attribute. The information gain for the different explanatory attributes contained in the data are given in Figure 1. From all datasets, the Adult datasets contains most attributes with a large information gain. Some attributes however, like *relationship* and *marital-status* are logically highly correlated with gender, and therefore it is questionable if these attributes give more objective information about the label than the

gender itself. On the other hand, *hours-per-week*, which contains the number of hours worked per week, does give objective information about someone's salary and therefore could cause some explainable discrimination. For every the dataset, the six explainable attributes with the highest information gain

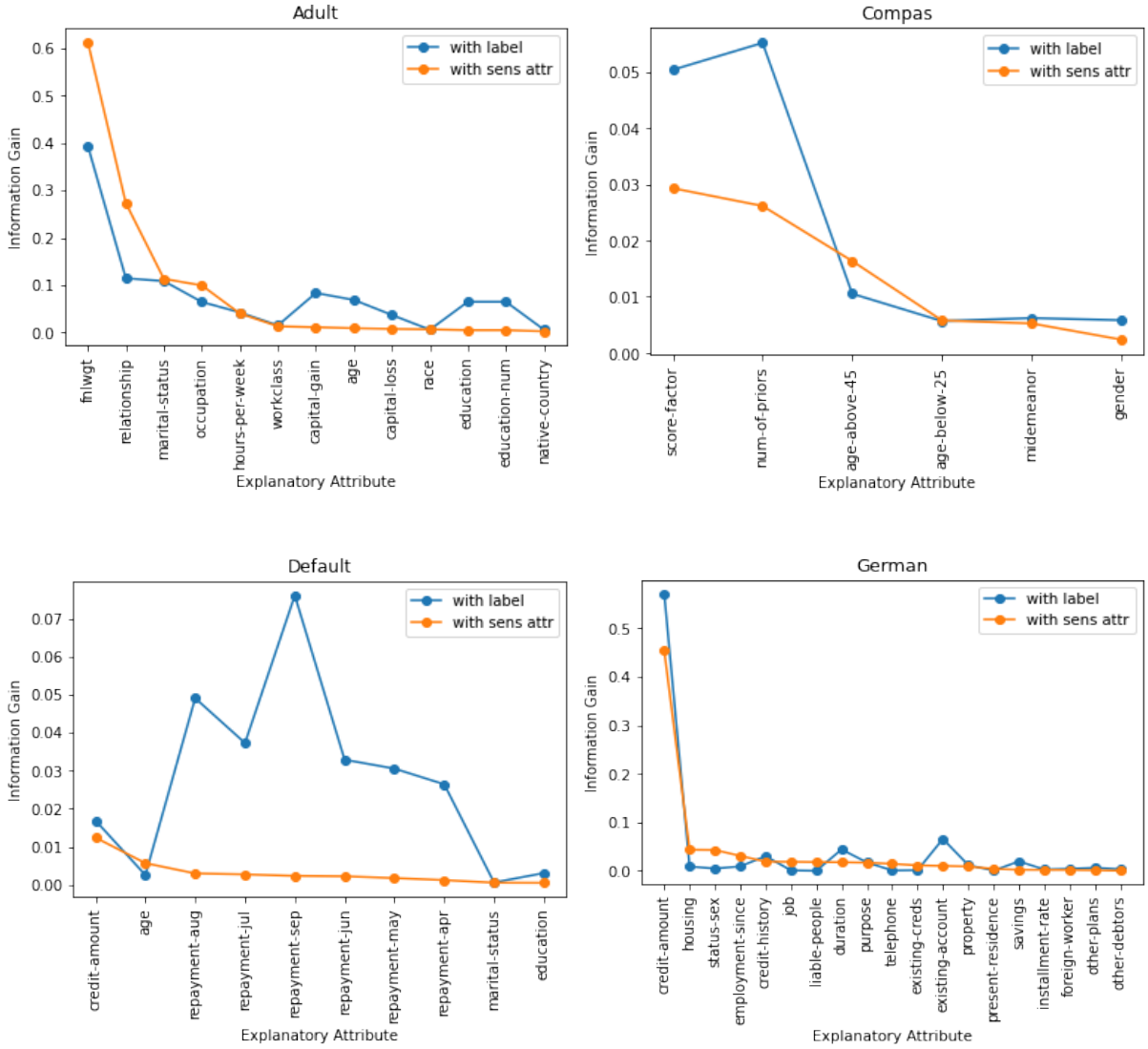

Figure 1: Relations between sensitive, explanatory attributes and labels

about the sensitive attribute are selected. For these attributes, the illegal discrimination is computed when considering the attribute as the explanatory attribute. The results are given in Figure 2.

In addition to the total discrimination and illegal discrimination present in the original data, Figure 2 also contains graphs for the illegal discrimination is the data after Local Massaging and Local Preferential sampling are performed. The graphs show a clear reduction in illegal discrimination after the methods are performed. For most datasets and attributes, the size of the reduction is slightly larger when using Local Massaging compared to Local Preferential Sampling, however the difference is minimal and could be dependent on the used datasets. Based on Figures 1 and 2, for every dataset one variable is selected as the explanatory attribute to use for Local Massaging and Local Preferential Sampling before CART is applied. The obtained results from CART on the pre-processed data can be compared to the results from

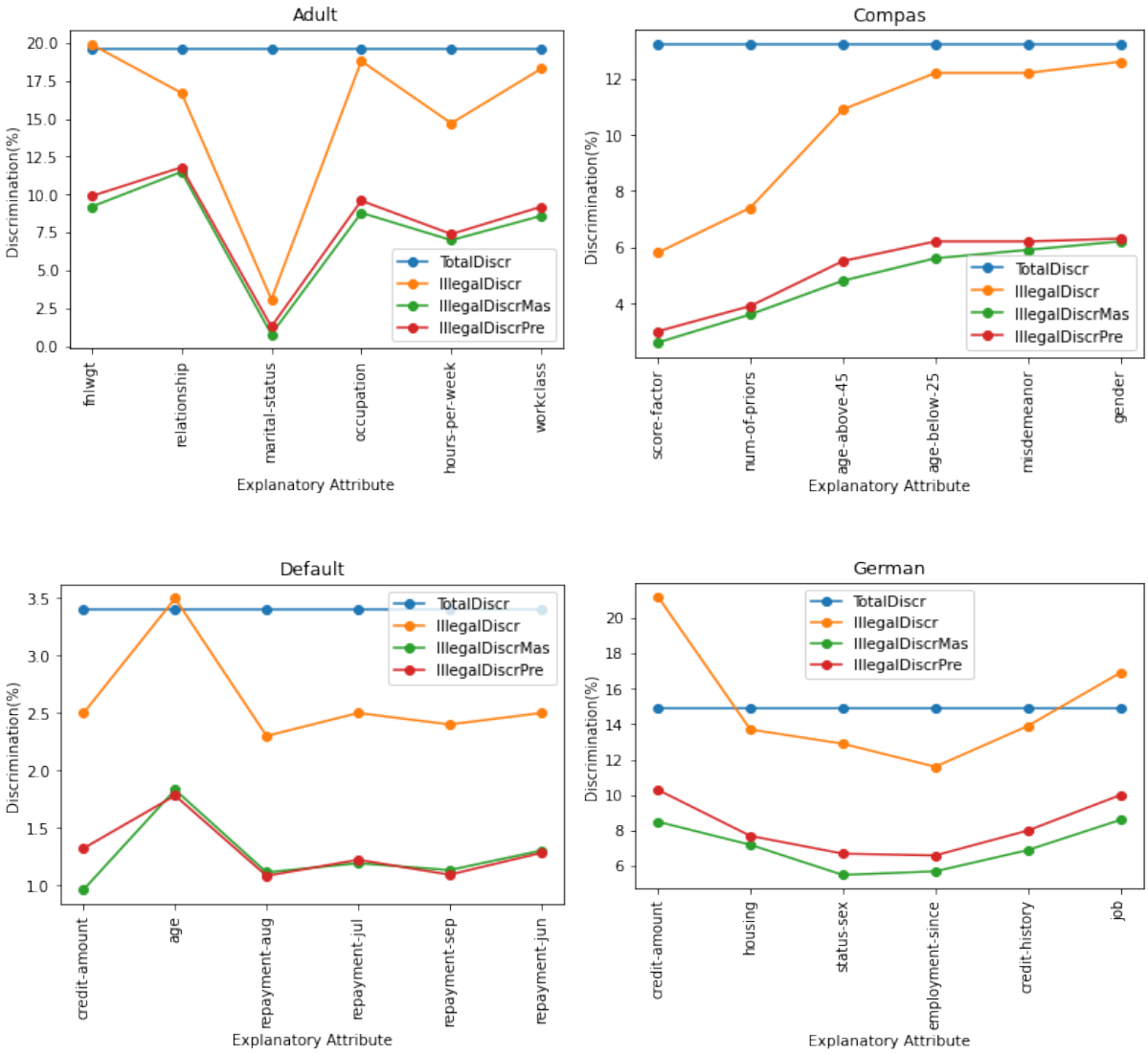

Figure 2: Discrimination in the dataset before and after Local Massaging and Local Preferential Sampling

FairCG. The chosen explanatory attributes are required to be correlated with the the sensitive attribute and give objective information about the label. Using this attribute as the explanatory attribute should lead to a reduction of the illegal discrimination in the dataset. The following variables are used for the different datasets: *hours-per-week* for Adult, *num-of-priors* for Compas, *credit-amount* for Default and *housing* for German. These variables are chosen since using them results in a significant reduction in illegal discrimination. In addition, these variables give objective information about the label and choices based on these attributes are generally legal. Note that it is also possible to use multiple explanatory attributes, as is described by Kamiran et al. (2013).

4.3.2 CART on pre-processed data

After Local Massaging and Local Preferential Sampling are preformed using the explanatory attributes mentioned above, a decision tree is trained on the processed training data. Values 2 and 5 are used for the maximum depth of the decision tree. The results for the best accuracy and best fairness are given in

Table 3. The first two rows contain the results for Local Massaging, the last two rows contain the results for Local Preferential Sampling. Similar to FairCG, fairness is defined as the difference in Equality of Opportunity for the two groups in the data.

From Table 3 can be derived that CART performs similar whether Local Massaging or Local Preferential Sampling is performed and there are only small differences in accuracy and fairness. The results from CART can be compared to the results from FairCG. For the Adult and Compas datasets, FairCG outperforms CART considering both accuracy and fairness. For the Default dataset, CART performs slighty better on both accuracy and fairness. For the German dataset, it is not possible to state which method performs better, since the results from the folds contained a high variability.

From the results from the used datasets, it seems that FairCG performs slightly better than CART trained on pre-processed data. Something that should be taken into account tho is that the FairCG method was designed in such a way that the fairness, defined as the difference in equality in opportunity, could be controlled for. When applying Local Massaging and Local Preferential Sampling, the main focus is on removing illegal discrimination, which is not equivalent to Equality of Opportunity. Therefore CART on pre-processed data might perform better if other fairness metrics are considered, like statistical parity. Also, generating the decision rules using the FairCG method is more time consuming than CART in combination with Local Massaging or Local Preferential Sampling.

Table 3: Mean Accuracy and Fairness results from CART on pre-processed data

	Adult		Compas		Default		German	
	Accuracy	Fairness	Accuracy	Fairness	Accuracy	Fairness	Accuracy	Fairness
				Local Massaging				
Depth: 2	81.9 (0.5)	14.1 (2.8)	63.5 (2.2)	12.0 (5.7)	82.0 (0.7)	0.1 (2.7)	67.7 (4.7)	2.4 (11.1)
Depth: 5	81.7 (0.6)	15.8 (4.6)	63.5 (2.2)	12.0 (5.7)	82.1 (0.6)	0.5 (3.7)	69.8 (3.9)	1.6 (12.3)
				Local Preferential Sampling				
Depth: 2	81.9 (0.6)	14.3 (4.9)	63.5 (2.2)	12.0 (5.7)	82.0 (0.9)	0.3 (4.0)	69.8 (6.5)	3.9 (15.3)
Depth: 5	81.6 (0.6)	16.9 (5.5)	63.5 (2.2)	12.0 (5.7)	82.1 (0.8)	1.3 (4.8)	71.8 (6.2)	1.1 (10.1)

4.3.3 FairCG and CART combined with Local Massaging

So far, FairCG and data pre-processing have been considered as separate methods to obtain a fair classification. In this final section, FairCG is applied to local massaged data to compare the performance of FairCG to CART when both methods are trained on data from which illegal discrimination is removed. To remove the discrimination, Local Massaging is used, since this results in less illegal discrimination in the used datasets compared to using Local Preferential Sampling (see Figure 2).

FairCG requires the choice of parameters ϵ and C, for respectively the maximum unfairness and complexity. As mentioned before, choosing $\epsilon = 0.00$ yields no valid results for most folds of most datasets. Therefore, now only $\epsilon = 0.01$ and $\epsilon = 1.0$ are used. For the complexity parameter C, the values are used which gave the best average results for FairCG on the original data, taking into account both accuracy and fairness. This implies using maximum complexities 100, 20, 20 and 40 for respectively the Adult,

Compas, Default and German datasets. The results from FairCG trained on local massaged data are given in Table 4.

The bottom row of Table 4 contains the results from CART applied on local massaged data. For all datasets the maximum depth for CART equals 2, except for the German dataset, for which this is 5. These depths yielded the best results for the datasets, as can be seen in Table 3.

Considering accuracy, the performance of FairCG is better for the Compas and German datasets, while CART performs better for the Adult dataset. When using strict fairness constraints for FairCG ($\epsilon = 0.01$), FairCG is more fair for the Adult and Compas datasets, but not for Default and German. Overall, in combination with Local Massaging, the results do not imply that one method clearly outperforms the other.

Table 4: Mean Accuracy and Fairness results from CART and FairCG on pre-processed data

	Adult		Compas		Default		German	
	Accuracy	Fairness	Accuracy	Fairness	Accuracy	Fairness	Accuracy	Fairness
				FairCG				
$\epsilon = 0.01$	81.3 (1.1)	13.0 (8.0)	66.6 (2.2)	8.5 (6.4)	82.0 (0.9)	0.7 (4.6)	72.7 (5.0)	8.6 (8.0)
$\epsilon = 1.00$	80.6 (0.9)	27.6 (6.0)	68.3 (2.4)	19.6 (8.2)	82.0 (0.9)	0.2 (2.2)	72.5 (3.0)	6.8 (9.6)
				CART				
	81.9 (0.5)	14.1 (2.8)	63.5 (2.2)	12.0 (5.7)	82.0 (0.7)	0.1 (2.7)	69.8 (3.9)	1.6 (12.3)

5 Conclusion

Over the last couple of years, people became more aware of the importance of classification methods to be both fair and interpretable. Some methods currently used in everyday life, appear to be unclear to those who have to use them, or they discriminate individuals based on their characteristics. This paper considers two methods for the task of fair and interpretable classification: Fair Column Generation (FairCG) and data pre-processing in combination with Classification and Regression Trees (CART). Both methods are described in detail and applied on standard machine learning datasets. Their performance in terms of accuracy and fairness is reviewed and the following research question is answered: *'How does FairCG perform compared to CART, where CART is applied to pre-processed data, regarding accuracy and fairness?'*

The FairCG method is introduced by Lawless and Günlük (2020) and uses integer programming to generate a set of decision rules for classification. The objective is set to minimize Hamming Loss and fairness constraints are added to put a bound and the maxiumum unfairness, where Equality of Opportunity is used as fairness metric. With data pre-processing, illegal discrimination is removed from the data by performing either Local Massaging or Local Preferential Sampling. Then CART is applied to the processed training data.

As was shown by Lawless and Günlük (2020), FairCG showed similar results regarding accuracy com-

pared to simple interpretable models for binary classification, but superior fairness results. This paper obtained similar results, with small differences which could be caused by the randomness component of the method, or by the usage of different hyperparameters. Pre-processing the data led to a reduction of illegal discrimination in the used dataset. The size of the reduction depends on the choice of the explanatory attribute. CART was applied to the processed data and the performance was comparable to FairCG, although being worse. Since Local Massaging and Local Preferential Sampling were not specifically designed to obtain Equality of Opportunity, using a different fairness metric to review the methods might lead to a more favorable outcome for the data pre-processing. Both methods gave promising results regarding accuracy and fairness and which method is optimal will depend on the exact application. FairCG and CART obtain similar results when they are both applied on local massaged data.

The methods showed in this paper could be useful in practice when sensitive decisions need to be made, like hiring decisions and loan approval. In further research, FairCG could be optimized using other fairness constrains, depending on the wish of the user. Also data pre-processing can be combined with other classification methods like Logistic Regression and Support Vector Machine, and multiple attributes could be used as explanatory attribute to remove a larger part of the illegal discrimination. The methods in this paper can also be compared to other fairness methods, like post-processing the classification results.

References

Aghaei, S., Azizi, M. J., & Vayanos, P. (2019). Learning optimal and fair decision trees for non-discriminative decision-making. *Proceedings of the AAAI Conference on Artificial Intelligence*, *33*(01), 1418–1426.

Barocas, S., & Selbst, A. D. (2016). Big data's disparate impact. *Calif. L. Rev.*, *104*, 671.

Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). *Classification and regression trees*. CRC press.

Cano, A., Zafra, A., & Ventura, S. (2013). An interpretable classification rule mining algorithm. *Information Sciences*, *240*, 1–20.

Carvalho, D. V., Pereira, E. M., & Cardoso, J. S. (2019). Machine learning interpretability: A survey on methods and metrics. *Electronics*, *8*(8), 832.

Chouldechova, A., & Roth, A. (2018). The frontiers of fairness in machine learning. *arXiv preprint arXiv:1810.08810*.

Dash, S., Günlük, O., & Wei, D. (2018). Boolean decision rules via column generation. *arXiv preprint arXiv:1805.09901*.

Dembczyński, K., Kotłowski, W., & Słowiński, R. (2010). Ender: A statistical framework for boosting decision rules. *Data Mining and Knowledge Discovery*, *21*(1), 52–90.

Dunkelau, J., & Leuschel, M. (2019). Fairness-aware machine learning.

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.

Günlük, O. (2020). Fair and interpretable decision rules for binary classification. http://www.ipam.ucla.edu/abstract/?tid=16780&pcode=DLC2021

Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. *arXiv preprint arXiv:1610.02413*.

Kamiran, F., & Calders, T. (2009). Classifying without discriminating. *2009 2nd International Conference on Computer, Control and Communication*, 1–6.

Kamiran, F., Calders, T., & Pechenizkiy, M. (2010). Discrimination aware decision tree learning. *2010 IEEE International Conference on Data Mining*, 869–874.

Kamiran, F., Žliobaitė, I., & Calders, T. (2013). Quantifying explainable discrimination and removing illegal discrimination in automated decision making. *Knowledge and information systems*, *35*(3), 613–644.

Lakkaraju, H., Bach, S. H., & Leskovec, J. (2016). Interpretable decision sets: A joint framework for description and prediction. *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 1675–1684.

Lawless, C., & Günlük, O. (2020). Fair and interpretable decision rules for binary classification.

Oneto, L., & Chiappa, S. (2020). Fairness in machine learning. *Recent trends in learning from data* (pp. 155–196). Springer.

Osoba, O. A., & Welser IV, W. (2017). *An intelligence in our image: The risks of bias and errors in artificial intelligence*. Rand Corporation.

Pallara, A. (1992). Binary decision trees approach to classification: A review of cart and other methods with some applications to real data. *Statistica Applicata*, *4*(3), 255–285.

Pedreshi, D., Ruggieri, S., & Turini, F. (2008). Discrimination-aware data mining. *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, 560–568.

Song, Y.-Y., & Ying, L. (2015). Decision tree methods: Applications for classification and prediction. *Shanghai archives of psychiatry*, *27*(2), 130.

Su, G., Wei, D., Varshney, K. R., & Malioutov, D. M. (2015). Interpretable two-level boolean rule learning for classification. *arXiv preprint arXiv:1511.07361*.

Ustun, B., Traca, S., & Rudin, C. (2013). Supersparse linear integer models for interpretable classification. *arXiv preprint arXiv:1306.6677*.

Zafar, M. B., Valera, I., Rogriguez, M. G., & Gummadi, K. P. (2017). Fairness constraints: Mechanisms for fair classification. *Artificial Intelligence and Statistics*, 962–970.

Zeng, J., Ustun, B., & Rudin, C. (2015). Interpretable classification models for recidivism prediction. *arXiv preprint arXiv:1503.07810*.

Appendix

Table 5: Results FairCG for Adult dataset

		$\epsilon = 0.0$		$\epsilon = 0.01$		$\epsilon = 1.0$	
	Complexity	Accuracy	Fairness	Accuracy	Fairness	Accuracy	Fairness
Test	80	81.1 (2.2)	0.8 (4.3)	82.5 (1.0)	1.4 (7.2)	82.1 (0.7)	6.0 (9.0)
	90	80.8 (2.2)	1.1 (3.9)	82.5 (1.0)	1.4 (7.5)	82.1 (0.7)	6.1 (10.1)
	100	81.0 (2.2)	0.3 (4.8)	82.5 (1.0)	1.2 (7.5)	82.1 (0.7)	6.0 (9.8)
Train	80	81.4 (2.1)	0.0 (0.0)	82.9 (0.6)	0.1 (0.9)	82.7 (0.5)	6.0 (8.5)
	90	80.9 (2.3)	0.0 (0.0)	83.0 (0.3)	0.3 (0.9)	82.7 (0.5)	6.0 (9.5)
	100	81.3 (2.4)	0.0 (0.0)	83.0 (0.3)	0.3 (0.9)	82.7 (0.5)	4.9 (10.0)

Table 6: Results FairCG for Compas dataset

		$\epsilon = 0.0$		$\epsilon = 0.01$		$\epsilon = 1.0$	
	Complexity	Accuracy	Fairness	Accuracy	Fairness	Accuracy	Fairness
Test	10	-	-	63.9 (1.7)	0.3 (6.0)	67.4 (2.9)	24.1 (3.3)
	15	59.5 (3.5)*	0.8 (3.4)*	64.8 (1.4)	0.2 (2.4)	68.0 (2.4)	24.3 (2.6)
	20	62.9 (2.4)*	0.9 (3.9)*	65.8 (2.1)	0.5 (4.8)	68.0 (2.2)	23.6 (6.9)
Train	10	-	-	64.7 (0.4)	0.0 (0.5)	67.8 (0.3)	24.2 (0.9)
	15	61.8 (3.6)*	0.0 (0.0)*	66.2 (0.4)	0.3 (0.6)	68.3 (0.3)	24.5 (0.9)
	20	63.4 (2.6)*	0.0 (0.0)*	66.7 (0.3)	0.3 (0.5)	68.8 (0.3)	23.8 (0.7)

*Only folds with valid results were used to obtain these values

Table 7: Results FairCG for Default dataset

		$\epsilon = 0.0$		$\epsilon = 0.01$		$\epsilon = 1.0$	
	Complexity	Accuracy	Fairness	Accuracy	Fairness	Accuracy	Fairness
Test	10	-	-	81.9 (0.7)	0.5 (4.2)	82.0 (0.6)	0.9 (4.2)
	15	-	-	81.9 (0.7)	0.4 (4.1)	82.0 (0.6)	2.3 (2.8)
	20	79.8 (3.0)*	0.1 (0.7)*	81.9 (0.8)	0.1 (4.5)	81.9 (0.7)	1.3 (4.2)
Train	10	-	-	82.1 (0.1)	0.3 (0.3)	82.1 (0.4)	0.5 (0.8)
	15	-	-	82.1 (0.1)	0.3 (0.4)	82.1 (0.1)	0.7 (0.9)
	20	80.1 (2.4)*	0.0 (0.0)*	82.1 (0.1)	0.3 (0.4)	82.1 (0.2)	0.8 (0.8)

*Only folds with valid results were used to obtain these values

Programming code: READ ME

The code for the thesis 'Fair and Interpretable Methods for Binary Classicfication' consits of two parts, FairCG and Data pre-processing.

FairCG

All code used for the FairCG is originally made by Connor Lawless, see https://github.com/conlaw. The code is adjusted for this research.

In the notebook Fair CG Rule Generation, the FairCG is performed with an empty rule set, to create a rule set which is used as starting point for FairCG in the FairCG Trials notebook. Both notbooks call the runSingleTest function from testhelpers.py. Here, the fit function from classifier.py is called. Here, rules are generated, added to the problem, and the master problem is solved. In addition to the fit function, in runSingleTest a classifier object is made. In turn, here a compact double sided master object, DNF Rule Model object, General Rule Generator object and Equality of Opportunity object are made.

The results and rules folders respectively contain the results and rules obtained from the trials. The databinerize folder containts the data and a notebook to binerize the data.

FairCG Trials Massaging.py contains the code for FairCG trained on local massaged data. The folder resultsmas contains the results from FairCG trained on local massaged data.

Data pre-processing

The procedure of data pre-processing consits of three parts:

1. Calculating the information gain
2. Computing the discrimination before and after performing Local Massaging and Local Preferential Sampling
3. Applying CART to the pre-processed data

Calculating the information gain

In the notebook calcInformationGain, the information gain is calculated for every possible explanatory attribute. The results are plotted in a graph for every dataset. The notebook uses the function calcInfoGain from IGCalculator.py to compute the information gain.

Computing the discrimination before and after performing Local Massaging and Local Preferential Sampling

In the notebook calcDiscrimination, for every dataset the total and illegal discrimination is computed. Then Local Massaging and Local Preferential Sampling are performed and the illegal discrimination is

computed again. The notebook uses the method computeDiscr from discrCalculator.py to compute the total, explainable and illegal discrimination. The notebook uses the methods localMassaging and localPrefSampling from processdata.py to perform local massaging and local preferential sampling.

Applying CART to the pre-processed data

In the notebooks CARTadult, CARTcompas, CARTdefault and CARTgerman, a classification tree is trained on pre-processed data. This is done for both local massaging and local preferential sampling. 10 fold cross calidation is used and the mean and standard deviations for accuracy and fairness are given. The notebook uses the methods localMassaging and localPrefSampling from processdata.py to perform local massaging and local preferential sampling.

To binerize the data, the methods in the file binerizer.py are used. This file is written by Conor Lawless, see https://github.com/conlaw