Student Number: 245309

1. Introduction

Machine learning fairness is an area which deals with reducing model inaccuracies and data biases to ensure equal opportunities to individuals irrespective of their sex, age, disabilities, etc [1][2]. In this assignment, the Random Forest classifier is used for binary classification on the Adult and German dataset. To find the right blend of hyperparameters, 5-fold cross validation is used and tested for accuracy and equal opportunity difference fairness metric. The hyperparameters with the highest accuracy and best fairness results are then used to create two separate models which are trained on the overall training set and tested on the testing data. The previously described method is then coupled with an ensemble method The first and second approach are then evaluated and compared using a proposed metric.

2. Dataset description

For this assignment, two datasets have been used, namely: Adult and German [3]. Both datasets have been split using 70% of the data for training and 30% for testing. These datasets have been used from AIF360 library and have been loaded using the pre-processing function which converts the education and age into respective binary groups for the Adult dataset and credit history, savings, employment for the German dataset. Sex is the protected attributed used for experimentation.

Table 1: Description of the datasets

Name	Columns	Rows	Feature names
Adult	19	48842	Race, sex, age, education & income
German	12	1000	Age, sex, credit history, savings, unemployment & credit

The German dataset was selected to analyze the effect of a smaller dataset on the fairness metrics. According to table 1, the German dataset is much smaller than the adult dataset and would be useful in analyzing the bias in a smaller dataset. To standardize the functionality range of the datasets, standard scaler has been used [4]. Using the equation below, each sample (x) is standardized by using the mean (u) and standard deviation (s).

$$z = \frac{(x - u)}{s}$$

3. Methods

To determine the effect of biases in machine learning models, the Random Forest classifier is chosen. The chosen classifier is trained and tested on both datasets. Further, this chosen classifier is coupled with an ensemble method comprising of pre-processing, in-processing and post-processing of transforming the dataset to reduce the bias.

3.1 Random Forest

The random forest classifier is a combination of decision tree classifier where in each of these trees are created by selecting a random set independently sampled vector from the original input vector. The method of bagging which creates training data by randomly drawing data from the input set with N replacements. To build the classifier, the following hyperparameters were used: (i) number of estimators, (ii) maximum depth, (iii) minimum sample split and (iv) splitting criterion. The variations of hyperparameters used have been described in table 2.

Table 2: Random Forest classifier hyperparameters

Name	Values
Number of estimators	10, 100 & 500
Max depth	5, 10, 20 & 50
Minimum sample split	20, 50 & 100
Criterion	Gini index & entropy

The number of estimators is the numbers of trees generated in the random forest. The max depth parameter defines the maximum depth to which each tree can be generated while the minimum number of sample split defines the smallest number of samples to be present in the respective internal node for splitting. Lastly, the criterion defines the measure according to which the impurity of a node is to be calculated and split. There are two criterions used for calculation which are:

Gini Index: In the equation below, p_j is the probability of the class j. This criterion calculates the frequency of an element when it is mislabeled randomly

$$Gini = 1 - \sum_j P_j^2$$

Entropy: In the equation below, p_j is the probability of the class j. This criterion measures the randomness associated with any randomly misclassified element.

$$Entropy = -\sum p \log_2 p_j$$

Each of the above mentioned hyperparameters are vital to building a random forest classifier with the aim of maximizing accuracy, reducing overfitting and underfitting, and finding the right criterion to find the impurities in the nodes.

3.2 Algorithmic Fairness models

Developing responsible and fair models are the need of the hour [5]. Most machine learning systems are driven by data and require this data to train itself where the underlying data is biased thus, resulting in biased predictions by the algorithm [6]. There are several algorithms created with the aim of mitigating biases. These methods ensure fair treatment of individuals by quantifying biases and

removing this from the data by utilizing pre-processing, in-processing or post-processing methods.

3.2.1 Ensemble fairness method

In this assignment, to utilize the best aspects and maximize fairness, an ensemble method has been used which consists of a pre-processing, in-processing and post-processing method.

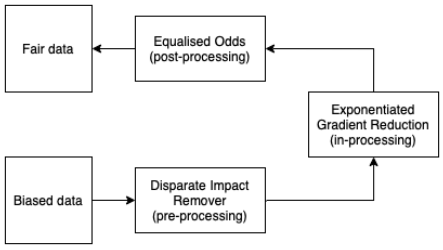


Figure 1: ensemble fairness classifier

Figure 1 describes the ensemble method used to maximize fairness. The process begins with the biased data being fed into the Disparate Impact Remover. The intermediary data is then passed into the Exponentiated Gradient Reducer which has been created using a Random Forest Classifier. Finally, the data is fed into the Equalized Odds method which removes any remaining bias present in the data.

3.2.1.1 Disparate Impact Remover

The first step of the proposed ensemble model is to remove the disparate impact. Disparate Impact can be defined using the equation given below. It can be defined as the ratio of the proportion of underprivileged individuals class individuals to the privileged class individuals.

$$\frac{Pr(Y=1|D=\text{unprivileged})}{Pr(Y=1|D=\text{privileged})}$$

To overcome this bias, [7] proposed a method which manipulates the data to increase the fairness by overlapping the distributions of the two classes using a repair level. For this assignment, the repair-level has been set to 1.0 indicating maximum overlapping of distributions with the aim of maximizing fairness.

3.2.1.2 Exponentiated Gradient Reduction

Once the biased data has less disparity within itself, the data is passed into the machine learning model. Exponentiated Gradient Reduction machine learning model selected for this task due to its ability to incorporate random forest classifiers into its methodology thus, removing bias as well as allowing a distinction to be made with the non-fairness random forest classifier.

This method was proposed by [8], returns a randomly generated binary classifier which has the least error with respect to the classification constraints. This method primarily reduces each of the classifiers into a sequence of cost-sensitive classifications. As described by the authors, this method uses a black box architecture which aims at optimizing the accuracy while improving the fairness.

In this assignment, the constraint of Equalized Odds has been used to train the classifier.

3.2.1.3 Equalized odds

The last step in the ensemble fairness method is to use optimization to calibrate the scores generated by the classifier [9]. This method calculates the probabilities of the scores which are further used to change the labels of the output labels with the objective of maximizing the equalized odds constraint.

3.3 Metrics

In this assignment, the classifiers are trained using 5-fold cross validation. These classifiers are tested using the validation data. The model with the best validation metrics is selected and further used to test the testing data. The following metrics have been used:

3.3.1 Accuracy

To find the performance of a model, its accuracy is calculated using the following formula:

$$Accuracy = \frac{T_P + T_N}{T_P + T_N + F_P + F_N}$$

Where T_P: True positive, F_P: False positive, T_N: True negative, and F_N: False negative.

3.3.2 Equal opportunity difference

The following metric is used to find the fairness:

$$EOD = TPR_{D=unprivileged} - TPR_{D=privileged}$$

This metric is the difference between the true positive rate of the privileged class subtracted by the true positive rate of the unprivileged class.

3.3.3 Proposed accuracy + fairness metric

Specifically for task 3, the following is proposed:

$$Criterion = Accuracy - |EOD|$$

This criterion is used to calculate the difference between the accuracy and the absolute equal opportunity difference where a score of 1 signifies optimal accuracy and fairness while a score of 0 or less signifies low accuracy and high amount of bias.

4. Results and analysis

The Adult and German datasets have been used for experimentation for the three tasks. For the first two tasks, the hyperparameters described in the previous section have been used to create the models and then, they are cross validated using the 5-folds on the training data. In the first task, a random forest classifier is used and the best parameters with the best accuracy and equal opportunity

difference are used to train and test the overall model. In task 2, an ensemble method is used the same sequence with an overall objective of increasing fairness. In the last task, the metric proposed has been used to find the most balanced model.

4.1 Adult dataset

According to appendix I, in terms of accuracy, there was a decrease when the number of estimators increased and an increase when the minimum number of sample split increased for task 1 and 2. Entropy worked better for the first task while Gini gave better results for the second task. As the max depth increase, for task 1 there was a fall while there was an increase in accuracy for task 2. In terms of fairness, the results had opposite trends for the splitting criterion and number of estimators with respect to the models maximizing accuracy.

4.1.1 Task 1

According to figure 2, the highest accuracy observed is 80.3% and its respective fairness is -46%. The model with the best fairness was at 36.45% while its accuracy was nearly 80%. The standard deviation and variance of the training accuracy was 0.001 and $1\times e^{-6}$ respectively. The fairness metric had a standard deviation and variance of 0.0226 and 0.0005 respectively.

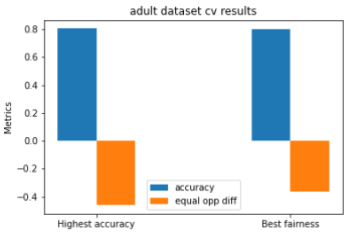


Figure 2: Training accuracy and fairness (Adult task 1)

The classifier was created using the best hyperparameters and trained. According to the results obtained (table 3), the testing results very similar to the training results of the best parameters for both cases. Although there is a betterment in the fairness metric for the testing set.

Table 3: Testing accuracy and fairness (Adult – task 1)

Type	Number of estimators	Max depth	Min samples split	criterion	Acc-uracy	Fair-ness
Accuracy	100	10	50	Gini	0.803	-0.43
Fairness	100	5	100	Gini	0.797	-0.30

4.1.2 Task 2

According to figure 3, the ensemble model training accuracy of both the models (accuracy centric and fairness centric) were like the task 1 accuracy of 80% while the equal opportunity difference is very close to 0 signifying a substantial improvement in the metric with a slight decrease in accuracy. These models also showed small variance and standard deviation.

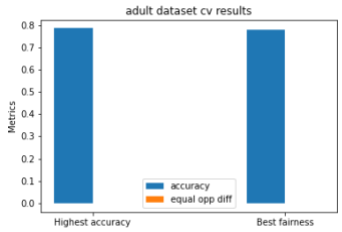


Figure 3: Training accuracy and fairness (Adult task 2)

The testing accuracy showed similar results for both the models. The testing accuracy of both the models were similar while the fairness metric of the fairness centric model was slightly better than the accuracy centric.

Table 4: Testing accuracy and fairness (Adult – task 2)

Type	Number of estimators	Max depth	Min samples split	criterion	Acc-uracy	Fair-ness
Accuracy	500	20	100	Entropy	0.788	$6xe^{-4}$
Fairness	500	5	20	Entropy	0.787	$-2xe^{-4}$

4.1.3 Task 3

According to the proposed criteria, the models from task 1 and task 2 were chosen and their respective training metrics are shown in table 5.

Table 5: Model selection based on proposed criteria (train)

Type	Accuracy	Fairness	Criterion
Model t1_1	0.798	-0.3640	0.434
Model_t2_1	0.784	0.00003	0.784

Upon training the models with the chosen hyperparameters, the models from the first task had a high accuracy with a low fairness metric like the task 1 trend (table 6). The fairness centric models showed a similar trend to the models in task 2. Overall, both the models were balanced with respect to their architecture.

Table 6: Model selection based on proposed criteria (test)

Type	Estim-ators	Max dep-th	Min samples split	Critr-ion	Accu-racy	Fair-ness
Criterion_t1	100	10	50	Gini	0.804	-0.43
Criterion_t2	500	20	20	Gini	0.784	$-2 \times e^{-3}$

4.2 German dataset

According to appendix II, for the German dataset, the Gini criterion outperformed the entropy for most models while the task 1 models had an increase in accuracy for an increase in max depth and a decrease in accuracy and it remained constant for both task 2 models. On increasing the min sample split, the task 1 models had an increase in accuracy while the task two models had a decrease. Finally, upon increasing the number of estimators, the task 1 models

had a decrease in accuracy while the task 2 models had an increase in accuracy.

4.2.1 Task 1

According to figure 4, the model with the best accuracy was at 70% with a fairness score of -0.09. The model with the best fairness score had an accuracy of 69% with a fairness score of -0.09. The standard deviation and variance for accuracy was very small and the same was true for the fairness metric.

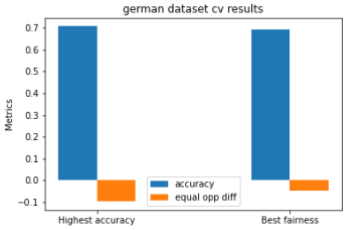


Figure 4: Training accuracy and fairness (German task 1)

According to table 7, the accuracy-oriented model had a better accuracy and a good fairness value. Notably, the fairness-oriented model had a slightly lesser accuracy but a better fairness score. This could possibly occur due to the lack of samples available in the training data and testing data.

Table 7: Testing accuracy and fairness (German – task 1)

Type	Number of estimators	Max depth	Min samples split	criterion	Accuracy	Fairness
Accuracy	100	5	20	Entropy	0.723	-0.07
Fairness	50	10	100	Entropy	0.723	-0.06

4.2.2 Task 2

The problem of lack of samples was also seen in this task. Despite classifying using the ensemble method, there was a reduction in the accuracy by 10% in the best model and it also had a significantly undesirable fairness score. The fairness-centric model too suffered a decreased accuracy but had a good fairness score. The accuracy and fairness had a larger variance and standard deviation in accuracy with respect to the task 1 models.

According to table 8, notably, the fairness centric model had a lower fairness score than the accuracy centric model signifying a failure in choosing the right hyperparameter due to the lack of samples. The accuracy centric model had a similar accuracy to the cross validated models.

Table 8: Testing accuracy and fairness (German – task 2)

Type	Number of estimators	Max depth	Min samples split	criterion	Accuracy	Fairness
Accuracy	500	5	50	Entropy	0.716	0.013
Fairness	50	10	100	Entropy	0.3933	0.027

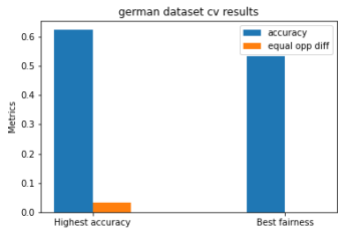


Figure 5: Training accuracy and fairness (German task 2)

4.2.3 Task 3

According to table 9 and 10, the proposed criterion was calculated, and the respective models were trained on the overall data. The proposed criterion worked well as it gave accurate as well as fair models, but the fairness model showed lesser fairness than the accuracy centric.

Table 9: Model selection based on proposed criteria (train)

Type	Accuracy	Fairness	Criterion_1
Model_t1	0.691	-0.045	0.645
Model_t2	0.604	-0.001	0.603

Table 10: Model selection based on proposed criteria (test)

Type	Estimators	Max depth	Min samples split	Critrion	Accuracy	Fairness
Criterion_t1	50	5	100	Entropy	0.713	0.000
Criterion_t2	500	5	50	Gini	0.710	$-8xe^{-3}$

5. Extension

To determine the impact of sensitive features on the classifier, the random forest classifier was trained with and without and without the sensitive features on the adult dataset (i.e., age and sex). According to Appendix III, the sensitive attributes play a key role in determining the accuracy and fairness. In the presence of the features, the accuracy was high, and the fairness was low while in the absence of the features, the accuracy and fairness were very good.

6. Conclusions

In this assignment, the presence of bias in machine learning algorithms were explored and methods to mitigated them were experimented. It was found that the size of the dataset plays a key role along with the way it is balanced. If the training set is biased and test set isn't, there is an improvement in the accuracy and fairness. The ensemble model used for task 2 increased fairness with a slight decrease in accuracy. Further, the proposed criterion for accuracy+fairness gave desirable results.

References

[1] Chouldechova, Alexandra, and Aaron Roth. "The frontiers of fairness in machine learning." arXiv preprint arXiv:1810.08810 (2018).

[2] Chouldechova, A., & Roth, A. (2018). The frontiers of fairness in machine learning. arXiv preprint arXiv:1810.08810.

[3] Fabris, A., Messina, S., Silvello, G., & Susto, G. A. (2022). Algorithmic Fairness Datasets: the Story so Far. arXiv preprint arXiv:2202.01711.

[4] https://scikitlearn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html

[5] Kleinberg, J., Ludwig, J., Mullainathan, S., & Rambachan, A. (2018, May). Algorithmic fairness. In Aea papers and proceedings (Vol. 108, pp. 22-27).

[6] Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. ACM Computing Surveys (CSUR), 54(6), 1-35.

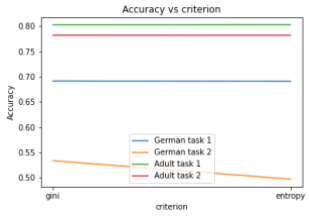
[7] Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., & Venkatasubramanian, S. (2015, August). Certifying and removing disparate impact. In proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining (pp. 259-268).

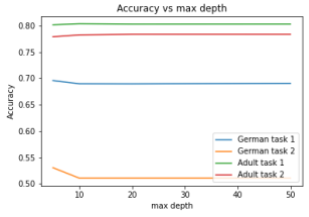
[8] Agarwal, A., Beygelzimer, A., Dudík, M., Langford, J., & Wallach, H. (2018, July). A reductions approach to fair classification. In International Conference on Machine Learning (pp. 60-69). PMLR.

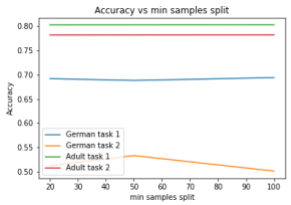
[9] Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. Advances in neural information processing systems, 29.

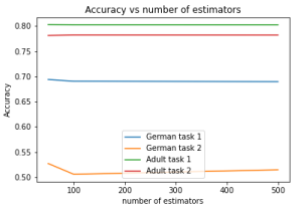
Appendix

I. Accuracy vs hyperparameters

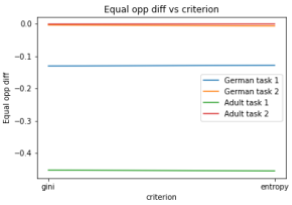


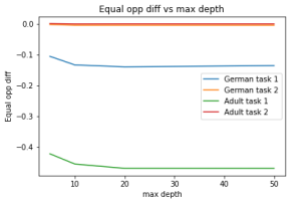


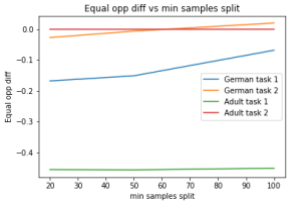


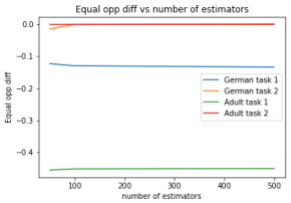


II. Fairness vs hyperparameters









III. Extension

Estimators	Presence of features		Absence of features	
	Accuracy	Fairness	Accuracy	Fairness
50	0.8049	-0.503	0.788	0.054
100	0.8049	-0.504	0.788	0.054
200	0.8048	-0.505	0.788	0.054
500	0.8049	-0.505	0.788	0.054