Student Number: 237778

1. Introduction

Data driven innovations (DDI), especially in the advent of Artificial Narrow Intelligence (ANI), have become increasingly popular. They are present in all areas of life and integral to our economies [1]. However, algorithmic biases within DDI processes are not fully understood [1]. There are three sources of algorithmic bias: societal bias, data bias and method bias [2].

Societal biases are present in arguably all instiutions including healthcare, education, finance and employment. For example, studies have shown that clinicians hold racial biases in relation to pain perception in black patients and this is reflected in treatment [3]. Such biases influence data, analyses and resulting data products. Data bias can take the form of emerging definitions including: historical bias, represenatation bias, measurement bias, population bias and more [4]. Datasets biased in these ways can produce prejudicial algorithms as seen in a Twitter's consistent cropping of black people's faces out of pictures [5]. Lastly, method bias is a result of failure of rigourisity in model/algorithm production inluding confirmation bias and correlation fallacy [6]. Algorithmic bias must be addressed to minimise adversarial impacts of DDIs and ensure they do not extend the impact discriminatory, unfair or unjust components of societies.

Fairness possesses multiple definitions but there are two competing schools of thought. We're All Equal (WAE) and What You See Is What You Get (WYSIWYG) [7]. WYSIWYG assumes data reflects reality whilst WAE holds the position that bias may be present in multiple ways. WAE is upheld in this report, thus, it follows that the probability positive outcomes for privileged and unprivileged groups is equal in a fair model.

$$\Pr\left(\hat{y}_{D=unprivileged} = 1\right) = \Pr\left(\hat{y}_{D=\,privileged} = 1\right)$$

However, equalising positive outcomes does not account for misclassifications in relation to the privileged and unprivileged groups. Misclassifications should be minimised to prevent unwitting discrimination. Thus, the definitions: disparate mistreatment and statistical parity (or demographic parity) are central to this report.

It is important to note that identities do not exist in isolation. For example, a non-white LGBT woman can be considered to be a member of multiple unprivileged groups. Intersectionality of identities is a core consideration.

This report explores regularisation-fairness trade-off, pre-processing methods which aim to minimise the influence of bias within data and aims to present a model selection strategy which considers both accuracy and fairness.

2. Method

2.1. Model Selection

Most Accurate

Accuracy is measured primarily with the accuracy metric and supplemented with the F1 score; measurement of true predictions is prioritised. F1 score is the harmonic mean of recall and precision. It enables the assessment of a model's ability to minimise misclassifications.

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

$$F1 = \frac{2(RecallxPrecision)}{Recall + Precision}$$

As seen in the formula above, a recall or precision of 0 will result in an f1 score of 0. The absence of false positives (FP) or negatives (FN) yields a score of 1. Accuracy score, on the other hand, describes the fraction of true positive (TP) and true negative (TN) predictions.

$$Accuracy(score) = \frac{TP + TN}{TP + TN + FP + FN}$$

Most Fair (non-binary sensitive features)

In this report, fairness of models is primarily measured using the disparate mistreatment definition and supplemented with the statistical parity difference (SPD) metric. However, bias within data is measured with the SPD only. Metrics will measure against multiple sensitive features, e.g., age and sex, as opposed to binary features.

Statistical parity requires equality in the probability of positive outcome prediction for both privileged and unprivileged groups [4]. Whist disparate impact is similar, they are calculated differently.

$$Statistical\ parity\ difference\ (SPD) \\ = \Pr\left(\hat{y}_{D=unprivileged} = 1\right) - \Pr\left(\hat{y}_{D=privileged} = 1\right)$$

By this definition models are fair if SPD is close to 0. However, this can be achieved even with misclassifications in relation to the groups. This represents an avenue for unfairness.

Disparate mistreatment requires misclassification for both the privileged and unprivileged group to be equal [4]. The chosen metric for measuring disparate mistreatment is error rate ratio.

$$Error\ Rate = \frac{FP + FN}{Negative(N) + Positive(P)}$$

$$Error\ Rate\ Ratio\ (ERR) = \frac{ERR_{D=unprivileged}}{ERR_{D=privileged}}$$

2.2.Support Vector Machine (SVM)

SVM is a supervised machine learning algorithm used to classify data by maximising the margin separating datapoints on a hyperplane. The regularisation parameter “C” is the argument of focus.

The value of C is used by the function determine the degree to which the model avoids misclassifying datapoints. Higher values correspond with lower tolerance for misclassification. In this investigation, the value of C will be varied around the default of 1.

SVM was selected because, unlike a probabilistic methods like logistic regression, it is not prone to outliers. The selection of SVM instead of multilayer perceptrons (MLP) is due to the ease of implementation of SVM especially with the need for 5-fold validation using the training data. Multilayer perceptions also possess a number of hyperparameters such as number of layers, neurons per layer, iterations and learning rate which must be optimised whilst SVM has fewer (gamma, C). SVM is effective in high-dimensional spaces and this, with the kernel technique, also makes SVM a suitable choice as it enables the production of non-linear solutions when required. The greatest weakness of SVM is the training time required for large datasets. The average time taken to complete validation using a single fold in the train splits of Adult (34189 rows) and German (700 rows) datasets is 50.18 and 0.02 seconds respectively.

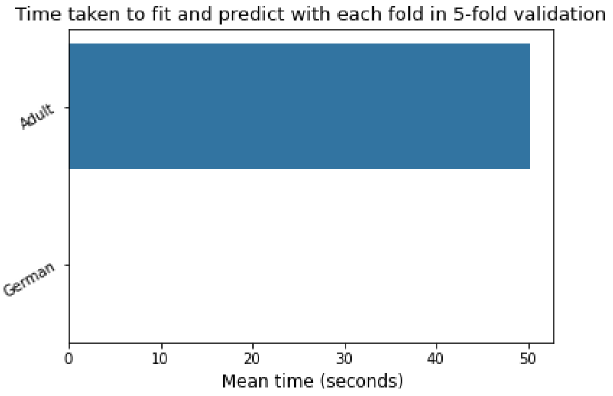

Figure 1 Comparison of time taken to complete validation with each dataset.

2.3.Datasets

The pre-processed “Adult” and “German” datasets available through the AIF360 library were used in this report. The Adult dataset contains 48842 records and 18 features. This dataset can be used to separate Adults into two classes; above 50-thousand and less than or equal to 50-thousand. The German dataset contains 11 features and 1000 records labelled as ‘good credit’ or ‘bad credit’. In Adult, sex and race are the protected attributes.

In German, sex and age. These protected attributes are linked to privileged and unprivileged groups within the data. The privileged groups in the Adult dataset are those who are white and male whilst the privileged groups in the German dataset are those who are above 25 years of age and male. In an attempt to complete holistic analysis, **both attributes within each dataset are utilised and considered simultaneously, not separately**.

2.4.Optimised pre-processing

Bias mitigation techniques may be implemented on the input data, the (classification) algorithm or the calculated outcome. They are known as pre-processing, in-processing and post-processing techniques. Optimised pre-processing increases fairness by minimising the dependence of classification models on protected attributes (disparate treatment) whilst limiting distortion of the data and maintaining utility through after randomised transformation of the labels and features of the input data [8]. Therefore, with these three considerations, the optimisation problem for the randomised transformation $P_{\hat{X},\hat{Y}|X,Y,D}$ of each sample (D_i, X_i, Y_i) to $(\widehat{X}_\iota, \widehat{Y}_\iota)$ according to [8] is:

$$min_{P_{\hat{X},\hat{Y}|X,Y,D}}\ \Delta\left(P_{\hat{X},\hat{Y}}, P_{X,Y}\right)$$

$$s.t.\ \ J\left(p_{\hat{Y}|D}(y|d), p_{Y_T}(y)\right) \leq \epsilon_{y,d}\ \ and$$

$$\mathbb{E}\left[\delta\left((x,y),(\hat{X},\hat{Y})\right)\middle| D = d, X = x, Y = y\right] \leq C_{d,x,y}\ \forall\ (d,x,y) \in \mathcal{D}, \mathcal{X}, \mathcal{Y},$$

$$P_{\hat{X},\hat{Y}|X,Y,D}\ is\ a\ valid\ distribution.$$

Where sensitive features are D and the distance function J(,) used in discrimination control is the probability ratio measure.

$$J(p,q) = \left|\frac{p}{q} - 1\right|$$

Utility preservation meets the requirement statistical closeness between the distributions of $(\widehat{X}_\iota, \widehat{Y}_\iota)$ and (X_i, Y_i) and is expressed as:

$$min_{P_{\hat{X},\hat{Y}|X,Y,D}}\ \Delta\left(P_{\hat{X},\hat{Y}}, P_{X,Y}\right)$$

and distortion control is:

$$\mathbb{E}\left[\delta\left((x,y),(\hat{X},\hat{Y})\right)\middle|D=d,X=x,Y=y\right] \le C_{d,x,y}\ \forall\ (d,x,y)\in\mathcal{D},\mathcal{X},\mathcal{Y},$$

After application of optimised pre-processing, the SPD within the train split of the Adult and German datasets increased (favourably) from -0.243 and -0.251 to -0.048 (80%) and -0.044 (82%) respectively.

Unlike Reweighing, Optimised pre-processing enables explicit control of individual fairness [8]. It also enables **optimisation for non-binary protected attributes** unlike methods such as Disparity Impact Remover.

3. Models

3.1.Most Accurate

Following 5-fold validation, the most accurate model for both datasets utilise a C value of 1. Adult (accuracy 0.804, F1 0.489), German (accuracy 0.71, F1 0.821). Thus, the chosen SVM model uses a c value of 1 with resulting cores of Adult (accuracy 0.804, F1 0.473, ERR 0274, SPD -0.215) and German (accuracy 0.693, F1 0.815, ERR 0.662, SPD -0.318).

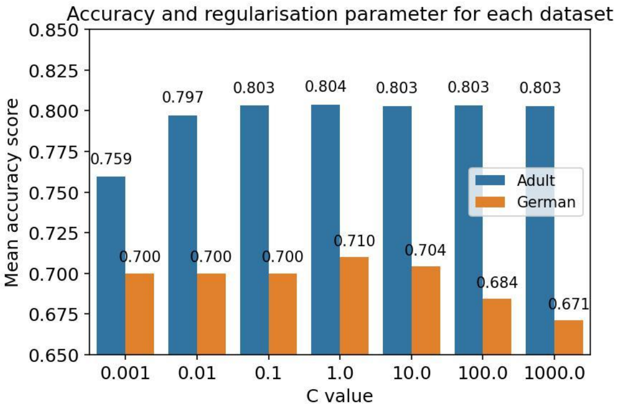

Figure 2 Mean accuracy scores after 5-fold validation for each dataset at each C value.

3.2.Most Fair

Using mean values after 5-fold validation, the most-fair models for German dataset utilises a C value of 1000 (ERR 1.373, SPD – 0.462). The most-fair Adult SVM model (ERR 0.303, SPD -0.237) uses C of 100 closely followed by 1000 (ERR 0.302, -0.242).

The chosen most-fair models, based on aggregated (both datasets) ERR (figure 3), is that which uses C of 1000 as it produces the lowest mean error rate ratio (figure below). The SVM models produced return Adult (accuracy 0.804, F1 0.479, ERR 0.275, SPD -0.22), which is unfair, and German (accuracy 0.687, F1 0.808, ERR 0.912, SPD -0.396).

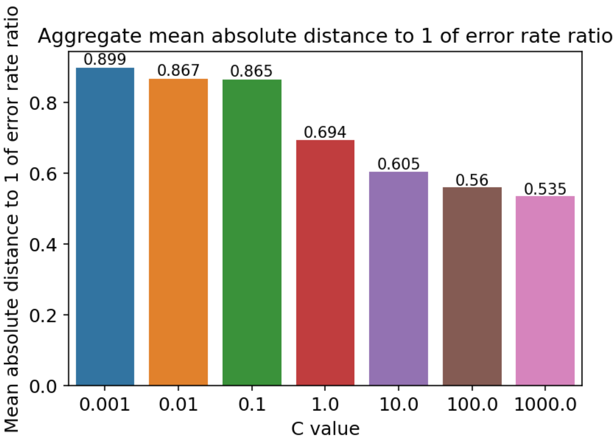

Figure 3 Mean absolute distance to 1 of the ERR of each model at each parameter. Aggregate (mean) of both datasets.

3.3.Optimised pre-processing (Accuracy)

The model with the highest mean accuracy after transformation for both Adult and German utilises C value of 1. The chosen model for accuracy, C = 1, yields accuracy score of Adult (accuracy 0.78, F1 0.501, ERR 0.962, SPD 0.015) and German (accuracy 0.697, F1 0.821, ERR 0.662, SPD 0.015).

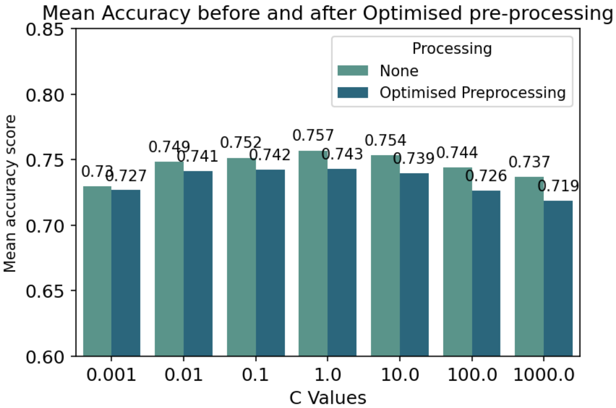

Figure 4 Mean accuracy score across both datasets with and without optimised pre-processing obtained from 5-fold validation at the tested c values.

3.4.Optimised pre-processing (Fairness)

The most-fair models after 5-fold validation utilise C values of 0.1 (ERR 0.011, SPD -0.124) and 1000 (German, ERR 0.004, SPD -0.344). The best model from aggregated metrics utilises C value of 1000 because the mean ERR for both datasets is 0.019. The chosen model (C=1000) yields Adult (accuracy 0.781, F1 0.502, ERR 0.986, -0.163) and German (accuracy 0.687, F1 0.811, ERR 0.782, SPD -0.105).

4. Model Selection Strategy and Discussion

Classical regularisation-accuracy trade-off is well understood. As such, an optimal C value of 1 for accuracy is expected. However, greater fairness (ERR) is achieved at higher values of C. An explanation is that lower tolerance for misclassification also results in lower errors in relation to the protected attributes.

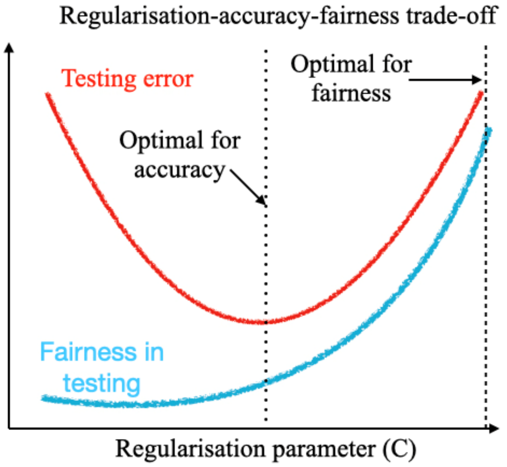

Figure 5 The regularisation-accuracy-fairness trade-off based on results and observations within this report.

It is important to note that selected models at each stage would differ with chosen metrics. For instance, models utilising C of 0.001 (Adult), 0.001 (German) and 0.01 (German) with transformed datasets achieved SPD of 0, not the chosen C of 1000 (table 1). Also, measured SPD values indicate bias in favour of the privileged group before and after pre-processing whilst ERR converges towards 1 after the transformation (figure 6). However, although F1 score and SPD are not used for model selection, they contextualise results.

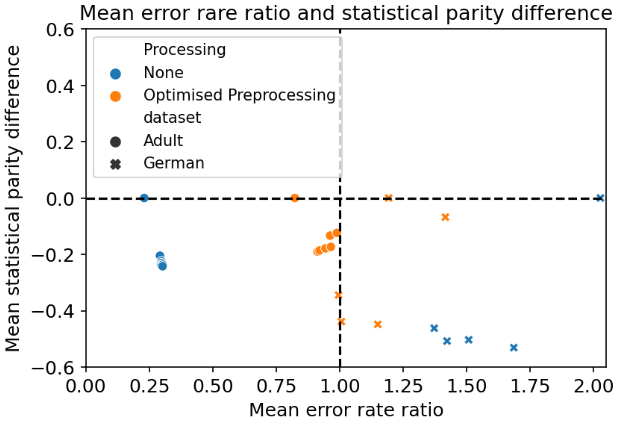

Figure 6 Mean ERR and SPD obtained using multiple values of C from 5-fold validation of both datasets with and without optimised pre-processing.

Optimised pre-processing demonstrably reduces bias in input data but it also reduces utility of the data (figure 4).

Using the results, the following strategy for model selection was devised. It is important to note that this strategy aims to minimise unfairness in early stages and optimise for accuracy in latter stages. Pre-processing is not advocated if bias is not measured in the data. Unrequired transformation will impact utility (figure 4).

Thresholds
Determine fairness, metrics and acceptable range(s) of input data bias with justification. Define accuracy and acceptable score(s).

Input data
Measure unfairness within input data. If unfairness exceeds defined range, apply Optimised pre-processing else, proceed. Optimised pre-processing enables optimisation with non-binary protected attributes and also provides explicit control for fine tuning.

Optimisation
Optimise most accurate and most fair models for accuracy using other available parameters. Also, adjust the key regularisation parameter (C) within the optimal accuracy and optimal fairness range (figure 6).

Model selection
Discard models with metrics outside acceptable ranges and select models with the most favourable metrics for accuracy and fairness.

If a defined maximum threshold of SPD +/- 0.2 in train splits is utilised following the strategy, the transformed datasets will be required as SPD of -0.243(Adult) and -0.251 (German) is measured the input. The previously selected models are then optimised for accuracy using the RandomisedSearchCV function available in the ScikitLearn library. This function is used to complete 5-fold validation using the SVM model only, in this instance, to find the optimal gamma parameter to complement the chosen C values. Accepted models exceed 0.7 in accuracy score and ERR = 1 +/- 0.2.

Chosen models
The German dataset's most fair model was produced with C of 1000 as expected, however, it fails to meet the accuracy threshold and would therefore not be deployed. The most accurate and most accurate and most fair models were produced using the adult data set at C of 1 and 1000 with identical metrics after both Optimised pre-processing and RandomSearchCV.
Most accurate and most fair– Adult dataset, C = 1 and 1000, gamma = 1, accuracy = 0.78, F1 = 0.5, ERR = 0.986, SPD = -0.171.

References

[1] S. Akter, G. McCarthy, S. Sajib, K. Michael, Y. K. Dwivedi, J. D'Ambra and K. N. Shen, “Algorithmic bias in data-driven innovation in the age of AI,” *International Journal of Information Management,* vol. 60, no. 102387, 21 October 2021.

[2] L. Floridi and M. Taddeo, “What is data ethics?,” *Philisophical transactions of the Royal Society. Mathemathica, Physical and Engineering Sciences.,* vol. 374, no. 2083, 2016.

[3] K. Hoffman, S. Trawalter, J. Axt and M. N. Oliver, “Racial bias in pain assessment and treatment recommendations, and false beliefs about biological differences between blacks and whites,” *Proceedings of the National Academy of Sciencesof the United States Of America,* vol. 16, no. 113, 2016.

[4] C. Haas and A. Ashokan, “Fairness metrics and bias mitigation strategies for rating predictions,” *Information Processing & Management,* vol. 58, no. 5, 2021.

[5] A. Hern, “Twitter apologises for 'racist' image-cropping algorithm,” Guardian, 21 September 2020. [Online]. Available: https://www.theguardian.com/technology/2020/sep/21/twitter-apologises-for-racist-image-cropping-algorithm. [Accessed 30 April 2022].

[6] A. Tsamados, N. Aggarwal, J. Cowls, J. Morley, H. Roberts, M. Taddeo and L. Floridi, “The ethics of algorithms: key problems and solutions,” *AI & Society,* vol. 37, pp. 215 - 230, 2021.

[7] S. Yeom and M. C. Tschantz, “Avoiding Disparity Amplification under Different Worldviews,” in *ACM Conferences on Fairness, Acountability and Transparency,* 2021.

[8] M. Belkin, D. Hsu, S. Ma and S. Mandal, “Reconciling modern machine-learning practice and the classical bias–variance trade-off,” *Proceedings of the National Academy of Sciences (PNAS),* vol. 116, no. 32, pp. 15849-15854, 2019.

[9] F. P. Calmon, D. Wei, K. N. Ramamurthy and K. R. Varshney, “Optimized Data Pre-Processing for Discrimination Prevention,” *Data Science Department, IBM Thomas J. Watson Research Center,* 11 April 2017.

Appendix

C	Dataset	Processing	Abs(ERR - 1)	ERR	SPD
1000	German	Optimised Pre-processing	0.004	0.996	-0.344
10	German	Optimised Pre-processing	0.007	1.007	-0.439
0.1	Adult	Optimised Pre-processing	0.011	0.989	-0.124
1000	Adult	Optimised Pre-processing	0.035	0.965	-0.173
0.01	Adult	Optimised Pre-processing	0.037	0.963	-0.133
100	Adult	Optimised Pre-processing	0.056	0.944	-0.179
10	Adult	Optimised Pre-processing	0.079	0.921	-0.186
1	Adult	Optimised Pre-processing	0.087	0.913	-0.19
100	German	Optimised Pre-processing	0.15	1.15	-0.448
0.001	Adult	Optimised Pre-processing	0.176	0.824	0
0.1	German	Optimised Pre-processing	0.193	1.193	0
0.01	German	Optimised Pre-processing	0.193	1.193	0
0.001	German	Optimised Pre-processing	0.193	1.193	0
1000	German	None	0.373	1.373	-0.462
1	German	Optimised Pre-processing	0.416	1.416	-0.068
100	German	None	0.423	1.423	-0.508
10	German	None	0.508	1.508	-0.503
1	German	None	0.686	1.686	-0.531
100	Adult	None	0.697	0.303	-0.237
1000	Adult	None	0.698	0.302	-0.242
1	Adult	None	0.702	0.298	-0.23
0.1	Adult	None	0.702	0.298	-0.224
10	Adult	None	0.702	0.298	-0.237
0.01	Adult	None	0.708	0.292	-0.205
0.001	Adult	None	0.77	0.23	0
0.1	German	None	1.027	2.027	0
0.01	German	None	1.027	2.027	0
0.001	German	None	1.027	2.027	0

Table 1 Fairness metric results for both datasets with and without optimised pre-processing. This is arranged in ascending order of abs(ERR-1) which is the absolute distance to one of the error rate ratio (ERR). SPD is statistical parity difference.