ERASMUS UNIVERSITY ROTTERDAM

Erasmus School of Economics

Bachelor Thesis: BSc² in Econometrics and Economics

2020 Major: Business Analytics and Quantitative Marketing

Fair and interpretable rules for binary classification

Student
M.H. Kroon
Student ID number
451067

Supervisor
dr. M.H. Akyuz
Second Assessor
MSc Utku Karaca

July 3, 2021

In this paper, a framework is presented for binary classification. The goal of this framework is to ensure fairness and interpretability without compromising accuracy. Research by Lawless and Günlük (2021) has provided a framework tested with multiple datasets. An integer programming formulation is used in combination with the large-scale optimization technique of column generation to ensure the model terminates. This research continues their work and extends it with a new smaller dataset and a different fairness metric. The new metric did not show any sign of improvement and the smaller dataset was too small for the model to perform well.

Contents

1 Introduction

The amount of decision-making done automatically by artificial intelligence has skyrocketed in the last decade. It has been making our lives easier, as computers are doing part of our thinking process. From banks deciding on whom to give out loans, colleges deciding which students to accept and legal systems to predict which criminals might fall into recidivism; all of these problems can be solved by machine learning. Machine learning uses several classification techniques to solve these problems. Classification is defined in the Cambridge Dictionary as: "The act or process of dividing things into groups according to their type" (The Cambridge Dictionary: English Dictionary, 2021). When the classification is binary, there are only two groups that can be chosen. As computer thinking in essence comes down to binary thinking, one would assume this is easy for these machines. However, as computers are trained to think like humans, human thinking biases can slip into the algorithm (Fuchs, 2018). In the research by Fuchs (2018), a study was discussed where black defendants were twice as likely to be predicted for recidivism as white defendants. Such a bias is one example of a problem happening more often, where algorithms are giving biased predictions.

Research done by Lawless and Günlük (2021) on binary classification has resulted in a framework where an integer programming approach is used to create a rule-based model. The input for this model are rulesets in disjunctive normal form (DNF). This research is recent and there is still much research to be done in this field. In this paper, the research by Lawless and Günlük (2021) will be extended. The research question will be: *Can DNF rulesets improve the fairness and interpretability of a model, without compromising the accuracy of the model?* To support this research question, three sub-questions are formulated.

- How does the work by Lawless and Günlük (2021) perform on a different computational platform?
- How does the model perform with a stricter fairness metric than the one applied by Lawless and Günlük (2021)?
- How does the model perform for a new smaller dataset?

The second and third sub-questions are the start of the extension. The previous research imposed a fairness metric, which only minimized the rate of people getting negative predictions while being positive. Another metric can be imposed, which also minimizes the

rate of people getting positive predictions while being negative. The three datasets researched by Lawless and Günlük (2021) are rather large. Smaller datasets can however also be of interest and the performance by the model of Lawless and Günlük (2021) is not known. Therefore, the model will also be performed on a relatively small dataset.

The reason why this research is highly relevant today is that more and more trust is being placed into machine learning; only the way the model makes assumptions and the fairness at play might sometimes be overlooked. Apart from the mathematical background, these are also socially important and heavily discussed themes in today's society. Institutional racism in the US alone has cost about 16 trillion dollars over the past 20 years, according to estimates of economists of Citigroup (Peterson and Mann, 2020). Fair and interpretable models might help in diminishing this problem.

The remainder of this paper is divided into multiple sections. Section 2 is a literature review, which will give a summary of the surrounding literature. In section 3 the methodology is presented, which introduces the framework used to create the model. The computational results are displayed in section 4, which starts with an exploration of the datasets and ends with the results. In the final section, a conclusion for the different research questions will be formulated and a suggestion for further research is proposed.

2 Literature

Classification is an important feature of data analysis. In 1994, Michie, Spiegelhalter, and Taylor (1994) stated that the classification and decision problem were the solutions to urgent problems arising in different fields, like science, medicine and commerce. After more than 25 years, the topic is still relevant and increased computational power has allowed an increasing number of problems to be solved. There is a great deal of diverse literature on this topic. This section will first consider different binary classification frameworks and explore why one is chosen above the other. There has been a significant amount of research on how to proceed with these kinds of formulations and optimal solutions, which will be discussed subsequently.

2.1 Binary Classification

Binary classification is a special form of statistical classification, where a multivariate dataset is analysed to find boolean results (Kumari and Srivastava, 2017). The information on the explanatory variables of the dataset is used to come up with rulesets, from which the model can make a prediction (Hand and Henley, 1997). There are several pieces of academic research of creating rulesets. One way is via decision trees, where research by Quinlan (1986) provides the best account. A decision tree starts for a certain explanatory variable and then all of the branches are the decisions in that step. In the next step, a new variable is analysed and from the previously created branches, new ones are generated. Without any form of pruning, these trees grow exponentially large and become too lengthy to solve. This is why Rivest (1987) present a way via decision lists, which consists of 'If-then-else if-...-else' statements. The implication hereof is that more complex decisions can occur at a node, as one does not need all the different choices at a node. An advantage of these models above decision trees is that the interpretability is better. The downside of this method is that the running time is polynomial with the input variables, thus will be cumbersome for large datasets.

Another way of creating rulesets is via decision rule sets, which can be done in two forms. The first form is in disjunctive normal form (DNF). A rule is in DNF if it is a disjunction of conjunctions (Pfahringer, 2017), in other words an 'OR-of-ANDs'. The second form is the conjunctive normal form (CNF), which is the opposite of the DNF; an 'AND-of-ORs'. Most current research in the field of binary classification uses the DNF (Dash, Günlük, and Wei, 2018; Lawless and Günlük, 2021). Another reason behind using the DNF above the decision list and decision trees is based on research by Lakkaraju, Bach, and Leskovec (2016). This research found it was as accurate as the other methods, yet the model was substantially smaller and with better user interpretability.

Apart from binary classification via rulesets, there are also other techniques popular which require different sorts of input. Support vector machines (SVM) is one of these techniques, where data is divided into certain groups based on certain characteristics.The method has grown in popularity in machine learning in the last 30 years and is still popular among mathematicians and economists (Meyer, Leisch, and Hornik, 2003; Gandhi, 2018). SVM has proven itself to be more efficient than several other techniques, especially in the field of pattern recognition (Pradhan, 2012). The only downside of this method is that by itself is not easy to interpret, which is why this paper will not be using it. (Martin-

Barragan, Lillo, and Romo, 2014)

2.2 Classification Framework

After these rulesets have been created, a clear framework is needed to come up with the optimal ruleset. In a ruleset in DNF form, there is a finite number of decision rules (Mendelson, 2009). Lawless and Günlük (2021) suggest minimizing over a finite number of rules can best be done by using integer programming. The distinction between a linear and integer program is that in an integer program some variables are constrained to be an integer (Vanderbei, 2015), which is in line with the datasets used in this paper. When there are some constrained to be an integer and some not, it is also referred to as a mixed integer program. It would however be difficult for this program to terminate, as by binarizing most features of the data, the problems would grow exponentially large.

Dash et al. (2018) tackle this problem by suggesting using the large-scale optimization technique of column generation. Usage of column generation in the literature dates back more than 50 years. An important contribution was made by Appelgren (1969), who, with limited computer power, was able to solve a linear program with column generation. Appelgren (1969) showed using Dantzig-Wolfe decomposition how column generation could be used for these programs (Dantzig and Wolfe, 1960). A clear column generation framework was formulated by Barnhart, Johnson, Nemhauser, Savelsbergh, and Vance (1998). However, there were not many experiments being done as computer power was lagging. In the first decade of 2000, this changed and with the availability of powerful computers and large-scale electronic data the theories could be better put into practice.

The goal of this IP formulation is minimizing the *Hamming Loss* (Dash et al., 2018). The hamming loss is described as the distance between the current rule and the closest rule that correctly specifies a sample and sparsity (Su, Wei, Varshney, and Malioutov, 2015). The two main factors Dash et al. (2018) focused on were accuracy and interpretability, there was no notion of fairness.

2.3 Fairness Metrics

In addition to the model of Dash et al. (2018), Lawless and Günlük (2021) added a fairness metric to the formulation. The metric was deemed necessary, as then the measurement error could be better controlled in both groups. The trade-off between accuracy and

fairness would be made less complicated. Hardt, Price, and Srebro (2016) defined two metrics, called equality of opportunity and equalized odds. These two metrics are based on type I and type II errors, which are described below.

Null hypothesis is	**True**	**False**
Rejected	Type I error	Correct decision
Not rejected	Correct decision	Type II error

Table 1: Type I and II errors

Equality of opportunity is based on type II errors (Wackerly, Mendenhall, and Scheaffer, 2014), which are also referred to as false negatives. The extension will look at the stricter metric, equalized odds. The difference between equality of opportunity and equalized odds is that the equalized odds on top of type II errors also include the type I errors. The name of this setting is Hamming Equalized Odds (Hardt et al., 2016).

3 Methodology

Two different formulations will be given in this section, which can both be used to find the optimal ruleset. The first formulation given will result in an exponential ruleset, which can be too extensive to solve. Hence, the second formulation will tackle this problem by implementing column generation.

3.1 Mixed inter program formulation

The goal of the mixed integer program (MIP) is to find the optimal ruleset with the best accuracy and fairness. The input for the model will be the binarized variables of each datapoint. The full set of possible rules is K, which is a finite set of all binary combinations of the different clauses. K_i denotes the starting ruleset obtained from the data. The goal of the formulation is to obtain the highest accuracy with the lowest number of rules. Lawless and Günlük (2021) derived the following formulation from the work of Dash et al. (2018).

$$
\begin{aligned}
z_{mip} = \quad & \text{minimize} && \sum_{i \in P} \zeta_i + \sum_{i \in Z} \sum_{\kappa \in K_i} w_\kappa && && (1)\\
& \text{subject to} && \zeta_i + \sum_{\kappa \in K_i} w_\kappa && \geq 1 \quad i \in P && (2)\\
& && C\zeta_i + \sum_{\kappa \in K_i} 2w_\kappa && \leq C \quad i \in P && (3)\\
& && \sum_{\kappa \in K} c_\kappa w_\kappa && \leq C && (4)\\
& && w \in \{0,1\}^{|K|}, \zeta \in \{0,1\}^{|P|} && && (5)\\
& && \frac{1}{|P_1|} \sum_{i \in P_1} \zeta_i - \frac{1}{|P_2|} \sum_{i \in P_2} \zeta_i && \leq \epsilon_1 && (6)\\
& && \frac{1}{|P_2|} \sum_{i \in P_2} \zeta_i - \frac{1}{|P_1|} \sum_{i \in P_1} \zeta_i && \leq \epsilon_1 && (7)
\end{aligned}
$$

In the objective function, the variables ζ_i and w_k are binary and respectively the number of misclassified data points and a variable indicating if rule $k \in K$ is used. This implies that if ζ_i were 1, it would be misclassified. A misclassification is in line with what was previously called a false negative. The first constraint (2) ensures that the false negative count goes up if a point is incorrectly specified. C in (3) and (4) is the maximum complexity allowed; complexity is the clause's length. When the rule is not being satisfied ζ_i needs to be 1, constraint (3) is there to ensure no other value can be taken. The bound for the complexity is in constraint (4). The final two constraints, (6) and (7), are obtained from the fairness metric, which were derived by Hardt et al. (2016). The name of this metric is equality of opportunity and is given by:

$$P(d(X) = 0|Y = 1, G = g) = P(d(X) = 0|Y = 1, G = g') \; \forall g, g' \in G \tag{8}$$

In equation (8) the probability of a prediction being false negative should be equal across groups. As this is difficult in practice, the difference between the two probabilities should be as small as possible. This can be expressed in the following equation:

$$\triangle(d) = \max_{g,g' \in G} |P(d(X) = 0|Y = 1, G = g) - P(d(X) = 0|Y = 1, G = g')| \tag{9}$$

In equations (6) and (7), the chances P_1 and P_2 are these probabilities for two different samples. The total sample is divided into two and the discrepancy between the two samples cannot be bigger than a predefined ϵ_1.

The metric used for the second subquestion is called equalized odds, also derived from Hardt et al. (2016) and is given by:

$$P(d(X) = 1|Y = 0, G = g) = P(d(X) = 1|Y = 0, G = g') \; \forall g, g' \in G \tag{10}$$

Together with equation (8), these two metrics are even stricter in ensuring fairness. The metric will be implemented in the same way as equation (8) is implemented in equations (6) and (7).

3.2 Column generation framework

The problem with this formulation is that, even though it is finite, the number of different rules grows exponentially. The column generation algorithm is used to be able to solve the problem in a reasonable time. A subset of the variables is taken from the LP relaxation of the MIP and this restricted master linear program (RMLP) is solved. Among the non-basic variables, in a principle way the most promising non-basic variables are searched. These are the variables with the most negative reduced cost (Bazaraa, 2010). The reduced cost is defined by Dash et al. (2018) as:

$$\sum_{i \in Z} \delta_i - \sum_{i \in P} (2\alpha_i - \mu_i)\delta_i + \lambda c \tag{11}$$

In the first term of equation (11) δ_i is a binary variable, which denotes for data point i if it rule wrongly classifies sample i. The second term is the same δ_i multiplied with $(2\alpha_i - \mu_i)$, these are the dual variables from the constraint of (3). The complexity, c is included in the final term, multiplied with the dual variables of (3), λ. When the reduced costs are minimized, no new rule can enter and the model is optimized. For this reason, it is the objective function in the column generation framework.

$$\begin{aligned}
z_{cg} = \quad & \text{minimize} \quad \sum_{i \in P} \delta_i + \sum_{i \in P} (2\alpha_i - \mu_i) + \lambda(1 + \sum_{j \in J} z_j) && (12)\\
& \text{subject to} \quad D\delta_i + \sum_{j \in S_i} z_j \quad \leq D \quad i \in I^- && (13)\\
& \qquad\qquad\quad \delta_i + \sum_{j \in S_i} z_j \quad \geq 1 \quad i \in I^+ && (14)\\
& \qquad\qquad\quad \sum_{j \in J} z_j \quad \leq D && (15)\\
& \qquad\qquad\quad z \in \{0,1\}^{|J|}, \delta \in \{0,1\}^{|P|} && (16)
\end{aligned}$$

The complexity coefficient denoted with C is now defined as $(1 + \sum_{j \in J} z_j)$, where z_j is a binary variable for $j \in J$ if the rule has feature j. Constraints (13) and (14) ensure that δ_i accurately classifies a data point. The next constraint (15) can be compared to constraint (4) and is there to bound the complexity of the problem, in the form of D.

4 Computational Experiments

4.1 Data

To test the models, four different datasets are going to be used. Three of these will be the same as Lawless and Günlük (2021), one new set will be added. Below a short overview is given.

Name	Data points	Features	Sensitive Variable
Adult	32,561	14	Gender
Compas	5,278	7	Race
Default	30,000	23	Gender
Heart	303	13	Gender

Table 2: Amount of touchpoints

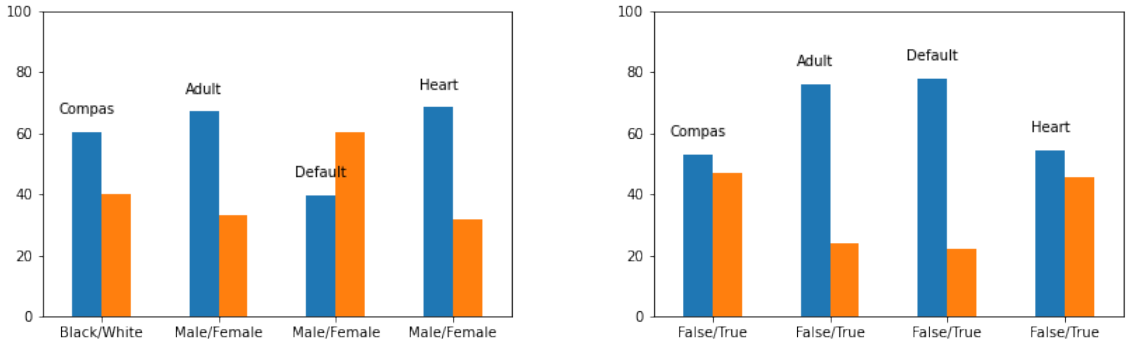
The largest dataset is the *adult* dataset. The data was extracted in 1994 from the US Census Bureau database by Ronny Kohavi and Barry Becker (Kohavi, 1996). The dataset includes information on individual-level on several character attributes, e.g. marital status, age and race. The goal of this database was to predict whether or not an individual's income exceeds $50,000. The sensitive variable in this case is gender.

The second-largest dataset is the *default*. Over 20 explanatory variables are given for Taiwanese credit cardholders in October 2005. More specifically, this dataset includes information on the person itself and its payment history. Yeh and Lien (2009) studied this dataset to predict default payments by cardholders. For this dataset, the sensitive variable is again gender.

The third dataset, referred to as *heart*, is the new dataset and is not included in the works of Lawless and Günlük (2021). The set is on cardiovascular data of patients in the Long Beach and Cleveland Clinic Foundation (Detrano, Janosi, Steinbrunn, Pfisterer, Schmid, Sandhu, Guppy, Lee, and Froelicher, 1989). Information on medical information of patients is presented to check if heart disease is present. Again, the sensitive variable is gender. These three datasets are obtained from the UCI machine learning depository (Dua and Graff, 2017).

The last dataset is the *Compas* dataset, where compas stands for *Correctional Offender Management Profiling for Alternative Sanctions* and it was first analysed by Angwin, Larson, Mattu, and Kirchner (2016). The dataset includes information on arrested people in

Florida in 2013 and 2014. The explained variable here is recidivism, which is the repetition of a previous crime. For this dataset, race is the sensitive variable, as the past has shown black defendants were more often predicted to be at a higher chance of recidivism than was true (Angwin et al., 2016). The dataset was obtained from Kaggle (Ofer, 2017).



(a) Division of sensitive values

(b) Division of explained values

Figure 1: Different divisions for the data

In figure 1a the division in the different datasets is given between the sensitive values. It is given on a scale of 100%, where the sum of the two bars is the full dataset. In the description of the compas dataset it was mentioned that black defendants were twice as likely to be predicted to fall into recidivism, this group is also the majority group in this set. The adult and heart set are both male-dominated, while the default set is female-dominated. The interesting part will be to see if this has any implications for the model, which will be discussed in the result section. Figure 1b works similarly to figure 1a, only now the division between explained variables is given. For *compas* and *heart* the two bars are quite close, which makes sense as arrestants are not unknown to fall into recidivism and heart diseases are on the list of the most common diseases. Contrary, it also makes sense that the other two sets are more imbalanced as in *adult* the threshold was above average income and in *default* as it is more likely that people pay off their debt.

In figure 2, the division within the sensitive values is specified. The adult and default set show similar patterns for both groups, only the difference in the adult set is more prominent. For the compas and heart set, the groups exhibit opposite results for the two different groups. It will be interesting to see if the results from the model move into the same direction for adult and default and for compas and heart.

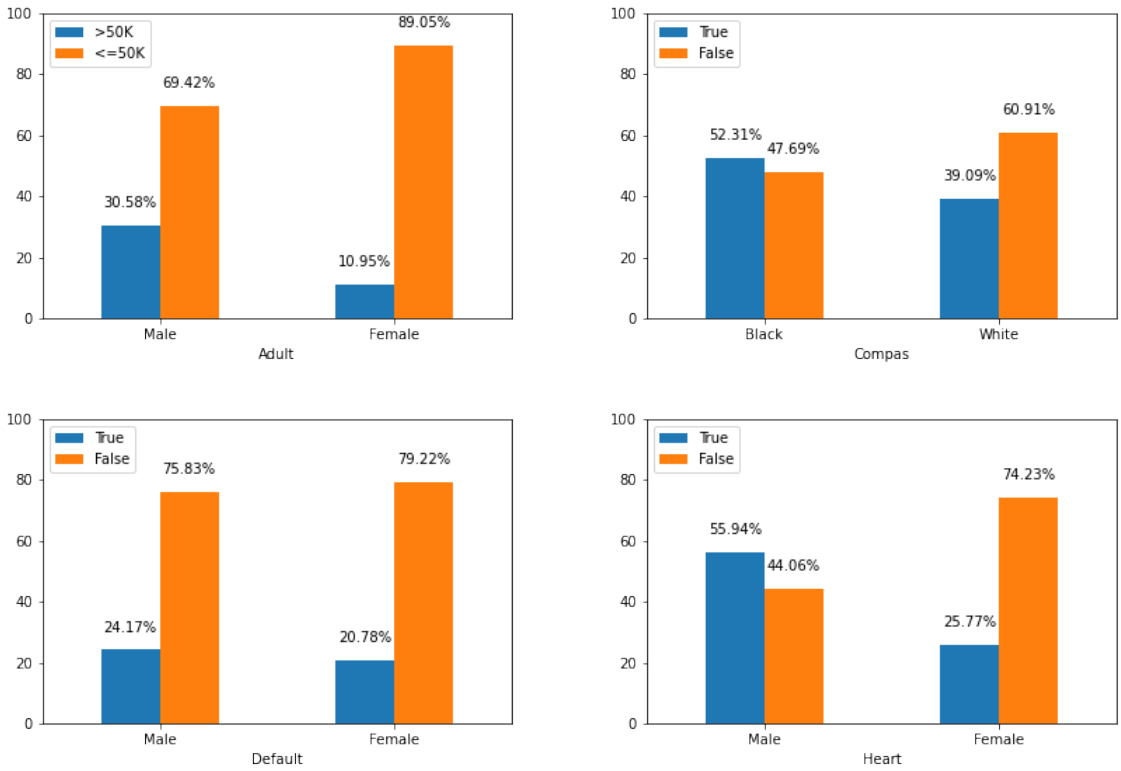


Figure 2: Division within sensitive variables

To make the rulesets, data will need to be adjusted. As discussed in the literature section, the rules will be put in DNF. The heart set will be used to give an example of how this works in practice. To predict whether someone is at risk of having a heart disease with a rule set of two, the following line could be an option:

$$[(\text{Age} \geq 60) \text{ and } (\text{Cholesterol} \geq 300)] \text{ or } [(\text{Chest Pain} = 4) \text{ and } (\text{Blood Sugar} > 120)]$$

To create these rulesets, all of the explanatory variables need to be binarized. Again, the heart dataset will be used to illustrate an example of how this binarization is done. Assume there is a 56-year-old patient with a cholesterol level of 275 and a chest pain of level of 3. As there are more than 40 different ages and more than 150 different cholesterol levels in the dataset, ordinary binarization would result in too large rulesets. Accordingly, thresholds will be implemented. Age can have thresholds at 40, 50, 60 and 70; cholesterol at 100, 200 and 300. For categorical features this can be done via "one-hot" encoding and for numerical features via thresholds. For one-hot encoding all the categories are given and a 1 is written down for the current category and a 0 for the rest. The binarization for the patient then becomes:

$$\begin{bmatrix} 56 \\ 275 \\ 3 \end{bmatrix} = \begin{bmatrix} (0,0,1,1),(1,1,0,0) \\ (0,0,1),(1,1,0) \\ (0,0,1,0),(1,1,0,1) \end{bmatrix}$$

Which as one rule results in:

$$\Big[56, \text{Cholesterol level}, \text{Chest Pain } 3\Big] = \Big[0011, 1100, 001, 110, 0010, 1101\Big]$$

4.2 Results

In this section, the results are presented for the column generation framework for the two different fairness metrics for the different datasets. To find these optimal values, different levels of epsilon and complexity were used. The results will be presented per subquestion.

4.2.1 Replication

	Adult		**Compas**		**Default**	
	Acc	Fair	Acc	Fair	Acc	Fair
Best accuracy	80.67(0.64)	4.05(2.30)	66.19(1.37)	20.23(5.34)	77.80(0.53)	0.01(0.09)
Best fairness	79.73(1.14)	1.80(1.11)	63.22(1.27)	3.50(2.33)	77.80(0.53)	0.01(0.09)

Table 3: Results Equality Of Opportunity

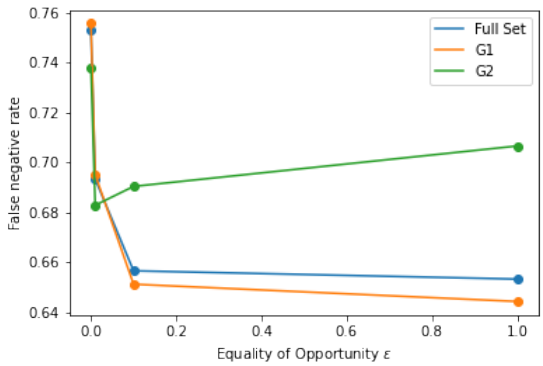
The first subquestion was on how the model of Lawless and Günlük (2021) would perform in a different computational program. The paper did mention a list of which epsilons and complexities were tested, only not the ones eventually used for best accuracy and fairness. In table 3, the results from the equality of opportunity metric, deducted from the research of Lawless and Günlük (2021), are shown. It will not be straightforward to compare these results with the results from Lawless and Günlük (2021), which are presented in table appendix 8 in A.1, due to the ten-fold cross-validation. As in these ten folds such a substantial degree of randomness is introduced. Even though, the results of the t-test between the two sets of results are presented. The reason behind this, is that it still will be helpful starting point for the analysis.

	Adult		**Compas**		**Default**	
	Acc	Fair	Acc	Fair	Acc	Fair
Best accuracy	2.38	6.63	2.84	**0.34**	29.31	**1.51**
Best fairness	4.77	3.91	10.95	2.88	**1.01**	5.52

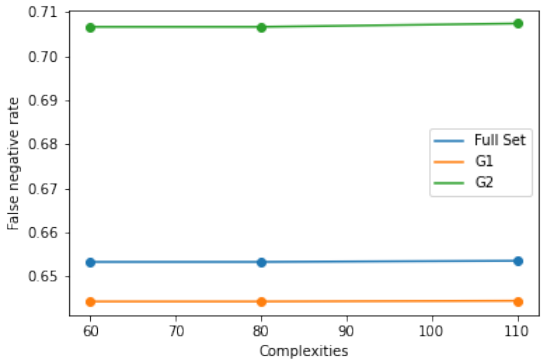
Table 4: Results Equality Of Opportunity t-test

The numbers in bold in table 4 are the only ones for which there is no significant difference in the two tables, at a 5% confidence level. Out of the 12 results, a quarter presents similar results. It is a good start, only not too much value should be attached to this result. When comparing the pure numbers in tables 3 and 7, the following results emerge. For the best accuracy, it can be observed that the accuracy is lower in the current results. Equivalently, the fairness is lower, even though not everywhere significant, which implies the new results are fairer. Reasons for the difference can lie in the choice for epsilon or complexity. Interestingly, it is reversed in the best fairness line. The accuracies and fairness are higher. The implication is that there is less of a trade-off in the model of this paper, as there was in the paper by Lawless and Günlük (2021). One of the reasons hereof could be that more extreme epsilons and complexities were used, however not mentioned in the paper. The epsilons and complexities tested and which are eventually used are presented in table 9 in appendix A.2.

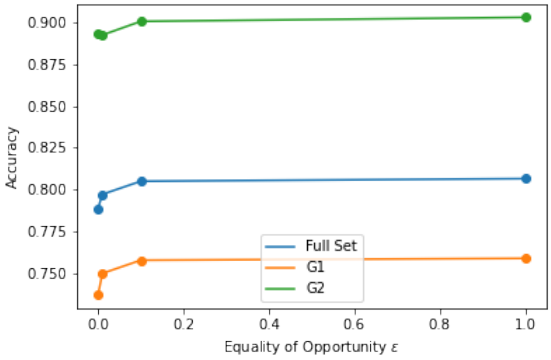
As looking at the numbers can be deceiving, it is better to look at how the different models behave under changing inputs. Three different plots are presented for each dataset. The first plot shows how the false negative rate reacts to a change in ϵ, to show how the choice for epsilon affects the relative amount of wrongly negatively identified samples. In the second plot, the number of complexities is compared to the false negative rate. Apart from the goal of making this model fair, interpretability is also important. Therefore, it is important to investigate how substantial the effect of an increase in complexities is on the false negative rate. The last plot uses ϵ again, now in comparison with accuracy. As a model can be fair and interpretable, when it is not accurate, it is insufficient. In every graph, a division is made in three groups, full set, G1 and G2. The full set is when all observations are included, G1 the set with the majority group for the sensitive variable and G2 the minority group.



(a)



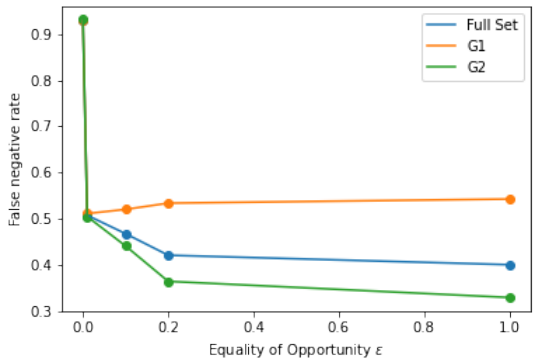
(b)



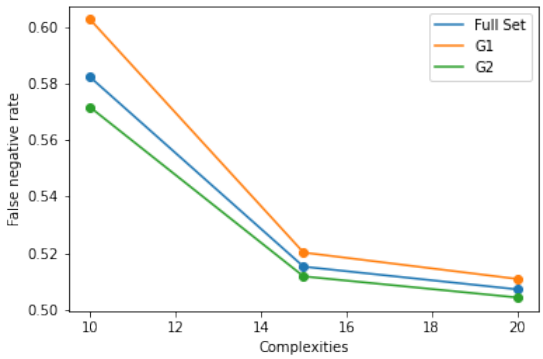
(c)

Figure 3: Different graphs for adult

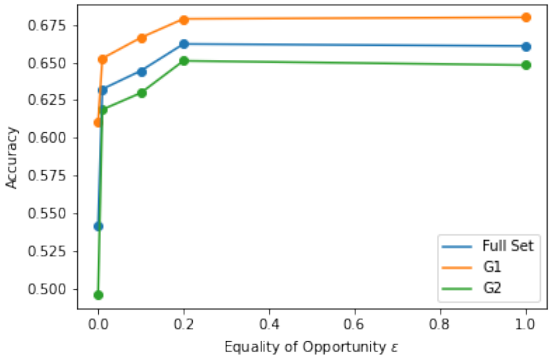
Epsilon being zero is the most strict constraint and assumes perfect fairness. For this dataset, it will be impossible under that constraint to find perfect fairness and no non-empty ruleset can be created. Therefore, all observations will be classified as the negative class. Figure 3a shows this phenomenon in the high ϵ at zero followed by a rapid drop hereafter. When the constraint is relaxed, a ruleset can be formulated and the false negative rate drops. For the complete set and the majority group, a gradual decrease is observed hereafter and remains stable around 0.65. The minority group behaves differently after the drop, a gradual increase can be observed. The difference in behaviour can be attributed to the algorithm classifying the majority group more correctly at the expense of the minority group. It can be observed in figure 3b that the number of complexities does not influence the false negative rate. In the final figure, figure 3c, the accuracy jumps after ϵ is 0 and then has a stable line for all three sets. The jump after epsilon zero has the same explanation as figure 3a.



(a)



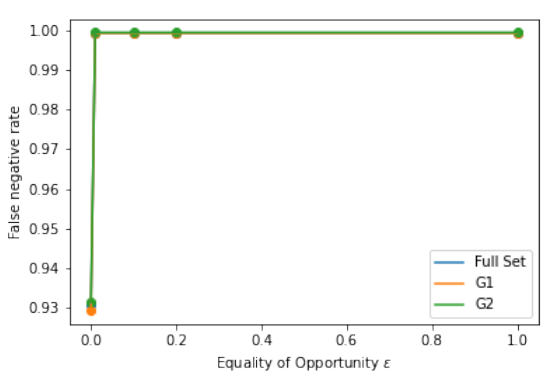
(b)



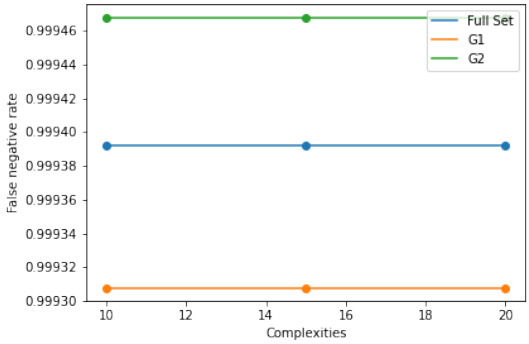
(c)

Figure 4: Different graphs for compas

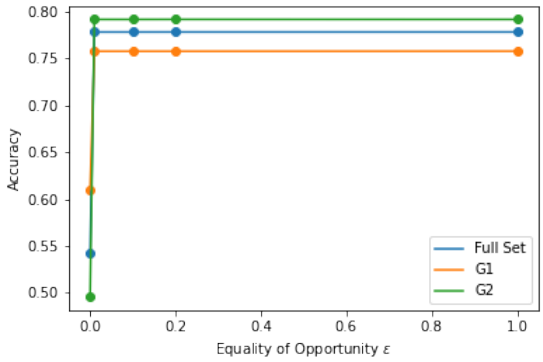
The first figure for the compas dataset exhibits the same shapes as the adult dataset. An increase in complexities has a substantial impact on the false negative rate, shown in figure 4b, which contrasts with the adult dataset. Longer rulesets lead to a lower false negative rate for this dataset. The last graph exhibits the same behaviour as the last figure of the adult dataset, only the jump is more significant after ϵ is zero. The trade-off between accuracy and fairness is visible here in the beginning, as when the fairness is relaxed, accuracy will go up. Important to note is that it stops after ϵ is 0.2.



(a)



(b)



(c)

Figure 5: Different graphs for default

The results for the default dataset are different from the other two datasets. In table 3 it can already be observed that the results do not differ for accuracy and fairness. Computational experiments with different epsilons and complexities have been done, however, did not yield a different outcome. As the resulting rulesets are not non-empty, it is difficult to pinpoint what was the reason behind this apparent optimal solution. The fact that this set was female dominated, instead of male as adult, could have been the reason. This is however unlikely, as the constraint (6) and (7) worked both ways.

A division can be observed between the results of the three datasets. The adult and compas dataset exhibit similar behaviour with each other and with the results of Lawless and Günlük (2021). Especially the graphs move into the same direction in the different results, which is most important as a direct comparison between the numbers is difficult. This is an interesting result, as it was expected in the previous section that the adult and default would be more in line with each other. One reason for this difference in these sets

could be that the male/female ratio was tilted, even though many other factors were the same. It also appears that the larger datasets, adult and default, produce more accurate and fair results in contrast to the compas dataset. Further research is however needed for this claim to hold, as figure 2 already showed that this dataset would be more challenging to predict as the division was not as evident as in the other two datasets.

4.2.2 Equalized Odds

	Adult		**Compas**		**Default**	
	Acc	Fair	Acc	Fair	Acc	Fair
Best accuracy	77.81(1.12)	1.41(0.64)	61.12(4.14)	9.93(4.86)	77.83(0.96)	0.10(0.06)
Best fairness	77.58(0.98)	0.44(0.38)	61.12(4.14)	9.93(4.86)	77.83(0.96)	0.10(0.06)

Table 5: Results Hamming Equalized Odds

In table 5 the results for the Hamming Equalized odds are presented. To make this a useful method, it should perform better than equality of opportunity for either accuracy or fairness. When looking at this table, the first thing that stands out is that there is almost no difference between the two lines, only for the adult set. The same was encountered for the default dataset in the previous section. As the values do not differ for any epsilon or complexity, no graphs will be presented. As they will all look the same and have no explanatory value. First, a look will be shed on what lies behind all these same values for the different epsilons and complexities. Hereafter, a choice will be made between what might then be the better metric to use for these kinds of models.

When the model does not differentiate for different complexities or epsilons, it could imply that the model has reached optimal fairness. The easiest way to reach this would be to give every datapoint the same classification. However, the results present different results for fairness when looking at table 5. Especially for the compas set, the unfairness is still relatively high. Furthermore, as none of the accuracy or fairness results is better than for equality of opportunity, it can be concluded that this metric is too strict and, it is better only to use equality of opportunity.

4.2.3 Heart dataset

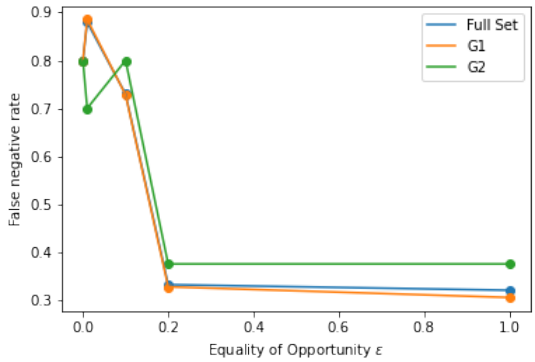
So far three datasets have been used, with two of the same size and one smaller, however still over 5,000 observations. For the third subquestion, a smaller dataset is going to be

used. This section will deal with the impact on the accuracy and fairness of this smaller model. As the Hamming Equalized Odds did not yield better results than the Equality of Opportunity, the latter method will again be used. As the dataset is limited in size, a 5-fold cross-validation is will be used. The results are presented in table 6.

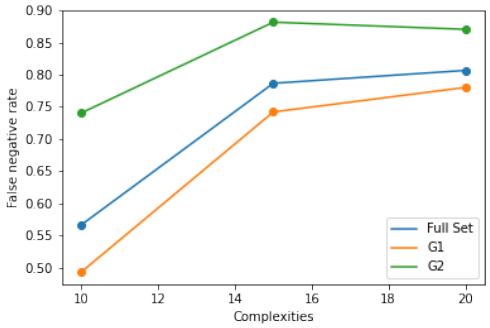
	Heart	
	Accuracy	Fairness
Best accuracy	80.67(9.83)	16.89(25.16)
Best fairness	56.67(6.67)	4.64(6.39)

Table 6: Results Heart

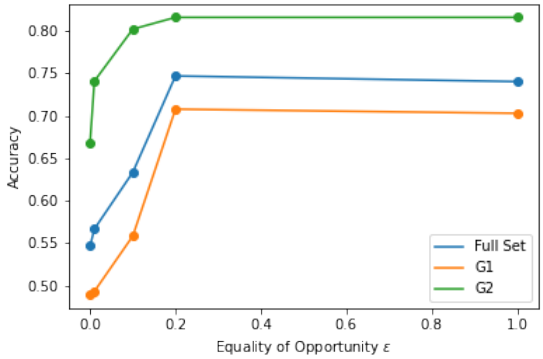
The mean values for accuracy and fairness seem to be not that different from the larger datasets. There is a reasonably high best accuracy combined with a relatively large unfairness and a lower best fairness accuracy combined with a substantially lower unfairness. However, only looking at the best values for accuracy and fairness can give a distorted view of reality. The standard deviations for fairness are telling. These values are larger than the mean values, which indicates that the spread of the different mean values per fold is substantial. The input for the model is insufficient and the model can only improve accuracy by disregarding fairness. In table 8 in appendix A.2 the best epsilons and complexities are presented. A consequence of this disregard of fairness is that for the best accuracy, such a small ϵ produces the best outcome.



(a)



(b)



(c)

Figure 6: Different graphs for heart

The three graphs presented exhibit, to a certain extent, the same behaviour as the compas dataset, which had the same traits in terms of the division in sensitive and explained variables. The second graph of the two datasets only presents contrasting behaviour, as the false negative rate and complexity are positively correlated in this case. In conclusion, it can be said that the model does not perform well for datasets of this size. The different values for the folds were ranging too far apart for the model to produce significant results.

5 Conclusion and Discussion

In this research, an attempt has been made to try and improve DNF rulesets with an increased focus on fairness and interpretability. Research from Lawless and Günlük (2021) introduced a framework, which served as the starting point of this research.

The same datasets have been used to see how the framework worked on a different computational platform. Direct comparison between the absolute results is difficult, as the 10-fold cross-validation introduced a large amount of randomness, almost never leading to the results being the same. Still, t-tests in a comparison between the results have lead to some not significant differences, hinting at a similarity. A better comparison can be made when looking at the graphs of the two result sets. For two of the three datasets, these exhibit substantially similar behaviour to conclude the research by Lawless and Günlük (2021) can be repeated on a different computational platform.Since the current fairness metric applied could be made stricter, it would be expected that fairness of the model can be improved. However, this metric makes significantly difficult for the model to find a solution, neither accuracy nor fairness is improved. Therefore, the findings from this research imply it would be better to keep on using the equality of opportunity metric; instead of equalized odds. Finally, a relatively smaller dataset is used. If the model would also work for this set, the model can be used in rich data environments and in environments where data might be sparse. Unfortunately, the model failed to present stable results. The main difficulty for the model was that it had too little information to produce a correct ruleset.

In conclusion, DNF rulesets can still be regarded as useful to improve the fairness and interpretability of a model. After replicating and trying to extend the work of Lawless and Günlük (2021), it can still be seen as an adequate standard in the field of binary classification. It would be interesting to add more datasets of relatively the same size as the adult, default and compas datasets for further reference. After all sorts of sets have been explored, one should know which epsilons and complexities to use when a set exhibits specific characteristics for both the sensitive and explained variable. Then a roadmap can be produced which would help an individual to create a binary classification model, which is both interpretable and fair.

References

Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias. *ProPublica, May*, 23(2016):139–159, 2016.

Leif H Appelgren. A column generation algorithm for a ship scheduling problem. *Transportation Science*, 3(1):53–68, 1969.

Cynthia Barnhart, Ellis L Johnson, George L Nemhauser, Martin WP Savelsbergh, and Pamela H Vance. Branch-and-price: Column generation for solving huge integer programs. *Operations research*, 46(3):316–329, 1998.

M. S. Bazaraa. *Linear programming and network flows*. Wiley, 2010.

George B Dantzig and Philip Wolfe. Decomposition principle for linear programs. *Operations research*, 8(1):101–111, 1960.

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

Robert Detrano, Andras Janosi, Walter Steinbrunn, Matthias Pfisterer, Johann-Jakob Schmid, Sarbjit Sandhu, Kern H Guppy, Stella Lee, and Victor Froelicher. International application of a new probability algorithm for the diagnosis of coronary artery disease. *The American journal of cardiology*, 64(5):304–310, 1989.

Dheeru Dua and Casey Graff. Uci machine learning repository, 2017. URL `http://archive.ics.uci.edu/ml`.

Daniel James Fuchs. The dangers of human-like bias in machine-learning algorithms. *Missouri S&T's Peer to Peer*, 2(1):1, 2018.

Rohith Gandhi. Support vector machine—introduction to machine learning algorithms. *Towards Data Science*, 2018.

David J Hand and William E Henley. Statistical classification methods in consumer credit scoring: a review. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 160(3):523–541, 1997.

Moritz Hardt, Eric Price, and Nathan Srebro. Equality of opportunity in supervised learning. *CoRR*, abs/1610.02413, 2016. URL `http://arxiv.org/abs/1610.02413`.

Ron Kohavi. Scaling up the accuracy of naive-bayes classifiers: a. In *Proceedings of the Second International Conference on*, 1996.

Roshan Kumari and Saurabh Kr Srivastava. Machine learning: A review on binary classification. *International Journal of Computer Applications*, 160(7), 2017.

Himabindu Lakkaraju, Stephen H Bach, and Jure Leskovec. Interpretable decision sets: A joint framework for description and prediction. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1675–1684, 2016.

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

Conor Lawless. Github repository. 2021. URL `https://github.com/conlaw?tab=repositories`.

Belen Martin-Barragan, Rosa Lillo, and Juan Romo. Interpretable support vector machines for functional data. *European Journal of Operational Research*, 232(1):146–155, 2014.

Elliott Mendelson. *Introduction to mathematical logic*. CRC press, 2009.

David Meyer, Friedrich Leisch, and Kurt Hornik. The support vector machine under test. *Neurocomputing*, 55(1-2):169–186, 2003.

D Michie, D.J. Spiegelhalter, and C.C Taylor. *Machine Learning, Neural and Statistical Classification*. 1994.

Dan Ofer. Compas recidivism racial bias, Jun 2017. URL `https://www.kaggle.com/danofer/compass`.

Dana Peterson and Catherine Mann. Closing the racial inequality gaps. `https://www.citivelocity.com/citigps/closing-the-racial-inequality-gaps/`, Sep 2020.

Bernhard Pfahringer. *Disjunctive Normal Form*, pages 371–372. Springer US, Boston, MA, 2017. ISBN 978-1-4899-7687-1. doi: 10.1007/978-1-4899-7687-1_223. URL `https://doi.org/10.1007/978-1-4899-7687-1_223`.

Ashis Pradhan. Support vector machine-a survey. *International Journal of Emerging Technology and Advanced Engineering*, 2(8):82–85, 2012.

J. Ross Quinlan. Induction of decision trees. *Machine learning*, 1(1):81–106, 1986.

Ronald L Rivest. Learning decision lists. *Machine learning*, 2(3):229–246, 1987.

Guolong Su, Dennis Wei, Kush R Varshney, and Dmitry M Malioutov. Interpretable two-level boolean rule learning for classification. *arXiv preprint arXiv:1511.07361*, 2015.

The Cambridge Dictionary: English Dictionary. *Classification*. Cambridge University Press, 2021.

Robert J Vanderbei. *Linear programming*, volume 3. Springer, 2015.

Dennis Wackerly, William Mendenhall, and Richard L Scheaffer. *Mathematical statistics with applications*. Cengage Learning, 2014.

I-Cheng Yeh and Che-hui Lien. The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. *Expert Systems with Applications*, 36(2):2473–2480, 2009.

A Appendix

A.1 Results Lawless and Günlük (2021)

	Adult		**Compas**		**Default**	
	Acc	Fair	Acc	Fair	Acc	Fair
Best accuracy	82.9(0.2)	9.4(0.4)	68.2(1.2)	24.5(5.3)	82.0(0.6)	0.5(1.2)
Best fairness	78.4(0.4)	0.3(0.3)	53.0(1.6)	0(0)	77.9(0.4)	0(0)

Table 7: Results Equality Of Opportunity

A.2 Tested Epsilons and Complexities

For the first rule generation phase, the same epsilon and complexities as used by Lawless and Günlük (2021) were used. These are:

$$\epsilon = \{0.2, 1\}$$
$$\text{Adult Complexity} = \{60, 80, 110\}$$
$$\text{Compas Complexity} = \{5, 15, 30\}$$
$$\text{Default Complexity} = \{5, 15, 30\}$$
$$\text{Heart Complexity} = \{5, 15, 30\}$$

For the evaluation of the accuracy, the following epsilons and complexities were tested.

$$\epsilon = \{0, 0.01, 0.1, 0.2, 1\}$$
$$\text{Adult Complexity} = \{60, 80, 110\}$$
$$\text{Compas Complexity} = \{10, 15, 20\}$$
$$\text{Default Complexity} = \{10, 15, 20\}$$
$$\text{Heart Complexity} = \{10, 15, 20\}$$

	Best Accuracy		Best Fairness	
Dataset	Epsilon	Complexity	Epsilon	Complexity
Adult	1	60	0.01	60
Compas	1	15	0.01	20
Default	-	-	-	-
Heart	0.01	20	0.01	10

Table 8: Epsilons and complexities chosen

Since the results for default did not differ per epsilon or complexity, there is nothing to report.

A.3 Computer Settings

The experiments were run on a MacBook Pro 2020 (M1 chip, 8 cores, 16 GiB memory). The python environment, version 3.8.5, was configured with anaconda and the package dependencies are presented in the table 9.

Package	Version
Gurobi	9.1.2
Numpy	1.19.2
Pandas	1.1.3

Table 9: Overview of packages

A.4 Python Scripts

The code from the paper of Lawless and Gunluk has been used for different programs to solve the model (Lawless and Günlük, 2021). It was obtained from the Github repository of Conor Lawless (Lawless, 2021). The code together with a short explanation file has been added in a zip file. Four different notebooks have been used, of which a short explanation is given below.

A.4.1 Split into Cells.ipynb

The first step is to binarize the data, which is done in this notebook. It utilizes another program, called *binerizer*. On basis of features of the column it can determine what binerization strategy to use.

A.4.2 Fair CG Rule Generation.ipynb

In this notebook, column generation is used to generate candidate rulesets. Different values for complexity and epsilon can be entered. The output of this notebook will be different text files containing the possible rulesets.

A.4.3 Fair CG Trials.ipynb

The rulesets from the previous program will be used to test the effectiveness of the model. A start will be made with one ruleset, hereafter the column generation is used.

A.4.4 Hamming Loss Experiment.ipynb

In this final program, tests are performed on the final rulesets to find the different measures to evaluate the model.