Student Number: 2444849

1. Introduction

Machine Learning (ML) and Artificial Intelligence (AI) have attracted a lot of attention in recent years; researchers, big tech companies and business owners are embeddings those technologies into their infrastructure. ML has a range of topics. One of the most widely researched topics, which is also considered relatively new, is the bias and fairness of ML systems and the ethical issues that may arise with the increased adoption of these systems. This field of ML introduces methodologies and evaluation metrics which minimise the biases in the data and the inaccuracies accompanied by the ML models. This field is essential as it aims to provide models with minimal biases to sensitive features in the datasets by learning fair representations which can be generalised to unseen data. Sensitive features include race, ethnicity, gender, age, disabilities, and other protected attributes.

This report will cover a comparison between using standard and fairness-based ML models, where both models will be trained and tested using two datasets from the UCI website, the adult dataset and the German credit data, which both contain sensitive features. The models will be evaluated for accuracy and fairness using scikit-learn (sklearn) and aif360 libraries. The main aim here is to study the feasibility of using discrimination aware classifiers using fairness techniques and their ability to minimise bias or discrimination compared to standard ML models.

2. Motivation

The need to develop and research discrimination-free ML algorithms is necessary for future applications, where ML and AI would replace various tasks currently done by humans. In fact, nowadays, many applications use ML for decision-making tasks, such as job requirements, and courts in the USA use the COMPAS algorithm for recidivism prediction, other substantial tech companies use recommender systems to show preferences for users, and many other applications. Developing AI and ML systems can benefit humans. Still, it is crucial that these systems are trained on unbiased data, as humans themselves are biased, or even processed with techniques that ensure zero or minimal discrimination, as having biased AI systems could lead to fatal outcomes, like the one investigated in 2016 which showed that COMPAS produced much higher false positive rate for black people compared to white people [5].

3. Models

3.1. Standard Logistic Regression (LR)

The classifier used with both datasets with the LR classifier from the sklearn library. The LR is one of the most widely used classifiers for supervised ML classification tasks, and class probability estimations, the methodology behind this classifier was formulated by David Cox back in 1958; this classifier takes in a linear combination of the attributes and then a nonlinear sigmoidal function is applied to them. For the particular classification task in this report, where the classification task is binary (either zero or one), the LR model tries to find a hyperplane which separates the outputs into two classes based on a loss function which can be optimised using backpropagation. However, the classifier can be used for multiclassification tasks and then it is called a multinomial logistic regression. The underlying architecture of this model is based on probabilistic mathematics, specifically on the concept of odds of an event, which can be defined as the probability of that event occurring divided by the probability of it not occurring [1,2]. Sklearn provides the user with the option to vary the hyper-parameters of all models to find the best combination that yields the optimal hyperplane and evaluation metrics, and with the aid of the sklearn built-in cross-validation (cv) function, which allows the user to split the training data into training and testing folds, where the model will be trained on n-1 folds and tested on the last one, here n=5 for the cv split were used. The varied hyperparameters will be explained in later sections.

3.2. LR with Reweighing

Reweighing is a data preprocessing technique which was first proposed in [3] and tries to minimise the possible discriminations in the outputs of a classifier. Reweighing is done by using the aif360 library. First, the privileged and unprivileged classes are defined, and then these classes are assigned weights along with the labels class; therefore, the dataset can be discrimination free with respect to the privileged and unprivileged groups assigned to the reweighing function to reimburse for any existing biases in the dataset lower weights will be assigned to inputs that have been deprived or favoured as explained in [4], where the goal is to obtain discrimination equal to 0. After the weights are fitted to the training data, they can be fitted to the standard LR model, which produces a discrimination-free classifier. Both the standard LR model and the discrimination-free LR will be evaluated for accuracy, which is equal to the correctly classified points divided by the incorrectly classified points, and discrimination or fairness using the equality of opportunity metric, which states that the classification of each data point should be based on specific qualifications but not sensitive attributes, so for example, a job hiring ML algorithm should have an equal acceptance rate for the qualified people from both groups say males and females.

3.3. LR Hyper-Parameters

To understand how the best accuracy and fairness are picked, the intuition behind tweaking the hyperparameters of the LR model needs to be explained. Three hyperparameters were tuned C, penalty, and the solver.

- C: it is the inverse of the regularisation strength; regularisation is used to minimise a model's overfitting to the training data by constraining the model's coefficients, resulting in a model which can generalise better to unseen data, like the held-out testing data, if C is a smaller number then regularisation increases, and vice-versa if C is a larger number, keeping in mind the default values in sklearn is 1.0. The range of C values varied across the folds was (0.001,0.005,0.01,0.1,1,10,20,100,200,1000) for the adult data set, and (0.001,0.1,1,1.5,3,5,10,20) for the german dataset.
- Penalty: The type of regularisation used, L1 and L2 penalties were used.
- Solver: It is the algorithm used to solve the optimisation problem of calculating the coefficients, so given a set of attributes and a target, the solver tries to find what coefficients should be used by the model here; saga and liblinear solvers were tested across the five folds, sklearn offers more options of solvers.
- Sample weight: the weights calculated by the reweighing method explained, assigning weights that ensure minimal discrimination for the sensitive class for tasks 2 and 3 to compare them to the standard LR classifier in task 1.

The combination of the hyper-parameters used to obtain the best accuracy and fairness values will be shown in the results section; more information about the hyper-parameter tuning can be found in [8].

4. Datasets

4.1. Adult Dataset (Census Income Dataset)

The Adult dataset was extracted by Barry Becker from the 1994 Census database and can be found on the UCI website. This dataset consists of 48842 instances and 14 attributes. The intention of this dataset is to use a supervised ML classification approach to be able to predict whether a person makes less or more than 50k a year. The aif360 library offers a preprocessed version of the adult dataset, with the protected attribute being 'sex'; the privileged individuals are men, while the unprivileged are women, both are given labels of 1 and 0, respectively, and the data was split into training (0.7) and testing (0.3). more information on this dataset can be found here [6]

4.2. German Dataset

The German dataset was compiled by Professor Dr. Hans Hofmann, which can also be found on the UCI website. This dataset is made up of 1000 instances and 20 attributes; again, aif360 offers a preprocessed version of the dataset with the protected attribute also being 'sex', where the privileged are men and the unprivileged are women, given the labels of 1 and 0 respectively, the aim is to use supervised ML algorithm to be able to predict whether a person has a good or bad credit risk, again the data was split into training (0.7) and testing (0.3) to carry out the experimentations. More information regarding this dataset can be found here [7].

5. Methodology

5.1. Task One

This task aims to investigate the accuracy and discrimination of the outputs generated by using a standard LR classifier. The classifier was trained on both the adult and German datasets; moreover, the hyper-parameter C, solver, and penalty were varied across five folds to develop the best combination that yields the models that scored the best accuracy and the best fairness. After successfully completing this task, four LR models will be obtained, two models for the best accuracy and two models for the best fairness for both datasets. All models were tested on the held-out testing sets for final evaluation.

5.2. Task Two

This task aims to investigate the accuracy and discrimination of the outputs generated by using the LR model, but with reweighing, the weights using the reweighing method explained in section 3.2. Moreover, the hyper-parameters were also varied to obtain the following four models, which score the best accuracy and fairness for both datasets. Then all models were tested on the held-out testing sets for final evaluation.

5.3. Task Three

Tasks one and two covered the comparison between the standard LR model and the discrimination-free LR using reweighing. Moreover, both models were evaluated using accuracy and equality of opportunity for fairness. For Task three, a different criterion was used to account for both fairness and accuracy of a model; based on that, the standard LR and the discrimination-free LR were evaluated, and the hyper-parameters were similarly varied across the five folds to obtain the model which scores the best accuracy and fairness values so that it can be thought of as a trade-off evaluation metric. The criterion used for evaluation was based on the concept of Pareto efficiency (PE). PE is used in various disciplines such as economics, engineering, and others. The definition of PE comes from the economic theory, which states that an alternation in an individual's state is said to be Perto efficient when it leaves at least one individual better off and no other individuals worse off; a point of Perto optimality is valid when no further PE changes can be made, this concept was first mentioned by the Italian economist Vilfredo Pareto as described in [9]. So, this theorem can be thought of as an evaluation/trade-off metric for accuracy and fairness of the classifiers by plotting curves of the false positive rate (FPR) denoting unfairness versus the error rate (1-accuracy) achieved by the classifiers with variations of hyper-parameters across the five folds, points which lie on the Pareto efficiency line/curve are considered Pareto efficient, if one variable needs to be optimised in value a trade-off or sacrifice will occur on the expense of the other value, i.e. if fairness is increased, then the error will also increase, as shown in the following research papers[10,11].

$$Error = \frac{False\ Positives + False\ Negatives}{\text{Actual Positives} + \text{Actual Negative}}$$

$$FPR = \frac{False\ Positives}{\text{False Positives} + \text{True Negatives}}$$

The values for the accuracy, FPR, error and other metrics can be obtained from the confusion matrix offered by the aif360 library.

6. Discussion and Results

CV and hyper-parameter tuning were carried out to obtain the optimal possible combinations on the training data. The results of all the models' performance on the held-out testing data for the three tasks are reported in table 1.

Data	Cls	Acc	Fairness	C	Penalty	Solver
A	1 S	0.804	-0.441	0.1	L2	Saga
A	2 F	0.804	-0.457	0.001	L1	Saga
G	3 S	0.717	-0.059	0.1	L1/L2	Saga/lib
G	4 F	0.717	-0.059	0.1	L1/L2	Saga/lib
A	5 S	0.791	0.035	0.001	L1	Saga
A	6 F	0.791	0.035	0.001	L1	Saga
G	7 S	0.717	-0.059	1.0	L1	Saga
G	8 F	0.717	-0.059	1.0	L1	Saga
A	9 F	0.788	-0.037	0.001	L2	Saga
A	10 S	0.805	-0.436	0.001	L2	Liblinear
G	11 F	0.653	-0.213	0.001	L2	Liblinear
G	12 S	0.717	-0.059	0.1	L2	Saga

Table 1: results of all classifiers on the testing data for both datasets.

In table 1, Cls denotes the classifier, and the "S" and "F" indicate whether the hyper-parameter combinations were used to score a standard or a fair LR model; in other words, "S" stands for the combination of hyperparameters to output the best accuracy, while "F" denotes the combination of hyperparameters to obtain a classifier with the best fairness. "A" in data represents the adult dataset, while "G" represents the german dataset. So, each task has 4 classifiers in total, one representing the optimal fairness while the other represents the optimal accuracy with the hyper-parameter variations for each dataset, the classifiers for each task are presented in order. The decision-making process of the models was based on the cv with hyperparameter tuning results which can be found in the appendix. Remember that an Accuracy closer to 1 and equal opportunity fairness closer to 0 are the optimal values for both evaluation metrics. From table 1, the following observations can be made:

- For task one, the classifiers seem to score better accuracy but worse fairness on the adult dataset compared to the German dataset. In both cases, the fairness of equality of opportunity shows bias towards the privileged group indicated by the negative sign accompanied with the fairness values, the classifier is scoring a better accuracy on the adult data due to the larger dataset compared to the German dataset size, which is considered very small for a ML problem, and that is a possible reason why the LR classifier is scoring better fairness when trained on it.
- For Task two, training the discrimination-free LR classifier on the adult dataset yielded a noticeable improvement in the equality of opportunity fairness; on the other hand, a minimal drop in accuracy occurred. As for the German dataset, the discrimination free LR classifier Yielded the

same results for both accuracy and fairness. As a result, using reweighing seems to produce better fairness results with the adult dataset and can be considered a successful trade-off compared to task one.

- For task three, comparing the discrimination-free classifiers 6 and 9 trained on the adult dataset, both scored equally close results for accuracy and fairness; classifier 9 showed minimal bias towards the privileged group while classifier 6 showed bias towards the unprivileged group denoted by the positive equality of opportunity fairness value.
- The standard LR classifier numbered 10 from task three scored a better accuracy and fairness than the standard LR classifier numbered 1 in task one, meaning that the proposed use of the PE trade-off can be considered a successful evaluation methodology.
- The discrimination-free LR classifier numbered 11 in task three scored worse accuracy and fairness results than the same classifier numbered 8 in task two, showing that the PE trade-off doesn't work properly compared to reweighing for the German dataset. This claim can also be backed up with the results of the standard LR classifier number 12 and the standard LR classifier number 3 in task one, as both score the same accuracy and fairness results. Therefore, using the PE trade-off with the German dataset is not recommended, while using reweighing is the preferred practice as it yielded better outputs. The PE plots for the adult dataset can be seen in figures 1 and 2.

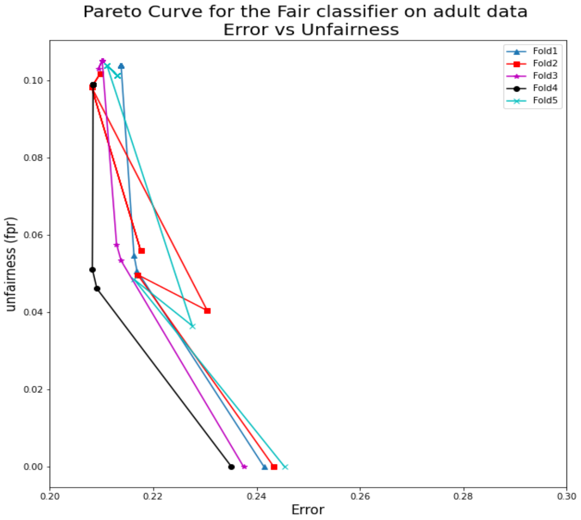

Figure 1: PE curve for fair classifier on the adult data, fpr vs error.

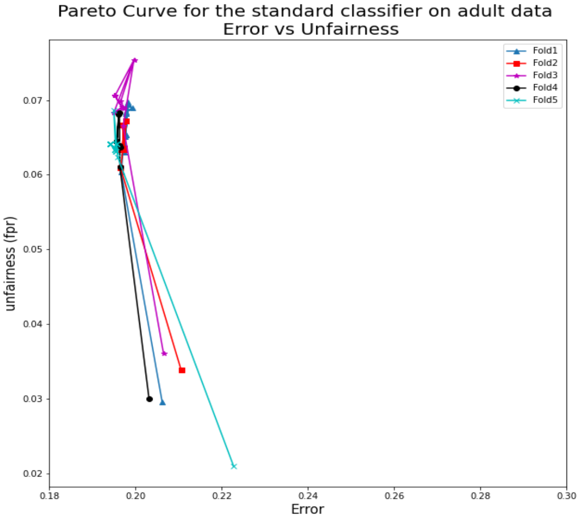

Figure 2: PE curve for standard classifier on adult data, fpr vs error.

From Figures 1 and 2, fold number 4 in black was chosen as the optimal PE line. It yielded the best fairness and accuracy for the adult dataset, with a fpr= 0.046 using the discrimination-free classifier and fpr= 0.062, keeping in mind that a smaller fpr is preferred.

7. Conclusion and further work

According to the PE theorem, a point is considered Pareto optimal when it makes one variable better and no other variable worse off. Consequently, whether to pick a ML model with better accuracy or fairness depends on the application of use. It all goes up to policy and decision makers to make that choice, keeping in mind that there are laws and regulations that protect sensitive attributes of individuals. After carrying out experimentation throughout the three tasks, a conclusion can be made that the LR model can be considered a good classifier, as shown in [12]. Furthermore, fairness and accuracy in ML models is a trade-off, varying hyperparameters of the models, and especially C controls how the decision boundary of the hyperplane is controlled to classify the points; this could be valid for the adult dataset, but not for the German dataset as carrying out cv on such a small dataset is not recommended, which yielded almost identical results for all variations of hyperparameters across the five folds, using larger datasets is recommended to carry out such experimentation. Finally, further investigation could be done by trying in-processing and post-processing fairness techniques, different classifiers, and trying a combination of other binary and non-binary sensitive features.

References

[1] AcademicianHelp. 2022. *AcademicianHelp*. [online] Available at: <https://www.academicianhelp.com/blog/logistic-regression-explained#:~:text=Logistics%20regression%20is%20a%20machine,minimizes%20the%20cross%20entropy%20loss.> [Accessed 13 May 2022].

[2] Bartosik, A. and Whittingham, H., 2021. *The Era of Artificial Intelligence, Machine Learning, and Data Science in the Pharmaceutical Industry*. Academic Press, pp.119-137.

[3] Calders, T., Kamiran, F. and Pechenizkiy, M., 2009. Building Classifiers with Independency Constraints. *IEEE International Conference on Data Mining Workshops*, pp.1-4.

[4] Kamiran, F. and Calders, T., 2011. Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems*, [online] 33(1), pp.1-18. Available at: <https://link.springer.com/content/pdf/10.1007/s10115-011-0463-8.pdf> [Accessed 11 May 2022].

[5] Ziyuan, Z., 2018. A Tutorial on Fairness in Machine Learning. [Blog] *Towards Data Science*, Available at: <https://towardsdatascience.com/a-tutorial-on-fairness-in-machine-learning-3ff8ba1040cb> [Accessed 11 May 2022].

[6] Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

[7] Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

[8] scikit-learn. 2022. *sklearn.linear_model.LogisticRegression*. [online] Available at: <https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html> [Accessed 4 May 2022].

[9] Law, J., 2008. *Pareto efficiency*. [online] Oxford Reference. Available at: <https://www.oxfordreference.com/view/10.1093/oi/authority.20110803100306253> [Accessed 11 May 2022].

[10] Kearns, M., Neel, S., Roth, A. and Steven Wu, Z., 2019. An Empirical Study of Rich Subgroup Fairness for Machine Learning. *ACM Digital Library*, [online] pp.1-7. Available at: <https://dl.acm.org/doi/pdf/10.1145/3287560.3287592> [Accessed 11 May 2022].

[11] Balashankar, A., Lees, A., Welty, C. and Subramanian, L., 2019. What is Fair? Exploring Pareto-Efficiency for Fairness Constrained Classifiers. *arxiv*, [online] pp.1-8. Available at: <https://arxiv.org/abs/1910.14120> [Accessed 11 May 2022].

[12] Zemel, R., Wu, Y., Swersky, K. and Dwork, C., 2013. Learning Fair Representations. *Proceedings of Machine Learning Research*, [online] 28, pp.6-7. Available at: <https://www.cs.toronto.edu/~zemel/documents/fair-icml-final.pdf> [Accessed 9 May 2022].

Appendix

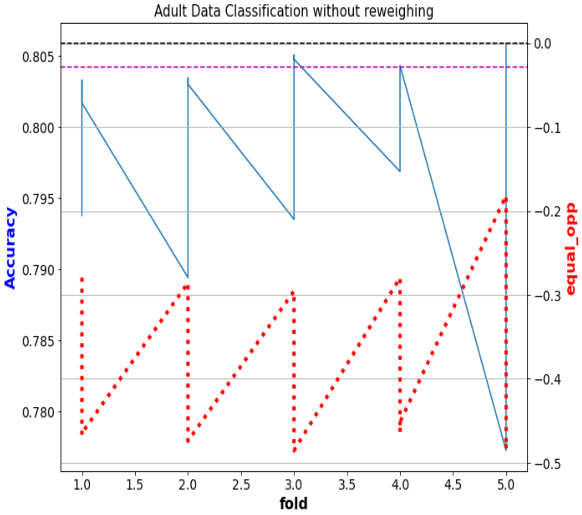

Figure 3: Accuracy vs fairness of the standard LR model on the cv of the adult dataset (task1), purple line is the accuracy obtained in the labs, acting as a benchmark.

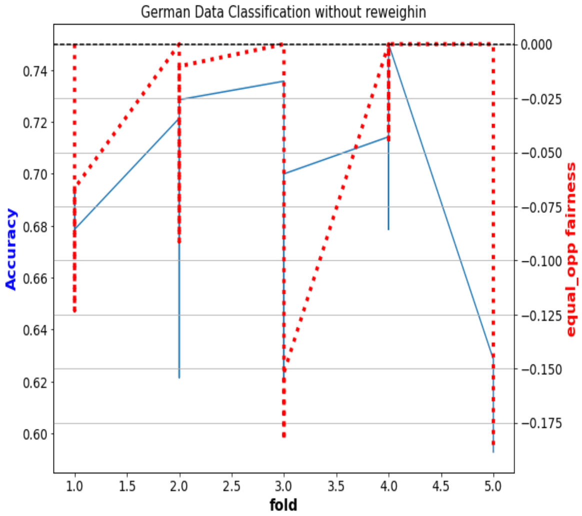

Figure 4: Accuracy vs fairness of the standard LR model on the cv of the German dataset (task1).