Student Number: 244837

1. Introduction

In this report we are discussing the model selection that considers accuracy and fairness metrics when training the machine learning models. The task is to compare standard machine learning models versus fairness-based machine learning models with respect to most fairness and most accuracy.

In numerous areas, such as job recruitment, health care, and even student evaluation, AI is rapidly being utilised to automate decision-making. When supervised machine learning is employed for this task and correlations are discovered, there is a special risk. Certain characteristics, like as colour, gender, and parental wealth, should not be utilised to determine a student's grades. Yet, because these variables correlate among non-protected data and because some supervised machine learning algorithms lack transparency, automated decision-making might result in biased outcomes. For this report we are using AI Fairness 360 toolkit which is an open-source library a complete set of indicators for testing for biases in datasets and models.

2. Related Work

There are many papers where AIF360 has been used for fairness research. In one paper by Bellamy and others they concluded that it provides a platform to experiment with and evaluate various existing bias detection and mitigation algorithms to acquire insights into their practical application within a shared framework. Contribute new datasets to share and examine for bias. Not just that it provides education on critical topics in bias detection and reduction, which measurements and mitigation methods should be used [5]
In another paper by Tor H. Aasheim, they reported certain fairness criteria, such as equal opportunity, are optimised using bias avoidance strategies. Bias mitigation in machine learning is now focused on intervention during the pre-processing, in-processing, and post-processing stages. [6]

3. Model

We have two primary objectives for this paper. The first aim is to determine whether higher generalization can lead to more equitable models. To acquire a satisfactory outcome, we need to choose a machine learning model and apply trade-off parameters; we also need to use training data to do 5-fold cross validation. We will select the model with the highest accuracy and the best fairness metric by adjusting the trade-off hyperparameter. In the second task, we repeat the same experiment but with an algorithmic fairness method like reweighing. Both the tasks should be performed on two data sets, that is the AI Fairness 360 adult dataset and another one of our choices, for which we have used the AIF 360 German dataset. For the model selection we are using logistic Regression.

Logistic Regression is a Machine Learning algorithm that is used to solve classification problems. It is a predictive analytic approach that is based on the probability notion.
A Logistic Regression model is similar to a Linear Regression model, except that it uses a more sophisticated cost function, which is known as the 'Sigmoid function' instead of a linear function. The function converts any real number into a number between 0 and 1. We utilise sigmoid to map predictions to probabilities in machine learning.

Data is used to learn model parameters, and hyper-parameters are tweaked to achieve the best fit. We employ a method in which we objectively search different values for model hyperparameters and select a subset that produces the best model on a given dataset.

Cross validation is a resampling process used to evaluate machine learning models on a limited data sample, and it will be employed in the models we are constructing. The process includes only one parameter, k, which specifies how many groups a given data sample should be divided into. As a result, the process is frequently referred to as k-fold cross-validation [1].

3.1 Implementation

As previously stated, we will use a logistic regression classification model with a 5-fold cross validation and hyperparameter tweaking. Only the parameter C, a hyperparameter value, will be used. Low value suggests that training data is more essential and represents real-world data, whereas C indicates the exact reverse. We use log space to pass the value of C and take five values in the range of 0 to -10. We have created a loop where it iterates through each value of C and does cross validation to find the accuracy and fairness for that model. Once the loop is completed, we will check which of the C values gave us the maximum accuracy and the best Equal opportunity value.

Since we used a 5-fold method we will see 5 values of C and corresponding accuracy and fairness. This is the process of hyperparameter tuning where we find the value of hyperparameter which gives us the best result.

Once that is done, we pass the C values for both the models, that is the one with the highest accuracy and the one with the best fairness metric across 5 folds into the full test, here we calculate the final accuracy and the fairness metric. This is how we implement task 1, in which we have to do for two datasets, one being AIF360 adult dataset, and the other one being the German dataset.

For task two, we will implement the identical things we did in task one, with the exception of implementing one fairness algorithm method, for which we will use reweighing. Reweighing is a pre-processing technique that Weights the examples in each (group, label) combination differently to ensure fairness before classification [2]. We'll be testing with both datasets that is adult dataset and the German dataset here as well.

4. Testing & Analysis

We did the experiment on two datasets for both the tasks, where we did the hyperparameter tuning first with the parameter C, A high C value instructs the model to give the training data more weight. A lower C value indicates that the model is giving complexity more weight at the expense of data fit. The analysis of both the tasks for both the datasets are given below.

4.1 Task 1 - Adult Dataset

After creating the cross validation function, we call it by passing the arguments for unscaled adult train set and the unscaled adult test set, along with that we pass the starting and ending values which are 0 and -10 into the log space function respectively and the number of values as 5, then the last argument which is being passed is the number of splits for the k fold for which we pass 5.

Within the function we loop through the different C values. For each C value we split the training set into 5 splits for the cross validation, out of which at one time one will be used as validation and the rest 4 for the training. This will go on for all 5 times since it's a 5 fold validation. This is done using another loop. And within this loop we scale the training and testing set in between 0 and 1. Once that's done we pass it to the model for the training where we use the logistic regression model. In turn we find the accuracy and Equal opportunity value which are appended on to a list for returning.

After this we plot the values we got and will get a graph like the one below.

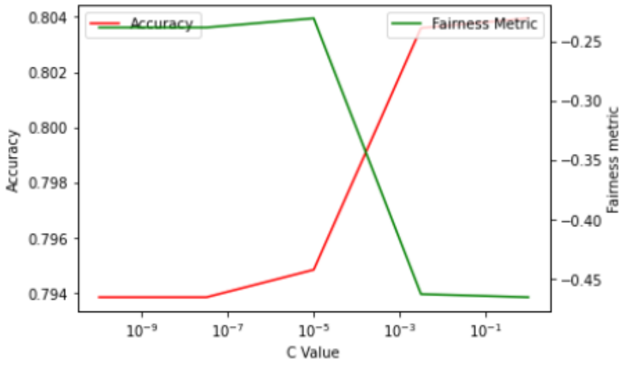


Figure 1: Accuracy and Fairness metric vs C value for task 1 adult dataset

From the above graph we can see that the Accuracy is highest when the value of C is 1 and we get the best fairness metric when the value of C is 10^{-5}. From this we can conclude that accuracy (Red) increases as the C value increases and the fairness metric(Green) increases as the C value decreases.

From the tuning, we found the values of C which gives us the best result, which will be passed to the full test function to find the corresponding accuracy and fairness metric. When we pass the value 1 which is the C value for the best accuracy into the full test we get accuracy and Fairness as 80.4% and -0.44. When we pass C value as 1e-05 we get the accuracy and fairness metric as 79.7% and -0.21 respectively. From this we observe that when we pas the C value with best accuracy into full test we get a good accuracy and but fairness isn't that good, but when we pass C value for best fairness into the model, we got the best Fairness metric till now. Hence we can say that better generalization can correspond to fairer models.

4.2 Task 1 – German Dataset

We do the same process as above here, but instead of the adult dataset we do it on the German dataset. The only difference is that the privileged and un privileged group will be based on age rather than sex since we are dealing with credit system here and the bias is on that. After doing the cross validation with different values of C and plotting the accuracy and the fairness metric against each other we get a graph like the one below.

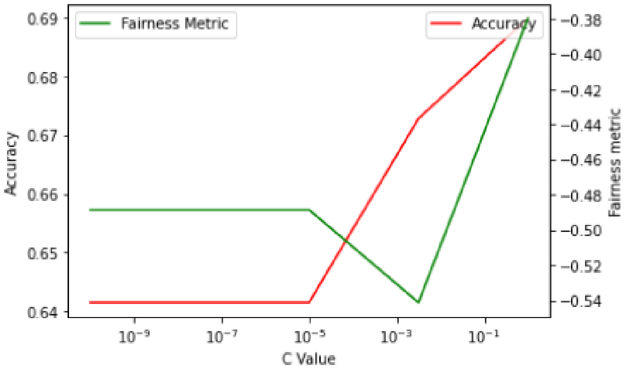


Figure 2: Accuracy and Fairness metric vs C value for task 1 German dataset

From this graph we can observe that the accuracy and fairness metric are the highest when the Value of C is 1. For fairness metric we can see that there is a sudden drop from 10^{-5} to 10^{-3} and then a sudden inclination. Also, the highest accuracy is 69% which might be because the German dataset is smaller compared to the adult one.

4.3 Task 2 – Adult Dataset

For task two we just modify it a bit, that is to bring a fairness algorithm to reduce the bias. For the same we are using reweighing. The function is almost the same except that we are adding the reweighing code into the inner loop before the prediction. Then we call the function with the same C values as the last two models and plot the resulting accuracy and fairness metric on the graph.

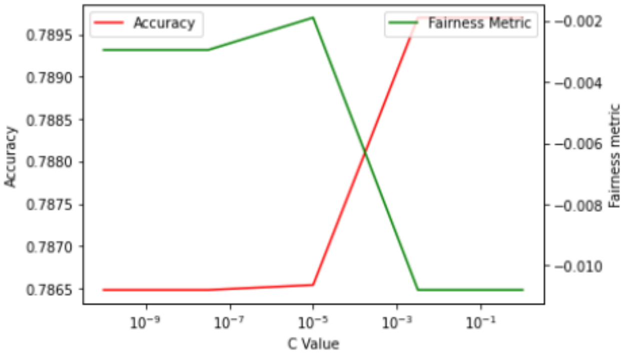


Figure 3: Accuracy and Fairness Vs C for Task 2 Adult Set

From the graph, we can observe that the accuracy is highest when C is 1 and fairness metric is highest when C is 10^{-5}. The highest accuracy is 78.96% and the best fairness metric is -0.001 which is really close to 0. Hence we can see that the reweighing algorithm is working and the bias is really less compared to the previous models.

4.4 Task 2 – German Dataset

For this task we use the German dataset which is the only difference. Everything else is the same, we are passing the same values of C, where we do cross validation and reweighing and then plotting the accuracy and fairness metric as below.

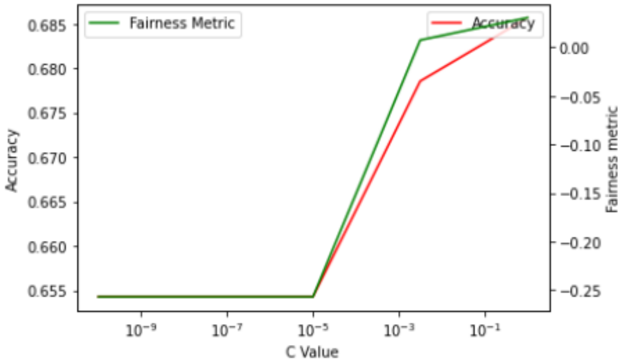


Figure 4: Accuracy and Fairness metric Vs C value for Task 2 German dataset

From the graph, we can observe that the highest accuracy is when C is 1 and best fairness metric is when C is $3*10^{-3}$. The highest accuracy is 68.5% which is less because of the small size of dataset and the fairness metric is 0.007 is nearer to 0 which is because of the reweighing algorithm to reduce the bias.

4.5 Result Table

MODEL	Task	C		Accuracy	Fairness Metric
Adult	1	Max Accuracy	1	80.4	-0.441
		Best Fainess	0.00001	79.7	-0.21
	2	Max Accuracy	1	79.05	0.035
		Best Fainess	0.00001	78.7	0.033
German	1	Max Accuracy	1	69.6	-0.43
		Best Fainess	1	69.6	-0.43
	2	Max Accuracy	1	70.6	0.011
		Best Fainess	0.003	70.6	0.066

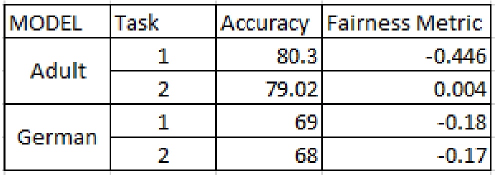
Table 1: Logistic regression results for both the datasets

The above table depicts the models, tasks and the C values we passed to the final testing data and the resulting accuracy and fairness metric. From that we can observe that the accuracy is less for German data in general. And for task 2, that is when we do the reweighing the fairness metric is nearer to 0, which means that the bias is reduced when we use the fairness algorithm. For the German dataset task 1 we can see that the C value is the same, that is 1 and hence the accuracy and the fairness metric is also the same. But here we can see that the accuracy is 69% which maybe because of the size of the dataset. For task 1 in adult dataset we can sat that taking C value as 0.00001 is better since accuracy doesn't change much but there is a significant change in fairness metric. For task 2 both are almost the same because of reweighing. For German data set with task 1 we take C value as 1 and for task 2 we take C value a 1 since the accuracy is the same but Fairness is closer to 0.

5. Extra Content

Beyond the lecture topics, we're looking into machine learning methodologies, for the same we are using Random Forest classifier. Random forest is a supervised machine learning algorithm that is commonly used to solve classification and regression problems. It creates decision trees from various samples, using the majority vote for classification and the average for regression. [3]

One of the most essential characteristics of the Random Forest Algorithm is that it can handle data sets with both continuous and categorical variables, as in regression and classification. It produces better outcomes for classification issues. We will be using this model for both the adult dataset and German dataset and will also do reweighing as we did in task 2. The results from the final tests are depicted in the table below.



MODEL	Task	Accuracy	Fairness Metric
Adult	1	80.3	-0.446
	2	79.02	0.004
German	1	69	-0.18
	2	68	-0.17

Table 2: Random Forest Model Results

From the table we can see that for the adult dataset the accuracy is high but for the German data set the accuracy is comparatively low, this is mainly due to the size of the dataset. Then the Fairness metric is near to 0 for the adult dataset after reweighing while German is -0.17, this is because in random classification we are already we are already doing bootstrapping, and since the data size is small for German dataset the reweighing doesn't make any change.

5.1 Advantages

-One of the main advantages of random forest is that we don't have to do cross validation. The random forest algorithm is made up of a series of decision trees, each of which is made up of a bootstrap sample of data selected from a training set with replacement. One-third of the training sample is set aside as test data, which will be utilised for cross validation.

-Contradicting to how we addressed the model, it can actually be used as a classifier and a regressor.

-In Random Forest, because it employs a rule-based approach, no data normalisation is necessary.

5.2 Disadvantages

In the same manner that we have benefits, we also have disadvantages.

-It a time-consuming process compared to the logistic regression model.

- It also takes a long time to train because it uses a number of decision trees to select the class.

- It also lacks comprehensibility due to the ensemble of decision trees and fails to determine the significance of each variable [4].

6. Future Work

We plan to investigate with even more datasets in the long term, as well as perform the tests with random training samples to include standard deviation. It can also be used to quantify and mitigate other facets of fairness, such as compensatory fairness, which refers to how well individuals are reimbursed for damages they have suffered. More effort is needed to expand the sorts of reasons provided as well as provide direction to professionals on when each type of explanation is most suitable.

7. Conclusion

This report looked at model selection using the accuracy and fairness metrics. We utilised a Logistic regression model to see how the value of the accuracy and fairness metrics changed when the hyperparameter was tweaked.

Then we utilised the Reweighing fairness method to remove the bias, and we noticed that the fairness metric changes and approaches zero. For the first task we concluded that better generalization could correspond to better results. And for the second take we saw how the fairness metric improved drastically. As extra work, we chose random forest as the ML model and ran the full test on both datasets, each with its own set of benefits and drawbacks.

1. References

[1] J. Brownlee, "A Gentle Introduction to k-fold Cross-Validation," 2018. [Online]. Available: https://machinelearningmastery.com/k-fold-cross-validation/. [Accessed 12 05 2022].

[2] aif360, "aif360.algorithms.preprocessing.Reweighing — aif360 0.4.0 documentation," [Online]. Available: https://aif360.readthedocs.io/en/latest/modules/generated/aif360.algorithms.preprocessing.Reweighing.html. [Accessed 12 5 2022].

[3] S. E. R, "Random Forest | Introduction to Random Forest Algorithm," 2021. [Online]. Available: https://www.analyticsvidhya.com/blog/2021/06/understanding-random-forest/. [Accessed 12 05 2022].

[4] Great Learning Team, "Random Forest Algorithm- An Overview | Understanding Random Forest," 2020. [Online]. Available: https://www.mygreatlearning.com/blog/random-forest-algorithm/. [Accessed 12 05 2022].

[5] K. D. H. C. H. S. H. R. K. E.Bellamy, "AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias," *IBM Journal of Research and Development*, vol. 63, 2019.

[6] K. S. H. B. M. S. Tor Aasheim, "Bias mitigation with AIF360: A comparative study," *No. 1 (2020): NIK Norsk informatikkonferanse* , 2020.