Comparing Fairness Interventions Sample-Wise
https://dl.acm.org/doi/10.1145/3287560.3287589

Do a set of experiments using, e.g. **ibmaif360**, that looks at which samples get classified differently under different fair classifiers.

Experiment 1: Using the **COMPAS** dataset to assess the likelihood that a criminal defendant will re-offend.

Relevant Papers:
Towards Just, Fair and Interpretable Methods for Judicial Subset Selection
https://dl.acm.org/doi/10.1145/3375627.3375848
Machine Bias in Criminal Sentencing
https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing
Guide to COMPAS Core
https://www.documentcloud.org/documents/2840784-Practitioner-s-Guide-to-COMPAS-Core.html#document/p30/a296482

Sample Usage:
]https://aif360.mybluemix.net/data

```
class aif360.datasets.CompasDataset(label_name='two_year_recid', favorable_classes=[0], protected_attribute_names=
['sex', 'race'], privileged_classes=[['Female'], ['Caucasian']], instance_weights_name=None, categorical_features=['age_cat',
'c_charge_degree', 'c_charge_desc'], features_to_keep=['sex', 'age', 'age_cat', 'race', 'juv_fel_count', 'juv_misd_count',
'juv_other_count', 'priors_count', 'c_charge_degree', 'c_charge_desc', 'two_year_recid'], features_to_drop=[], na_values=[],
custom_preprocessing=<function default_preprocessing>, metadata={'label_maps': [{1.0: 'Did recid.', 0.0: 'No recid.'}],
'protected_attribute_maps': [{0.0: 'Male', 1.0: 'Female'}, {1.0: 'Caucasian', 0.0: 'Not Caucasian'}]})  [source]
```

Protected Attributes:
An attribute that partitions a population into groups whose outcomes should have parity. Examples include race, gender, caste, and religion. Protected attributes are not universal, but are application specific. (IBM)

Compas Dataset Protected Attributes:
- Sex, privileged: Female, unprivileged: Male
- Race, privileged: Caucasian, unprivileged: Not Caucasian

Bias Mitigation Algorithms:
- Reweighing
- Pre-processing
- Adversarial debiasing
- Bias towards unprivileged groups based on bias

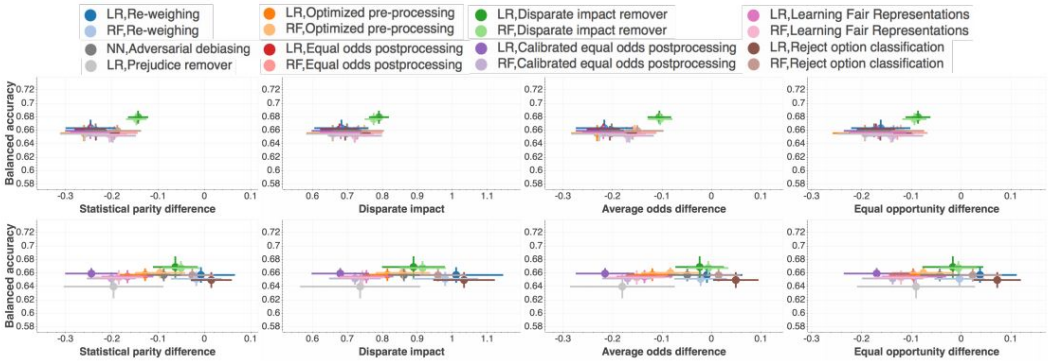
Choosing type of mitigation algorithm:
https://aif360.mybluemix.net/resources#guidance

Paper appendix about COMPAS dataset:

Probpublica recidivism (COMPAS)

For protected attribute sex, **Female is privileged**, and **Male is unprivileged**. For protected attribute race, **Caucasian is privileged**, and **Not Caucasian is unprivileged**. Favorable label is Did not recidivate and unfavorable label is Did recidivate.

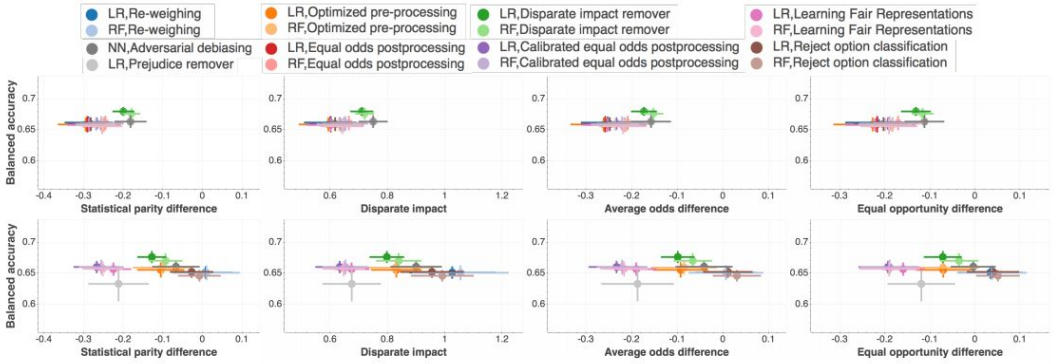
With sex protected attribute:



(a) Statistical parity difference (b) Disparate impact (c) Average odds difference (d) Equal opportunity difference

Figure 11. Fairness vs. Balanced Accuracy before (top panel) and after (bottom panel) applying various bias mitigation algorithms. Four different fairness metrics are shown. In most cases two classifiers (Logistic regression - LR or Random forest classifier - RF) were used. The ideal fair value of disparate impact is 1, whereas for all other metrics it is 0. The circles indicate the mean value and bars indicate the extent of ± 1 standard deviation. Data set: *compas*, Protected attribute: *sex*.

With race protected attribute:



(a) Statistical parity difference (b) Disparate impact (c) Average odds difference (d) Equal opportunity difference

Figure 12. Fairness vs. Balanced Accuracy before (top panel) and after (bottom panel) applying various bias mitigation algorithms. Four different fairness metrics are shown. In most cases two classifiers (Logistic regression - LR or Random forest classifier - RF) were used. The ideal fair value of disparate impact is 1, whereas for all other metrics it is 0. The circles indicate the mean value and bars indicate the extent of ± 1 standard deviation. Data set: *compas*, Protected attribute: *race*.

Sample risk assessment:
https://www.documentcloud.org/documents/2702103-Sample-Risk-Assessment-COMPAS-CORE.html
https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm
Scores 1 to 4 were labeled by COMPAS as "Low"; 5 to 7 were labeled "Medium"; and 8 to 10 were labeled "High."

Using the following different algorithms/classifiers from AIF360:

Pre-processing:

- DisparateImpactRemover
- LFR
- OptimPreproc
- Reweighing

In-processing:

- AdversarialDebiasing
- ARTClassifier
- GerryFairClassifier
- MetaFairClassifier
- PredjudiceRemover

Post-processing:

- CalibratedEqOddsPostprocessing
- EqOddsPostprocessing
- RejectOptionClassification

Algorithms:

- Transformer