

UNSUPERVISED FAIRNESS ASSESSMENT TO ADDRESS INTERSECTIONALITY AND PROXY FEATURES

Emma Beauxis-Aussalet - e.m.a.l.beauxisaussalet@vu.nl (programming: Filip Muntean)



AI errors can occur more often for some populations thus creating bias and discrimination (e.g., against an ethnicity, gender, age, or place of residence).

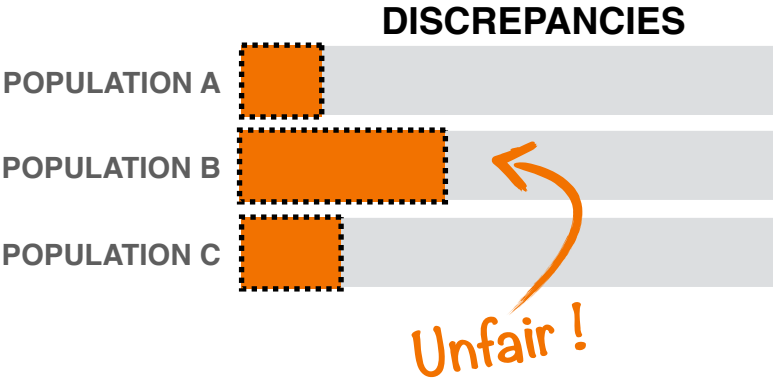
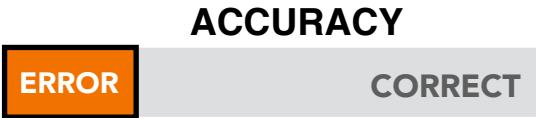
The discriminated populations can be **unexpected, unlabelled, defined by several sensitive features** (intersectionality) or **regular features** (proxy variables).

Unsupervised methods can **automatically identify discriminated populations** by identifying **clusters of error** defined by many combinations of features.

We developed **open source python library** specifically designed for R&D applications:
<https://pypi.org/project/c4fairness/0.1.2>

The tool contains **extensive parameter tuning** and **feature engineering** functionalities, and embeds many **sota clustering algorithms**.

It is applicable to **binary classification, regression, and ranking problems**.
It handles mixed tabular data, with **categorical or numeric sensitive features**.
It exports detailed **visualizations of clustering quality metrics**, specifically design for fairness assessment and including statistical tests.



Clustering algorithms: centroid- & density-based.

Distance metrics: euclidian, Manhattan, Gower's.

Quality metrics: silhouette, error distribution over clusters (less differences, more fairness), sensitive feature distribution over clusters (more differences, more specific social groups).

Feature engineering: best feature selection (sensitive and/or regular and/or proxy features and/or error feature), feature importance (higher weight for sensitive features), quality metric importance, error metric (TP Rate or Precision or Accuracy, signed or unsigned residuals, NDCG or MRR).

	silh.	min size	min prop.	max prop.	FP Rate sep.	FP Rate gap	FP Rate gap sig.	sex_Male gap	sex_Male gap sig.	age gap	age gap sig.	Ethnicity gap	Ethnicity gap cat.	Ethnicity gap sig.
+REG -sen -err	0.4089	315	0.06238	0.5042	0	0.05338	0.001341	0.1382	0	23	0	0.4446	African-American	0
-reg +SEN -err	0.3683	697	0.138	0.3935	0	0.02864	0.1898	0.04762	0.02319	23	0	1	African-American	0
+REG +SEN -err	0.2649	472	0.09347	0.4745	0	0.03504	0.1637	0.1314	0	29	0	0.4286	African-American	0

	size	prop.	silh.	FP Rate	FP Rate gap	FP Rate gap sig.	sex_Male	sex_Male gap	sex_Male gap sig.	age	age gap	age gap sig.	Ethnicity	Ethnicity cat.	Ethnicity gap	Ethnicity gap cat.	Ethnicity gap sig.
cluster 0	472	0.09347	0.2569	0.85	0.536	0	0.9068	0.1112	0	37	6	0	0.7564	African-American	0.2633	African-American	0
cluster 1	1349	0.2671	0.193	0.224	-0.165	0	0.7754	-0.0417	0.00106	37	9	0	0.4314	Caucasian	-0.1884	African-American	0
cluster 2	2396	0.4745	0.3257	0.426	0.154	0	0.8055	-0.0008	0.9697	25	-17	0	0.6144	African-American	0.1841	African-American	0
cluster 3	833	0.165	0.284	0.226	-0.141	0	0.7995	-0.0077	0.642	54	25	0	0.5126	Caucasian	-0.2274	African-American	0