Student Number: 250945

1. Introduction

In recent years the advancement of AI (artificial intelligence) especially ML (machine learning) has increased automation for tasks in different domains. For example, recommender systems, student admission and job recruitment. With such has risen the concern for bias and discrimination created by these systems [1]. For instance, in the US nonwhites (Black, Hispanic, and Asian) are more likely to be convicted for a similar crime committed by Caucasian people. Moreover, Amazon's job recruitment systems demonstrated bias toward female applicants [2]. These challenges inspired the creation of many tools. Such as AIF360 from IBM and Ethicml from the University of Sussex. To overcome existing fairness and accuracy controversy.

Model name	Accuracy	Fairness
Tree Gini depth 2	0.7862	-0.0026
Tree Gini depth 26	0.8036	0.4648
Garry fair linear regressor c = 1e-05	0.8030	0.0208
Garry fair linear regressor c = 1	0.7298	-0.0031
Garry fair tree classifier c = 1e-05 and depth of 2	0.7871	0.0220
Garry fair tree regressor c = 1e-05 and depth of 26	0.8037	-0.4868

Table1: Result of the best approaches investigated using adult income dataset. Fairness score is equal opportunity difference.

This coursework will focus on understanding the effects of regularization on the accuracy-fairness trade-off. The standard ML method of this research is the decision tree classifier [3] while the fairness model is Garry fair [8].

2. Methodology

2.1. Datasets used for analysis

The main dataset is Adult income with two main sensitive features race and sex and is the largest dataset with 48842-row instances. The second dataset is German credit data (Statlog) with age, sex and foreign worker as sensitive features. Which differ in attribute characteristics such as credit history, and credit amount and is the smallest dataset with 1000 rows. In the adult dataset, 32% are female while 68% are male while the German dataset has 31% female and 69%, male. Lastly, the datasets from AIF360 were preprocessed while sklearn was not [5]. German dataset was selected to understand the effects of the size of the dataset.

2.2. Models analyzed

2.2.1 Machine learning

The main machine learning model used during this coursework was a decision tree classifier. A non-parametric supervised learning algorithm is used for regression and classifications. The objective of a Decision tree classifier is to create a model based on a series of decision rules (if-then-else). Which is used to predict the class for a given target variable by inferring the features of the training data. From this, a structure resembling a tree is created.

2.2.1.1 Understanding decision trees

There are two types of decision trees categorical variable decision which has categorical variables as its predicting target and continuous variable decision which has a continuous variable as the target. Nonetheless, all decision trees stem from the same principle. The root node represents the starting point of the decision tree. Where the main query to fit our target variable is made. After this process, the node goes through the process of splitting which is then subdivided into two sub-notes (more queries). From which the sub-nodes are called decision nodes. Where the branch (new node position) stops producing further nodes is called terminal nodes. While the nodes that continue to subdivide keep going further to create new subtrees. As the tree increases in-depth, the more complex the rules become, and optimal features are selected.

$$\text{Entropy} = -\sum_{i=1}^{n} p_i * \text{Log}_2(p_i)$$

$$\text{Gini Index} = 1 - \sum_{i=1}^{n} (P_i)^2$$

Figure1: Formula of Entropy and Gini Index [7]. Where pi in entropy is the probability of being a function of entropy. While pi in Gini index is the probability of being classified as distinct.

The splitting algorithm used in this coursework implemented by sklearn is based on CART (Classification and Regression Trees) with a Gini index similar to C4.5 [7]. The Gini's impurity index is the main splitting criterion. Where features are not required to be categorical in addition to converting the trained trees to sets of if-then decisions. Pruning or branch removal during selection is performed based on the score of each rule.

Given the nature of the coursework, only Gini and Entropy criteria were used on this coursework. The greater the Entropy the more difficult it becomes to conclude from the data being processed. The higher the Gini index the more diverse our tree partition becomes.

Hyper parameter	Parameter values testes
Pruning strategy	Entropy and Gini
Depth size	None, 2, 26, 51, 100.

Table 2: Parameters tested in regular classifier

2.2.2 Fairness based model

The main fairness model used in this research was the Garry fair classifier [8]. Designed primarily for understanding the notion of rich subgroups. This is an algorithm that bridges the gap between understanding individual and statistical fairness. By picking a fairness constraint such as equalized false-positive rates of a protected group and standardising this constraint to multiple subgroups bounded by a VC (Vapnik and Chervoneskis) dimension. Which are a set of multiple functions (scenarios) essentially learning from a statistical binary classification algorithm. Thus, considering absent fairness constraints.

2.2.2.1 Understanding Garry fair classifier

This is considered an in-processing method which is essentially a classifier built with the goal of fairness classification. In normal classifiers, the protected attributes such as gender (male and female) are for instance uniform and randomly distributed. One can simply take a standard statistical definition of fairness such as equal odds condition and equalize the false-positive rates of the protected groups of the dataset. However, according to [9] this process is not favourable. Since there will be subgroups within the dataset which will not be covered by this process. Moreover, only the 2 groups will be protected and will not account for the intersections created with the protected attributes. The proposed approach covers the problem of gerrymandering in fairness. Instead of asking questions from statistical fairness which only covers two subgroups. It asks questions which will cover all possible combinations. Avoiding overfitting the fairness definition used by the fairness model.

Algorithm 3 FairFictPlay: Fair Fictitious Play

Input: distribution $\mathcal{P}$ over the labelled data points, CSC oracles CSC($\mathcal{H}$) and CSC($\mathcal{G}$) for th classes $\mathcal{H}(S)$ and $\mathcal{G}(S)$ respectively, dual bound C, and number of rounds T

Initialize: set h^0 to be some classifier in $\mathcal{H}$, set λ^0 to be the zero vector. Let $\overline{D}$ and $\overline{\lambda}$ be the poi distributions that put all their mass on h^0 and λ^0 respectively.

For $t = 1,\ldots,T$:

Compute the empirical play distributions:

Let $\overline{D}$ be the uniform distribution over the set of classifiers $\{h^0,\ldots,h^{t-1}\}$

Let $\overline{\lambda} = \frac{\sum_{t'<t}\lambda^{t'}}{t}$ be the auditor's empirical dual vector

Learner best responds: Use the oracle CSC($\mathcal{H}$) to compute $h^t = \text{argmin}_{h\in\mathcal{H}(S)}\langle \text{LC}(\overline{\lambda}), h\rangle$

Auditor best responds: Use the oracle CSC($\mathcal{G}$) to compute $\lambda^t = \text{argmax}_{\lambda}\, \mathbb{E}_{h\sim\overline{D}}[U(h,\lambda)]$

Output: the final empirical distribution $\overline{D}$ over classifiers

Figure2: Algorithm for the Garry fair classifier [8].

Hyper parameter	Parameter values testes
Regularization (C)	1e-05, 0.001, 0.001, 100, 10000
Number of epochs	10 and 20
Regressor strategy	Linear and Tree

Table 3: Parameters tested in fairness classifier

The other method used was preprocessing which focuses on balancing the dataset i.e., removing discrimination before the classifier is learned. The two methods used in this coursework were DI (Disparity impact remover) and RW (Reweighting). While the in-processing methods were Garry fair classifier and Adversarial debiasing. Lastly, the post-processing methods were CEO (Calibrated equalized odds). CEO calibrates the outputs scores of the classifier to find probabilities which change the predicted output in an equalized manner.

3. Results and analysis for adult income

3.1.1 Task 1

In the adult dataset, the standard machine learning method with the highest accuracy score of 0.80 had a branch with a depth of 26 and used entropy as a pruning strategy. However, it presented the lowest equal opportunity difference score of -0.46 which favoured the privileged group (figure 5 in the appendix). The best fairness criterion had a depth of 2 while Gini and entropy presented a similar fairness result of -0.002658 and accuracy of 0.78. After evaluating the testing data, the decision trees with the highest branch had the worse fairness score. While the decision tree with a branch of 2 presented balanced accuracy and fairness. Concluding that generalization does not correspond to fairness on a large dataset.

3.1.2 Task 2

The fairness methods used in this task were DI and CEO. From which the first test focused on understanding the impact of DI. In the adult dataset applying preprocessing while using the decision tree classifier, proved to be beneficial. With the scores now at 0.78 and fairness of 0.0133 and for the adult dataset (Table 4 in the appendix). While 0.69 accuracy and fairness of -0.11 for the German dataset. Moreover, the depth (2 and 26) and the criterion (entropy) remained the same. Moreover, applying post-processing produced improvements in the fairness score with the score from all the different configurations now reaching 0.0. Decision trees with depth less than 10 and entropy criterion lead to a fair model while greater depth size prioritizes accuracy. However, the post scores of the fairness model tend to decrease and be around 0.60 when testing against the predefined testing data. Thus, is best to use the fairness metrics instead of relying on accuracy from the test data. Since the groups now have equal opportunity.

3.1.3 Task 3

FOR (false omission rate) is the proportion of individuals that the models predict as a false negative label while the actual label is positive. This variable focuses on model prediction rather than model outcomes. For example, a family not receiving food subsidies can be labelled false positive or negative. This variable focuses on the ratio of groups not within the false negative. Namely, unprivileged group [10].

In probabilistic terms,

$$P(\text{missed by program} \mid \text{no subsidy, group } i) = C \quad \forall\, i$$

Where C is a constant value. Or, alternatively,

$$\frac{FN_i}{FN_j} = \frac{n_i - k_i}{n_j - k_j} \quad \forall\, i, j$$

Figure 3: example of the calculation of false omission rate using the subsidy analogy [10].

Thus, the proposed metric for this coursework is as follows:

$$ProposedMetric = \left(\frac{FOR(False)}{FOR(False) + FOR(True)} \right)$$

FOR (false) is the non-privileged individuals while FOR (True) is the proportion of individuals categorised as privileged. However, they were assigned to the unprivileged group. Instead of looking at single groups, the goal of the proposed metric is to find the proportion of the different groups affected by the accuracy of our model. Give a score ranging from 0 to 1. Where a score of 0 is given to the best model similar equal opportunity difference. As seen in figure 8 (in the appendix). For the adult dataset models with an accuracy score above 0.80 but a lower efficiency score. Are consistently able to register lower fairness using the metric.

4. Results and analysis for German credit

4.1.1 Task 1

On the smaller German dataset, a depth of 2 on both configurations (Gini and entropy) presented the highest accuracy score of 0.69 and the second fairness of 0.0. The second highest model with an accuracy of 0.68 had None as the depth and -0.21 as the fairness score. After evaluation of the testing data, a score of 0.72 was achieved on the unfair method while 0.723 on the fair model. The lowest equal opportunity score was 0.0 while the score of 0.033 favoured the unprivileged. Thus, generalization does not correspond to fairness on a small dataset.

4.1.2 Task 2

Testing on a smaller dataset (German loan) the accuracy using decision trees dropped by 8 points from 0.78 of the adult datasets using Gini and depth 2 after preprocessing. However, the fairness score improved to 0.0. As seen in table 4 (located in the appendix) this dataset reached this level more times than the adult dataset. Thus, testing the data on a smaller dataset or small proportions of the dataset allows the model to better understand the data both in pre and post-processing.

4.1.3 Task 3

The newly proposed metric when working with the German dataset was at times inconsistent. For instance, due to the size of the German dataset, some scores were attributed to NaN. To overcome this challenge, the values above a certain range or inf had to be scaled or attributed to a score of 0. This was a common factor among fairness metrics [1]. Which in small datasets tend to not produce the best results.

5. Extension results and analysis

5.1.1 Removing sensitive feature

Applying a decision tree classifier on the training folds without sensitive features produced nonbiased fairness scores. This was to be expected as explained by [2]. Moreover, the equal opportunity score stayed in the range of -0.008 and 0.01. While accuracy was the same in all the parameter configurations using both datasets. Staying at 0.7868 for the adult. This leads to conclude that sensitive features are important to the classifier. Removing the sensitive feature or known as unawareness is not the best solution. Certain variables because of the nature of a problem are correlated to the outcome of the solution provided by the ML method. For instance, highly correlated features such as neighbourhood may be a proxy for a sensitive attribute for instance race [4].

5.1.2 Beyond binary features

Working with variables beyond sensitive features enables the study of other variables which may influence the decisions of our model. For instance, in the German dataset using the job categorical variable is a protected attribute. With adversarial debiasing, the fairness metrics were favouring the minority with scores ranging from -0.01 to 0.01. While the accuracy stayed below 0.70 reaching as low as 0.30. Nonetheless, studying the effects of relationship variables on the adult dataset proved to be fruitful. Given the size of the dataset, the accuracy ranged from 0.82 to 0.75 while the fairness score was -0.04 to 0.

The best performing criterion was Adversarial debiasing with C=1000 on both datasets. Which contradicts the norm of favouring lower regularization values. Thus, new metrics are necessary for this type of experiment. Lastly, reweighing as an in-processing method is important when working with non-binary features to achieve plausible results.

5.1.3 Gerry Fair classifier

Compared to the other in processing classifier learned during the term time Adversarial Debiasing. The Gerry fair classifier produces almost similar results when it comes to the regularization parameter (figure 8 in the appendix). For instance, as the C parameter increases our model is less fair and prioritizes balanced accuracy. On the other hand, for small values of C, the model focuses on fairness. Moreover, as the number of epochs increases, this model tends to better understand the data. Furthermore, this classifier can integrate regression algorithms. Such as tree regression which compares the scores of our normal decision tree model. The Gerry fair version optimizes on fairness rather than accuracy [8].

The performances of Garry fair and adversarial debiasing are improved after pre and post-processing techniques are applied. For instance, DI and CEO allow the fairness score to stay below the 0.1 to -0.1 range. Moreover, as the classifier learns more about the data the accuracy of the model tends to stabilize.

6. Conclusion

Decision trees allow us to visualize the features that influence the decisions made by the model and improve fairness after disparate impact preprocessing [6]. In the German dataset credit history from another bank and savings amount (Figure 4 in the appendix). Nonetheless, decision trees are not robust because of their stochastic nature. Fairness-based methods such as Gerry fair are the best alternative to normal decision tree models. However, they are not robust to different data splits and datasets. Moreover, most metrics only favour binary sensitive features [2]. All in all, ML, fairness research and its corresponding metrics are intertwined [1]. Entirely dependent on the number of sensitive features of a given case study.

References

[1] Castelnovo, A., Crupi, R., Greco, G., Regoli, D., Penco, I.G. and Cosentini, A.C. (2022). A clarification of the nuances in the fairness metrics landscape. *Scientific Reports*, 12(1). doi:10.1038/s41598-022-07939-1.

[2] Richardson, B. and Gilbert, J. (2021). A Framework for Fairness: A Systematic Review of Existing Fair AI Solutions. *Journal of Artificial Intelligence Research*, [online] 1. Available at: https://arxiv.org/pdf/2112.05700.pdf.

[3] Kotsiantis, S.B., Zaharakis, I.D. and Pintelas, P.E. (2006). Machine learning: a review of classification and combining techniques. *Artificial Intelligence Review*, [online] 26(3), pp.159–190. doi:10.1007/s10462-007-9052-3.

[4] Zhang, B.H., Lemoine, B. and Mitchell, M. (2018). Mitigating unwanted biases with adversarial learning. [online] New Orleans, LA, USA: Association for Computing Machinery, pp.335–340. doi:10.1145/3278721.3278779.

[5] IBM Research Trusted AI (n.d.). *AI Fairness 360*. [online] aif360.mybluemix.net. Available at: https://aif360.mybluemix.net.

[6] M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian, “Certifying and removing disparate impact.” ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2015.

[7] scikit learn (2009). *1.10. Decision Trees — scikit-learn 0.22 documentation*. [online] Scikit-learn.org. Available at: https://scikit-learn.org/stable/modules/tree.html.

[8] Kearns, M.J., Neel, S., Roth, A. and Zhiwei Steven Wu (2017). Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. *CoRR*, [online] abs/1711.05144. Available at: http://arxiv.org/abs/1711.05144.

[9] Kearns, M.J., Neel, S., Roth, A. and Zhiwei Steven Wu (2018). An empirical study of rich subgroup fairness for machine learning. *CoRR*, [online] abs/1808.08166. Available at: http://arxiv.org/abs/1808.08166.

[10] 저자: Ian Foster, Rayid Ghani, Jarmin, R.S., Frauke Kreuter and Lane, J.I. (2020). *Big data and social science : a practical guide to methods and tools*. 출판사: Boca Raton: Chapman & Hall/Crc.

Appendix

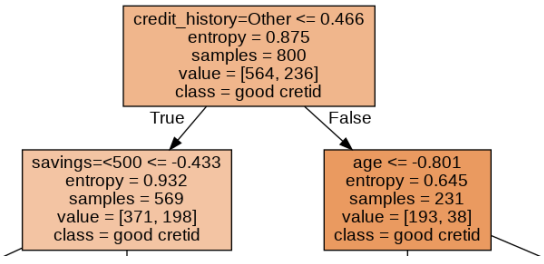

Figure 4: Example of feature selection using decision tree classifier on German loan dataset.

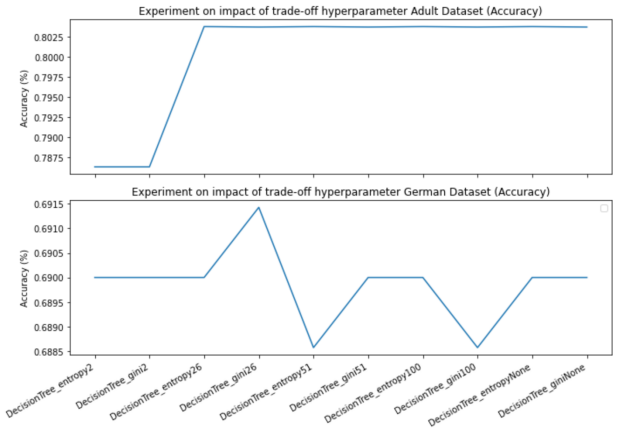

Figure 5: Example of trade off hyper parameter with respect to accuracy score using decision tree classifier.

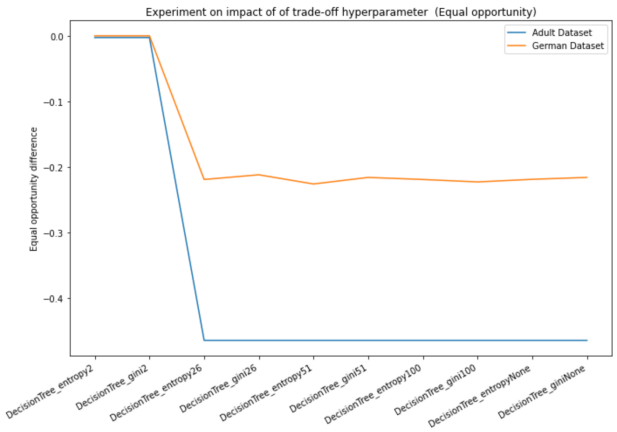

Figure 6: Example of trade off hyper parameter with respect to equal opportunity score using decision tree classifier.

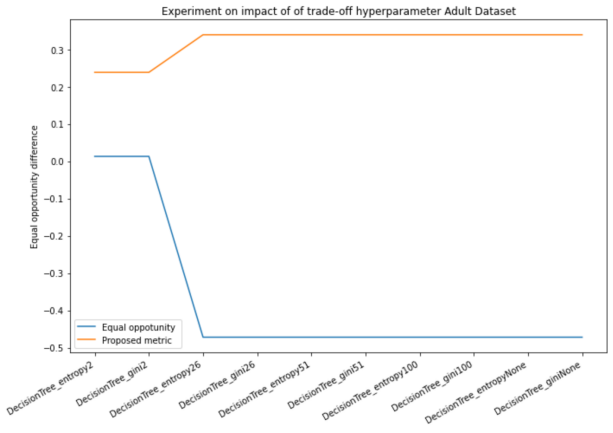

Figure 7: Example of trade off hyper parameter comparing proposed metric against equal opportunity difference.

Model name	Accuracy	Fairness
Tree Gini depth 2 after DI	0.78 and 0.70	0.0133 and 0.0
Tree Gini depth 26 after DI	0.80 and 0.69	-0.4619 and -0.11
Tree Gini depth 100 after DI and CEO	0.80 and 0.72	0.0 and 0.0
Tree Gini depth 2 after DI and CEO	0.78 and 0.71	0.0 and 0.0

Table 4: Result of Adult and German dataset after preprocessing using DI and pre plus post processing using both DI and CEO. The first value is the adult income dataset.

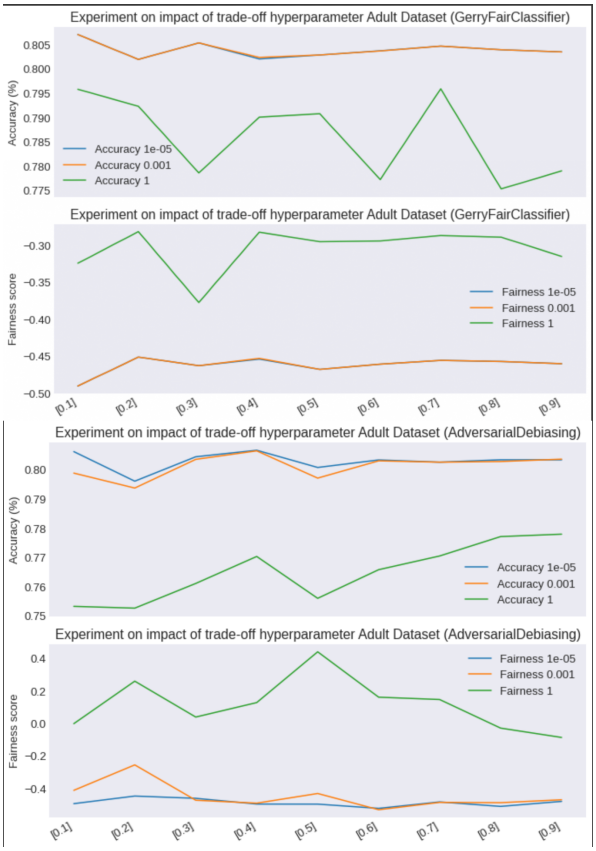

Figure 8: Trade off hyper parameter comparing different train and test splits on Gerry fair and Adversarial debiasing.