**Student Number: 401316**

# 1. Introduction

This report shall analyse the effects of altering trade-off hyperparameters for two chosen machine learning classification algorithms on accuracy [1] and a chosen fairness metric, here being Equal Opportunity Difference [2] or True Positive Rate (TPR) Difference between the privileged and unprivileged groups. The relationship between accuracy and fairness will be discussed, as well as Random Forest and Logistic Regression algorithms compared for two chosen datasets.

Following this, the chosen datasets will undergo reweighing, a pre-processing technique which weights the data samples based on the outcomes for the protected attributes in order to reduce bias [3]. Subsequently, the analysis will be repeated, with comparisons made with the results from the un-reweighed datasets, concluding that reweighting improves fairness, but often at the cost of accuracy.

Finally, using the previous observations, a model selection criterion will be proposed, to optimally account for both accuracy and fairness, due to the trade-off between accuracy and fairness that will be observed before and after reweighing.

Random Forest and Logistic Regression have been chosen as the two classification algorithms to compare here, because of the theoretical and practical pros and cons comparing the two, which will be discussed in the conclusions. Potential further analysis of other machine learning classification algorithms will be outlined in the conclusion, and their applicability to the chosen datasets critiqued.

## 1.1 Method

The chosen datasets for analysis were the pre-processed adult [4] and German [5] datasets imported from the AIF360 library. The machine learning classification algorithms used are Logistic Regression [6] and Random Forest [7], imported from the Sci-Kit Learn library. Custom functions were written to take the chosen dataset and relevant inputs and automate a grid search to find the parameters for a chosen classification algorithm that produce the optimum accuracy and TPR difference. The chosen protected characteristic for both datasets was sex.

For Logistic Regression, there are no hyperparameters that significantly affect accuracy, particularly, however it has been found in previous research that the ‘solver’ and ‘c-values’ can have some marginal impact on results [8]. Subsequently, solvers 'newton-cg', 'lbfgs' and 'liblinear' were tested as well as c-values of 100, 10, 1.0, 0.1 and 0.01, following a logarithmic scale with base 10.

For Random Forest, the chosen hyperparametersfor having the most impact on algorithmic performance, according to Brownlee [8], were ‘n estimators’ i.e. the number of ‘trees’ in the random forest, ‘max depth’ i.e. the number of levels in a tree, and ‘max features’ i.e. the number of features to consider at every split. Values of 10, 50, 100, 500 and 1000 were chosen for ‘n estimators’; 10, 50 and 100 for ‘max depth’; and 'auto' and 'log2' for ‘max features’.

The same parameter choices for the two classification algorithms were used throughout all analysis.

5-fold cross-validation [9] was used, for accuracy utilising ‘GridSearchCV’ [10] and for TPR Difference another custom algorithm was developed, as the AIF360 classification metrics are incompatible with GridSearchCV as it requires scorers that take 'pandas dataframes as inputs. AIF360 has a ‘make scorer’ method that could be used to accomplish this, however this solution was chosen as it preserved additional AIF360 functionality.

The final model selection criterion proposed to account for accuracy and fairness, used a method written to compare multiple possible parameters for multiple classification algorithms. Currently only the two classification algorithms previously mentioned have been included, however the method was written so that the possible algorithms can easily be expanded upon, and how this could be done is clearly defined, but comparing more than two algorithms was deemed beyond the scope of this report. Also, this would be significantly improved with the use of match-case statements instead of if-elif-else statements, introduced in python 3.10, however Jupyter Notebooks do not currently have Python 3.10 compatibility.

# 2. Initial Analysis of the Effect of Varying Trade-off Hyperparameters on Classification

The optimum parameters for accuracy and TPR Difference for Logistic Regression on the adult dataset, were found to be the same parameters for both metrics [Figure 1]. For Random Forest, the difference between accuracy was much smaller than the difference between TPR Difference values for the fairest and most accurate parameter choices [Figure 2]. Clearly, the previous assertion that hyperparameters have little effect on Logistic Regression accuracy and fairness is supported by the results here, and the relatively small variation in accuracy compared to TPR Difference for Random Forest is understandable, as the parameters were almost identical for the most accurate and fairest models.

Similarly, the German dataset parameters for the most accurate and fairest models for Logistic Regression were the same [Figure 3]. Interestingly, the accuracy on the test set for Random Forest [Figure 4] was the same for the distinct parameters that provided the most accurate and fairest fits on the train set. One might assume that the significantly smaller sample size of the German dataset compared to the adult dataset increases likelihood that a chosen classification method may be biased [11]. However, despite the consistently lower accuracy for the German dataset compared to the adult dataset for both Logistic Regression and Random Forest, results showed greater fairness in the German dataset for both algorithms before reweighing. After analysis of the effect of reweighing the datasets is presented, this observation shall be reconsidered.

## 3. Analysing the Effect of Reweighing on Accuracy and TPR Difference

To analyse the effect of the pre-processing technique, 'reweighing' [2] on fairness or TPR Difference, reweighed copies of the adult and German dataset were created using a built-in AIF360 dataset method. Following this, and using the same functions as before, the analysis was repeated for both datasets, providing some contrasting results.

Analysis on the Adult dataset showed that the fairest and most accurate parameters for both Logistic Regression and Random Forest produced a slight decrease in accuracy after reweighing, but a large increase in the TPR Difference; even having a positive TPR Difference for the fairest Logistic Regression model [Figure 5] and for both the most accurate and fairest Random Forest models [Figure 6]. These last two models produced identical accuracy and fairness when tested on the test set, although the near-identical input parameters imply this agreement is not unlikely.

For the German dataset, there was a smaller decrease in accuracy for Logistic Regression [Figure 7] after reweighing than the adult dataset but interestingly, when optimal parameters were tested on the test set after training, the fairness was higher for the previously assumed most accurate model, rather than the fairest. Furthermore, for Random Forest on the German dataset [Figure 8], reweighing did not produce an improvement on fairness for the fairest model, and the accuracy increased for the most accurate model. It must be repeated that the sample size of the German dataset is relatively small, and as such, further analysis might be required using a variety of random states for the cross-validation, to determine if this result is consistent. However, let us consider another possible explanation for these observations. Referring to the previous observation that prior to reweighing, fairness for the German dataset was greater, and noting that the adult dataset consists of census income data, whereas the German dataset consists of credit data, sex should understandably have less of an impact on outcome. Unfair income disparities between genders, 37% less globally for women, remain commonplace [12], and the adult dataset has income as the outcome to predict. Conversely, the German dataset maps to good or bad credit, with credit history as a classifying variable, which is a result of management of spending after income. The average credit score for women is 652 compared to 705 for men, which is approximately a 7.5% decrease. Th ese percentage differences of 37% and 7.5% are notably not wildly out of range of the reported fairness metrics before reweighing, on the adult and German datasets respectively. Further analysis, reweighing the datasets corresponding to income for individual samples could provide insights into whether the differences in pre-reweighting fairness are caused by the fact that income is not a variable withing the German dataset, however many of the variables are linked to income. Unfortunately, this would require data which the datasets do not provide, but this is a topic that should be researched further in this regard, if appropriate datasets can be provided.

## 4. Creating a Model Selection Criterion Accounting for Both Accuracy and Fairness

From the results in the two prior sections, it appears that reweighing decreases accuracy but increases fairness. This should be as expected, as the use of machine learning classification in this instance is with the aim of identifying and removing bias from human decisions in data analysis, specifically in categorising individuals with protected s [14]. Consistently, accuracy in the results was not affected as much by parameters as fairness, with very little difference between the test results for the fairest parameters and the most accurate parameters.

Subsequently, the proposed criterion for model selection from this study, is to find the most accurate model and parameter choice, compute its accuracy, then take a user defined tolerance threshold to choose the fairest model, with accuracy within the threshold. A custom function was written to accomplish this, and the results show that by selecting the correct threshold value, a user can determine whether to prioritise accuracy and fairness at their digression, based on the characteristics of the dataset and the analysis being performed [Figure 9], although the relatively small variations in accuracy for the two pairs of datasets and models chosen for this study ensure that a likely unnecessarily small tolerance value is required to achieve a different result, in this instance. For algorithm and dataset combinations where the most accurate and fairest parameters were identical, the

result has been provided for verification, however no tolerance value changes analysed as clearly there will be no change. Again, analysis of different algorithms on different datasets would be particularly useful in determining the validity of this proposed criterion.

## 5. Conclusion

To summarise, Logistic Regression was shown to be less sensitive to hyperparameter changes, as expected. Moreover, accuracy did not vary a great amount for any model on a particular dataset with parameter changes either. The size of the German dataset might be assumed to increase bias, however a higher fairness for both algorithms before reweighing than the adult dataset implies that the adult dataset must be inherently more biased with regards to sex as a protected characteristic, which supports consistent observations of a persisting global gender pay gap.

Furthermore, reweighing dramatically improved fairness, with a small trade-off for accuracy in the adult dataset, however the results after reweighing for the German dataset were inconsistent, albeit with some improvement in fairness for Logistic Regression, but not Random Forest.

The criterion for considering both accuracy and fairness in model and parameter selection proved to be precise and showed promise, however the scope for use within these datasets, with the chosen algorithms may not make best use of it. Both Logistic Regression and Random Forest had similar accuracies throughout, thus without a very fine tolerance value, the outright fairest model and parameters would be chosen consistently.

The culmination of this report is to conclude that, while the suitability of different algorithms to different classification tasks is well known, there is some scope for data scientists to determine whether accuracy or fairness should be prioritised for a given task. For a task such as reducing gender pay disparity, using data such as the adult dataset analysed in this report, fairness should be prioritised, and even after reweighing, there is only a small decrease in accuracy, compared to a large increase in fairness. Further analysis is recommended into the trade-off between accuracy and fairness for different algorithms in different situations especially.

# Appendix

| | C | penalty | solver | Accuracy | Fairness |
|---|---|---|---|---|---|
| **Most Accurate** | 0.01 | l2 | newton-cg | 0.686667 | -0.016667 |
| **Fairest** | 0.01 | l2 | newton-cg | 0.686667 | -0.016667 |

Figure 1: Most Accurate and Fairest Parameters with Test Accuracy and Fairness for Logistic Regression on the Adult Dataset.

| | bootstrap | max_depth | max_features | min_samples_leaf | min_samples_split | n_estimators | Accuracy | Fairness |
|---|---|---|---|---|---|---|---|---|
| **Most Accurate** | True | 100 | auto | 1 | 2 | 10 | 0.800996 | -0.446872 |
| **Fairest** | True | 10 | auto | 1 | 2 | 10 | 0.800450 | -0.427673 |

Figure 2: Most Accurate and Fairest Parameters with Test Accuracy and Fairness for Random Forest on the Adult Dataset.

| | C | penalty | solver | Accuracy | Fairness |
|---|---|---|---|---|---|
| **Most Accurate** | 0.01 | l2 | newton-cg | 0.686667 | -0.016667 |
| **Fairest** | 0.01 | l2 | newton-cg | 0.686667 | -0.016667 |

Figure 3: Most Accurate and Fairest Parameters with Test Accuracy and Fairness for Logistic Regression on the German Dataset.

| | bootstrap | max_depth | max_features | min_samples_leaf | min_samples_split | n_estimators | Accuracy | Fairness |
|---|---|---|---|---|---|---|---|---|
| **Most Accurate** | True | NaN | log2 | 1 | 2 | 1000 | 0.7 | -0.132653 |
| **Fairest** | True | 10.0 | auto | 1 | 2 | 500 | 0.7 | -0.132653 |

Figure 4: Most Accurate and Fairest Parameters with Test Accuracy and Fairness for Random Forest on the German Dataset.

| | C | penalty | solver | Accuracy | Fairness |
|---|---|---|---|---|---|
| **Most Accurate** | 0.01 | l2 | newton-cg | 0.786597 | -0.036695 |
| **Fairest** | 100.00 | l2 | newton-cg | 0.786597 | 0.004274 |

Figure 5: Most Accurate and Fairest Parameters with Test Accuracy and Fairness for Logistic Regression on the Reweighed Adult Dataset.

| | bootstrap | max_depth | max_features | min_samples_leaf | min_samples_split | n_estimators | Accuracy | Fairness |
|---|---|---|---|---|---|---|---|---|
| **Most Accurate** | True | 50 | log2 | 1 | 2 | 10 | 0.774722 | 0.104279 |
| **Fairest** | True | 50 | auto | 1 | 2 | 10 | 0.774722 | 0.104279 |

Figure 6: Most Accurate and Fairest Parameters with Test Accuracy and Fairness for Random Forest on the Reweighed Adult Dataset.

| | C | penalty | solver | Accuracy | Fairness |
|---|---|---|---|---|---|
| **Most Accurate** | 0.01 | l2 | newton-cg | 0.69 | 0.000000 |
| **Fairest** | 0.01 | l2 | liblinear | 0.68 | -0.022789 |

Figure 7: Most Accurate and Fairest Parameters with Test Accuracy and Fairness for Logistic Regression on the Reweighed German Dataset.

| | bootstrap | max_depth | max_features | min_samples_leaf | min_samples_split | n_estimators | Accuracy | Fairness |
|---|---|---|---|---|---|---|---|---|
| **Most Accurate** | True | 10 | log2 | 1 | 2 | 10 | 0.703333 | -0.139456 |
| **Fairest** | True | 10 | auto | 1 | 2 | 500 | 0.700000 | -0.132653 |

Figure 8: Most Accurate and Fairest Parameters with Test Accuracy and Fairness for Random Forest on the Reweighed German Dataset.

| | model | solver | penalty | C | n_estimators | max_features | max_depth | min_samples_split | min_samples_leaf | bootstrap | accuracy | TPR difference |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Adult 0.1 Tolerance** | RForest | N/A | N/A | N/A | 10 | auto | 10 | 2 | 1 | True | 0.800450 | -0.427673 |
| **Adult 0.001 Tolerance** | LogReg | newton-cg | l2 | 0.01 | N/A | N/A | N/A | N/A | N/A | N/A | 0.802361 | -0.443231 |
| **Reweighted Adult 0.1 Tolerance** | RForest | N/A | N/A | N/A | 10 | auto | 50 | 2 | 1 | True | 0.774722 | 0.104279 |
| **Reweighted Adult 0.001 Tolerance** | LogReg | newton-cg | l2 | 0.01 | N/A | N/A | N/A | N/A | N/A | N/A | 0.786597 | -0.036695 |
| **German 0.1 Tolerance** | LogReg | newton-cg | l2 | 0.01 | N/A | N/A | N/A | N/A | N/A | N/A | 0.686667 | -0.016667 |
| **Reweighted German 0.1 Tolerance** | LogReg | liblinear | l2 | 0.01 | N/A | N/A | N/A | N/A | N/A | N/A | 0.680000 | -0.022789 |
| **Reweighted German 0.01 Tolerance** | LogReg | newton-cg | l2 | 100 | N/A | N/A | N/A | N/A | N/A | N/A | 0.680000 | -0.022789 |

Figure 9: Best Accuracy/Fairness Trade-Off Parameters for Given Datasets with User-Defined Accuracy Tolerances.

# References


(1) MISHRA, A. Metrics to Evaluate your Machine Learning Algorithm. *Towards Data Science* [online]. 2018 [18/05/22]. Available from: https://towardsdatascience.com/metrics-to-evaluate-your-machine-learning-algorithm-f10ba6e38234

(2) AIF360. Classification Metric. *Aif360* [online]. 2021 [18/05/2022]. Available at: https://aif360.readthedocs.io/en/latest/modules/generated/aif360.metrics.ClassificationMetric.html#aif360.metrics.ClassificationMetric.true_positive_rate_difference

(3) KAMIRAN, F and CALDERS, T. Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems*, 33(1):1–33, 2012

(4) AIF360. Adult Dataset. *Aif360* [online]. 2021 [18/05/22]. Available at: https://aif360.readthedocs.io/en/stable/modules/generated/aif360.datasets.AdultDataset.html#aif360.datasets.AdultDataset

(5) AIF360. German Dataset. *Aif360* [online]. 2021 [18/05/22]. Available at: https://aif360.readthedocs.io/en/stable/modules/generated/aif360.datasets.GermanDataset.html#aif360.datasets.GermanDataset

(6) SCIKIT LEARN. Logistic Regression. *Scikit learn* [online]. 2022 [18/05/2022]. Available at: https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html

(7) SCIKIT LEARN. Random Forest Classifier. *Scikit learn* [online]. 2022 [18/05/2022]. Available at: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

(8) BROWNLEE, J. Tune Hyperparameters for Classification Machine Learning Algorithms. *Machine Learning Mastery* [online]. 2020 [18/05/2022]. Available at: https://machinelearningmastery.com/hyperparameters-for-classification-machine-learning-algorithms/

(9) BROWNLEE, J. A Gentle Introduction to k-fold Cross-Validation. *Machine Learning Mastery* [online]. 2020 [18/05/2022]. Available at: https://machinelearningmastery.com/k-fold-cross-validation/

(10) SCIKIT LEARN. Grid Search CV. *Scikit learn* [online]. 2022 [18/05/2022]. Available at: https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html

(11) VABALAS, A ET AL. Machine learning algorithm validation with a limited sample size. *PLOS ONE,* 14(11): e0224365

(12) MASTERSON, V. 6 surprising facts about the global gender pay gap. *World Economic Forum* [online]. 2022 [18/05/2022]. Available at: https://www.weforum.org/agenda/2022/03/6-surprising-facts-gender-pay-gap/

(13) CREDIT KARMA. 'Mind the gap. The gender credit one that is.'. *credit karma* [online]. 2021 [18/05/2022]. Available at: https://www.creditkarma.co.uk/insights/i/gender-credit-gap/

(14) MAHMOUDIAN, H. Reweighing the Adult Dataset to Make it "Discrimination-Free". *Towards Data Science* [18/05/2022]. Available at: https://towardsdatascience.com/reweighing-the-adult-dataset-to-make-it-discrimination-free-44668c9379e8