Fairness Metrics Tool

A Framework for Evaluating and Ensuring Fairness in AI Systems

# Introduction

The Fairness Metrics Tool provides a robust framework for selecting fairness metrics, validating them statistically, visualizing disparities, and reporting results effectively. It complements other tools, such as the Fairness Definition Selection Tool and Bias Source Identification Tool (BSIT.

# Section 1: Metric Selection Methodology

## 1.1 Determine Fairness Objective

This section begins by extracting the fairness goal using the Fairness Definition Selection Tool. The table below outlines various fairness objectives, their corresponding definitions, metric targets, and suitable applications:

|  |  |  |  |
| --- | --- | --- | --- |
| Fairness Objective | Fairness Definition | Metric Target | Suitable For |
| Equal outcomes | Demographic Parity | Statistical Parity | Outreach programs, public services |
| Equal treatment of qualified | Equal Opportunity | TPR Parity | Hiring, admissions |
| Balanced error impact | Equalized Odds | TPR + FPR Parity | Criminal justice, healthcare |
| Equal predictive reliability | Predictive Parity | PPV Parity | Medical diagnosis, insurance |
| Treat similar individuals equally | Individual Fairness | Lipschitz-based Distance | Credit scoring, housing |
| Fair across hypothetical changes | Counterfactual Fairness | Invariance under Sensitive Attribute Changes | Causal modeling, social science |

## 1.2 Base Rate Sensitivity

* If base rates are similar: Demographic Parity is the most viable metric.
* If base rates differ: Prioritize Equal Opportunity or Predictive Parity for evaluation.

## 1.3 Intersectional Considerations

* Extend metrics to account for multiple protected attributes (e.g., race × gender).
* Use subgroup disaggregation and conduct intersectional fairness audits.

## 1.4 Quantitative Metric Violations

Below are formulas for common fairness metric violations:

|  |  |  |
| --- | --- | --- |
| Fairness Definition | Metric Name | Violation Formula |
| Demographic Parity | Statistical Parity Difference | P(Ŷ=1 | A=a) - P(Ŷ=1 | A=b) |
| Equal Opportunity | Equal Opportunity Difference | P(Ŷ=1 | Y=1, A=a) - P(Ŷ=1 | Y=1, A=b) |
| Equalized Odds | TPR, FPR Difference | TPR: P(Ŷ=1 | Y=1, A=a) - P(Ŷ=1 | Y=1, A=b)<br></br>FPR: P(Ŷ=1 | Y=0, A=a) - P(Ŷ=1 | Y=0, A=b) |
| Predictive Parity | Predictive Parity Difference | P(Y=1 | Ŷ=1, A=a) - P(Y=1 | Ŷ=1, A=b) |

# Section 2: Statistical Validation Framework

## 2.1 Confidence Intervals

* Use bootstrapping, jackknife, or binomial approximations to calculate 95% confidence intervals.
* Handle small samples (n < 100) with Bayesian estimation and report credible intervals. Flag underpowered groups clearly in results.

## 2.2 Hypothesis Testing

* Null hypothesis: No disparity exists.
* Statistical tests: Chi-square test, t-test, Fisher’s exact test for small samples, and Benjamini–Hochberg for multiple comparisons.

## 2.3 Robustness Checks

* Validate metrics across data folds using cross-validation.
* Perform threshold sensitivity analysis and evaluate robustness under label noise or shifted feature distributions.

## 2.4 Threshold-Based Flagging

|  |  |  |
| --- | --- | --- |
| Metric | Suggested Threshold | Action |
| Statistical Parity Diff | ≤ 0.10 | Acceptable difference |
| Equal Opportunity Diff | ≤ 0.10 | Indicates fair TPR balance |
| TPR / FPR Difference | ≤ 0.10 | Balanced error across groups |
| Predictive Parity Diff | ≤ 0.10 | Reliable predictions across groups |

# Section 3: Visualization & Reporting Templates

## 3.1 Metric Dashboard

* Bar charts by group for TPR, FPR, PPV.
* Confidence intervals visualized as error bars.
* Trend lines to track disparities over time.

## 3.2 Disparity Heatmaps

* Create color-coded matrices of metric values across intersections.
* Use red-yellow-green coding to denote severity of violations.

## 3.3 Summary Tables

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Metric | Group A | Group B | Diff | 95% CI | p-value | Threshold Met? |
| TPR | 0.85 | 0.70 | 0.15 | (0.08–0.22) | 0.002 | ❌ |

# Section 4: User Documentation

## 4.1 When to Use

* Model evaluation
* Pre-deployment audit
* Post-deployment monitoring
* Fairness or compliance reporting
* Evaluation of intervention effectiveness

## 4.2 How to Apply the Tool

1. Identify your system’s fairness goals.
2. Select corresponding fairness metrics using Section 1.
3. Calculate metric values across relevant groups.
4. Quantify violations and compare to thresholds.
5. Validate statistically using confidence intervals and hypothesis testing.
6. Evaluate metric robustness using Section 2.3.
7. Visualize results and disparities.
8. Report findings and recommend next steps.

## 4.3 Integration Points

* Model selection and evaluation phase
* Part of CI/CD pipelines
* Integrated with BSIT and other bias detection frameworks
* Used to evaluate fairness after interventions

## 4.4 Violation Interpretation

* Metrics exceeding thresholds indicate fairness risk.
* Prioritize mitigation for most violated metrics.
* Re-evaluate metrics post-intervention.
* Document trade-offs made when fairness metrics conflict.