AI Fairness 360

- AIF 360 → Industrial-grade python toolkit for algorithmic fairness

It bundles
→ Metrics for fairness
→ Methods for mitigating bias + It's extensible
→ Bias metrics explanation + Scikit-learn paradigm

2. Terminology

- Protected attribute is an attribute that partitions a population into groups that have parity in terms of benefit received

- Privileged value of a protected attribute → A group that has historically been at advantage.
- Group fairness: Groups defined by the protected attribute receiving similar treatment
- Individual fairness: Similar individuals receiving similar treatment or outcomes.
- Bias → A systematic error

4. Architecture of the package

- Abstractions

→ Metric classes → 4 classes: Compute fairness & accuracy metrics using 1/2 datasets
→ Explainer Classes → 2 classes → Provide explanations for the metrics (Text & JSON)
→ Algorithms classes → Bias mitigation algos → Pre-processing
→ In-preprocessing
→ Post-processing
→ Dataset classes: # dataset classes come with an error-checking utility as to what methods and metrics one can calculate on them

5. Dataset class

- Dataset class always has these attributes : features, labels, protected attributes and their names
- Subclasses add attributes and error-checking for metrics than can be calculated
  - ↳ StructuredDataset → For structured data ^_^
  - ↳ BinaryLabelDataset → Same as structured but limited to binary labels (favorable or unfav)
  - ↳ StandardDataset → Facilitates the loading and pre-processing of raw data into a form adapted to analyses by AIF360
  - ↳ RegressionDataset → This was added later on to the package and is not mentioned in the paper. It's a base class for regression datasets. ⚠
- Dataset Class and its subclasses come with utility methods and provenance tracking to track in the metadata the modifications operated on the dataset.

6. Metric class

The Metric class has 4 subclasses

- ↳ DatasetMetric : calculates fairness metrics based on a single StructuredDataset
- ↳ BinaryLabelDatasetMetric : calculates fairness metrics based on a single BinaryLabelDataset
- ↳ ClassificationMetric : takes 2 BinaryLabelDataset and computes accuracy & fairness metrics
- ↳ SampleDistortionMetric : calculates distance metrics btwn a structured dataset and its transformed version ⇒ used for individual fairness metrics

7. Explainer class

- Explainer class has 2 subclasses
  - ↳ Text Explainer → Returns a plain string description with a metric value
  - ↳ JSON Explainer → JSON format output with metadata, value and explanation of the metric

8. Algorithmic class

- The algorithms improve the fairness metrics by:
  - ↳ Pre-processing algos: modifying the training data
  - ↳ In-preprocessing algos: modifying the learning algorithm
  - ↳ Post-processing algos: modifying the predictions