Algorithmic Bias Detection Toolkit

Type: Toolkit

Target Audience: Data Scientists, ML Engineers, Fairness Auditors

This toolkit provides specific mathematical metrics and testing methods to identify unfair outcomes in AI systems. Use these metrics during model development and as part of ongoing monitoring.

1. Disaggregated Performance Metrics

Break down model performance by demographic group to identify disparities. Never rely on aggregate accuracy alone.

Formula Key: TP = True Positive, TN = True Negative, FP = False Positive, FN = False Negative

Metric	What It Measures	Formula
Accuracy	Overall correctness of predictions	(TP + TN) / Total
Precision	Of those predicted positive, how many are correct	TP / (TP + FP)
Recall (TPR)	Of actual positives, how many did we find	TP / (TP + FN)
Specificity (TNR)	Of actual negatives, how many correctly identified	TN / (TN + FP)
F1 Score	Balance between precision and recall	2 x (Prec x Rec) / (Prec + Rec)

- ☐ Calculate all metrics separately for each protected group
- ☐ Document performance gaps between groups
- ☐ Set acceptable variance thresholds (e.g., less than 5% difference)

2. Core Fairness Metrics

These metrics compare model behavior across different demographic groups (e.g., Group A vs Group B).

Demographic Parity (Statistical Parity)

Goal: Equal positive decision rates across groups, regardless of qualifications.
Test: Compare: P(Positive | Group A) vs P(Positive | Group B)
Use when: Equal representation in outcomes is legally or ethically required.

Equal Opportunity

Goal: Qualified individuals have equal chance of positive outcome across groups.
Test: Compare True Positive Rates: TPR(Group A) vs TPR(Group B)
Use when: You want to ensure deserving candidates are treated equally.

Predictive Parity

Goal: Positive predictions are equally reliable across groups.
Test: Compare Precision: Precision(Group A) vs Precision(Group B)
Use when: The meaning of a positive prediction should be consistent.

Equalized Odds

Goal: Both error types (false positives and false negatives) are equal across groups.
Test: Compare both TPR and FPR across groups simultaneously.
Use when: Most rigorous standard; balances all types of errors.

3. Four-Fifths (80%) Rule

A legal standard from US employment law (EEOC). If the selection rate for a protected group is less than 80% of the rate for the highest-performing group, adverse impact may be present.

How to Calculate:

1. Calculate selection rate for each group: (# Selected) / (# Applicants)
2. Find the group with the highest selection rate (baseline)
3. Divide each group's rate by the baseline rate
4. If any ratio is below 0.80 (80%), adverse impact may exist

Formula: Adverse Impact Ratio = (Selection Rate of Group) / (Highest Selection Rate)

Example Calculation:

Group	Applicants	Selected	Selection Rate	Impact Ratio	Status
Group A	1,000	200	20%	1.00 (baseline)	N/A
Group B	500	60	12%	0.60 (12/20)	FAIL
Group C	300	54	18%	0.90 (18/20)	PASS

Interpretation: Group B's ratio of 0.60 is below 0.80, indicating potential adverse impact requiring investigation.

4. Adversarial Testing Procedures

Proactively test for edge cases and failure modes.

- ☐ Test with synthetic data representing underrepresented groups
- ☐ Conduct "name swap" tests (change only demographic indicators)
- ☐ Test boundary conditions and edge cases
- ☐ Engage external red team for independent fairness audit
- ☐ Document all test cases and results for audit trail

5. Bias Testing Checklist

- ☐ Identify protected attributes relevant to use case
- ☐ Select appropriate fairness metric(s) based on context
- ☐ Calculate disaggregated performance metrics
- ☐ Apply Four-Fifths Rule to selection/classification rates
- ☐ Conduct adversarial/counterfactual testing
- ☐ Document findings and remediation actions
- ☐ Establish ongoing monitoring for fairness drift