A Multi-objective settings

We consider the problem of optimizing multiple objectives. Specifically, for our use case, we leverage balanced accuracy [Pedregosa *et al.*, 2011] as a quality metric for the performance, and demographic parity ratio and equalised odds ratio [Weerts *et al.*, 2023] for fairness.

Given a binary classification problem, let $\hat{Y}$ represent the predictions made by a specific model on a given dataset, and let Y denote the corresponding ground truth labels. Then, we define:

- **True Positive** (TP) predictions: the number of cases when $\hat{y}_i = 1$ and the corresponding $y_i = 1$, where $\hat{y}_i \in \hat{Y}$ and $y_i \in Y$;
- **True Negative** (TN) predictions: the number of cases when $\hat{y}_i = 0$ and the corresponding $y_i = 0$, where $\hat{y}_i \in \hat{Y}$ and $y_i \in Y$;
- **False Positive** (FP) predictions: the number of cases when $\hat{y}_i = 1$ and the corresponding $y_i = 0$, where $\hat{y}_i \in \hat{Y}$ and $y_i \in Y$;
- **False Negative** (FN) predictions: the number of cases when $\hat{y}_i = 0$ and the corresponding $y_i = 1$, where $\hat{y}_i \in \hat{Y}$ and $y_i \in Y$.

Thus, we can introduce the balanced accuracy as:

$$\text{Balanced Accuracy} = \frac{1}{2}\left(\frac{\text{TP}}{\text{TP}+\text{FN}} + \frac{\text{TN}}{\text{TN}+\text{FP}}\right)$$

As to the fairness metrics, let us now represent the sensitive attribute with $X_s \in \mathbb{X}$ and the set of all possible values of the sensitive attribute with $\mathcal{X}_s$. We define the demographic parity ratio and equalised parity ratio.

$$\text{Demographic Parity Ratio} = \frac{\min_{x_s \in \mathcal{X}_s} \mathbb{P}(\hat{Y} = 1 \mid X_s = x_s)}{\max_{x_s \in \mathcal{X}_s} \mathbb{P}(\hat{Y} = 1 \mid X_s = x_s)}$$

$$\text{True Positive Rate Ratio} = \frac{\min_{x_s \in \mathcal{X}_s} \mathbb{P}(\hat{Y} = 1 \mid X_s = x_s, Y = 1)}{\max_{x_s \in \mathcal{X}_s} \mathbb{P}(\hat{Y} = 1 \mid X_s = x_s, Y = 1)}$$

$$\text{False Positive Rate Ratio} = \frac{\min_{x_s \in \mathcal{X}_s} \mathbb{P}(\hat{Y} = 1 \mid X_s = x_s, Y = 0)}{\max_{x_s \in \mathcal{X}_s} \mathbb{P}(\hat{Y} = 1 \mid X_s = x_s, Y = 0)}$$

$$\text{Equalized Odds Ratio} = \min\left(\text{True Positive Rate Ratio}, \text{False Positive Rate Ratio}\right)$$

In particular, in [Weerts *et al.*, 2023], the probability is estimated with the frequencies on the real dataset.

When considering multiple objectives, we cannot definitively decide which solution is the best, as improvements in one objective may lead to degradation in another. We seek a Pareto front, which represents the solutions that have the best trade-offs with respect to the objectives (i.e., the set of non-dominated solutions). A solution is considered non-dominated if there is no other solution that is better in at least one objective, without being worse in another. The Pareto front $P_{\mathcal{D}_{val}}(\mathcal{H})$ for a given set of solutions $\mathcal{H} \subset \mathbb{H}$ evaluated on dataset $\mathcal{D}_{val}$ is defined as:

$$P_{\mathcal{D}_{val}}(\mathcal{H}) = \left\{ H \;\middle|\; \begin{array}{l} H \in \mathcal{H}, \nexists H' \in \mathcal{H} : \\ \forall m \in \{1, \ldots, M\} : \\ \mathcal{M}_m(H', \mathcal{D}_{val}) \geq \mathcal{M}_m(H, \mathcal{D}_{val}), \\ \exists j \in \{1, \ldots, M\} : \\ \mathcal{M}_j(H', \mathcal{D}_{val}) > \mathcal{M}_j(H, \mathcal{D}_{val}) \end{array} \right\}.$$

Given a Pareto front, and the quality metric values of the models within $v = \{(\mathcal{M}_1(H), \ldots, \mathcal{M}_m(H)) : H \in P_{D_{val}}(\mathcal{H})\}$, the hypervolume is defined as:

$$\text{Hypervolume} = \text{Volume}\left(\bigcup_{v_i \in v} \{\boldsymbol{x} \in \mathbb{R}^d \mid v_i \preceq \boldsymbol{x} \preceq \boldsymbol{r}\}\right)$$

where $\boldsymbol{r}$ is the optimal reference point (i.e., best value for each quality metric).

B Search Space

The tests are run on datasets from OpenML [Vanschoren *et al.*, 2013]—a well-known repository for data acquisition and benchmarking. As it provides already-encoded datasets, we do not consider the encoding step. As to Imputation, the adult dataset has no missing values, and in the COMPAS dataset, we drop the few instances with missing values. Except for that, we included all the Data Pre-processing steps available in the scikit-learn [Pedregosa *et al.*, 2011] Python library (plus imbalance-learn [Lemaitre *et al.*, 2017] for Rebalancing transformations). The leveraged steps, algorithms per step, and hyperparameters per algorithm are reported in Table 2.

Table 2: Algorithms and number of hyperparameters for each of the steps in HAMLET4Fairness. Algorithm names and hyperparameters (No. Hps) are imported from the scikit-learn Python library.

Step	**Algorithm**	**No. Hyperparameters**
Mitigation	CorrelationRemover	1
	LearnedFairRepresentation	2
Normalization	StandardScaler	2
	MinMaxScaler	0
	RobustScaler	2
	PowerTransformer	0
Discretization	Binarizer	1
	KBinsDiscretizer	3
Feature Eng.	SelectKBest	1
	PCA	1
Rebalancing	NearMiss	1
	SMOTE	1
Classification	KNeighborsClassifier	3
	RandomForestClassifier	7
	MLPClassifier	6

Table 3: Discovered rules and statistics for COMPAS.

Approach	**Metric**	**ID**	Rule Type		S_1	S_2	**A**	**Metric Thres.**	**No. Conf.**	**Supp.**
			`mand.`	`mand. order`						
PKB+IKA	DMR	1	✓	✓	$\mathcal{M}it$	$\mathcal{N}$	$\mathcal{R}f$	0.8	80	0.7
		2	✓	✓	$\mathcal{M}it$	$\mathcal{F}e$	$\mathcal{R}f$	0.8	80	0.7
		3	✓	✓	$\mathcal{M}it$	$\mathcal{N}$	$\mathcal{K}nn$	0.6	56	0.64
		4	✓	✓	$\mathcal{M}it$	$\mathcal{F}e$	$\mathcal{K}nn$	0.6	56	0.77
		5	✓	✓	$\mathcal{M}it$	$\mathcal{N}$	$\mathcal{M}lp$	0.6	53	1.0
		6	✓	✓	$\mathcal{M}it$	$\mathcal{F}e$	$\mathcal{M}lp$	0.6	53	0.62
	BA	7	✓	✓	$\mathcal{M}it$	$\mathcal{N}$	$\mathcal{R}f$	0.6	101	0.66
		8	✓	✓	$\mathcal{M}it$	$\mathcal{F}e$	$\mathcal{R}f$	0.6	101	0.6
PKB	DMR	9		✓	$\mathcal{M}it$	$\mathcal{N}$	$\mathcal{R}f$	0.8	73	0.53
		10	✓	✓	$\mathcal{M}it$	$\mathcal{F}e$	$\mathcal{R}f$	0.8	73	0.6
		11	✓	✓	$\mathcal{M}it$	$\mathcal{N}$	$\mathcal{K}nn$	0.6	55	0.69
		12	✓	✓	$\mathcal{M}it$	$\mathcal{F}e$	$\mathcal{K}nn$	0.6	55	0.82
	BA	13	✓		$\mathcal{M}it$	-	$\mathcal{R}f$	0.6	83	1.0

DMR = Demographic Parity Ratio, BA = Balanced Accuracy
$\mathcal{M}it$ = Mitigation, $\mathcal{N}$ = Normalization, $\mathcal{F}e$ = Feature Eng.,
$\mathcal{K}nn$ = KNeighborsClassifier, $\mathcal{R}f$ = RandomForestClassifier,
$\mathcal{M}lp$ = MLPClassifier

C Extensive Set and Statistics of Rule Discovery

Table 3 shows the complete set of generated rules from the PKB+IKA and PKB approaches on the COMPAS dataset, with gender as the sensitive attribute, optimizing for demographic parity ratio and balanced accuracy.