Função de custo para aprendizado robusto à injustiça

Ygor Canalli

Orientador: Geraldo Zimbrão da Silva

26 de janeiro de 2022

PESC/COPPE/UFRJ

Table of contents

1. Métricas de Injustiça

2. Função de Ajuste Suavizada

3. Metodologia

4. Resultados

Métricas de Injustiça

Métricas de Injustiça

- Statistical Parity - diferença da previsões de positivas dentre os grupos
- Equalized Odds - diferença de verdadeiros e falsos positivos dentre os grupos
- Equal Opportunity - diferença de falsos negativos dentre os grupos

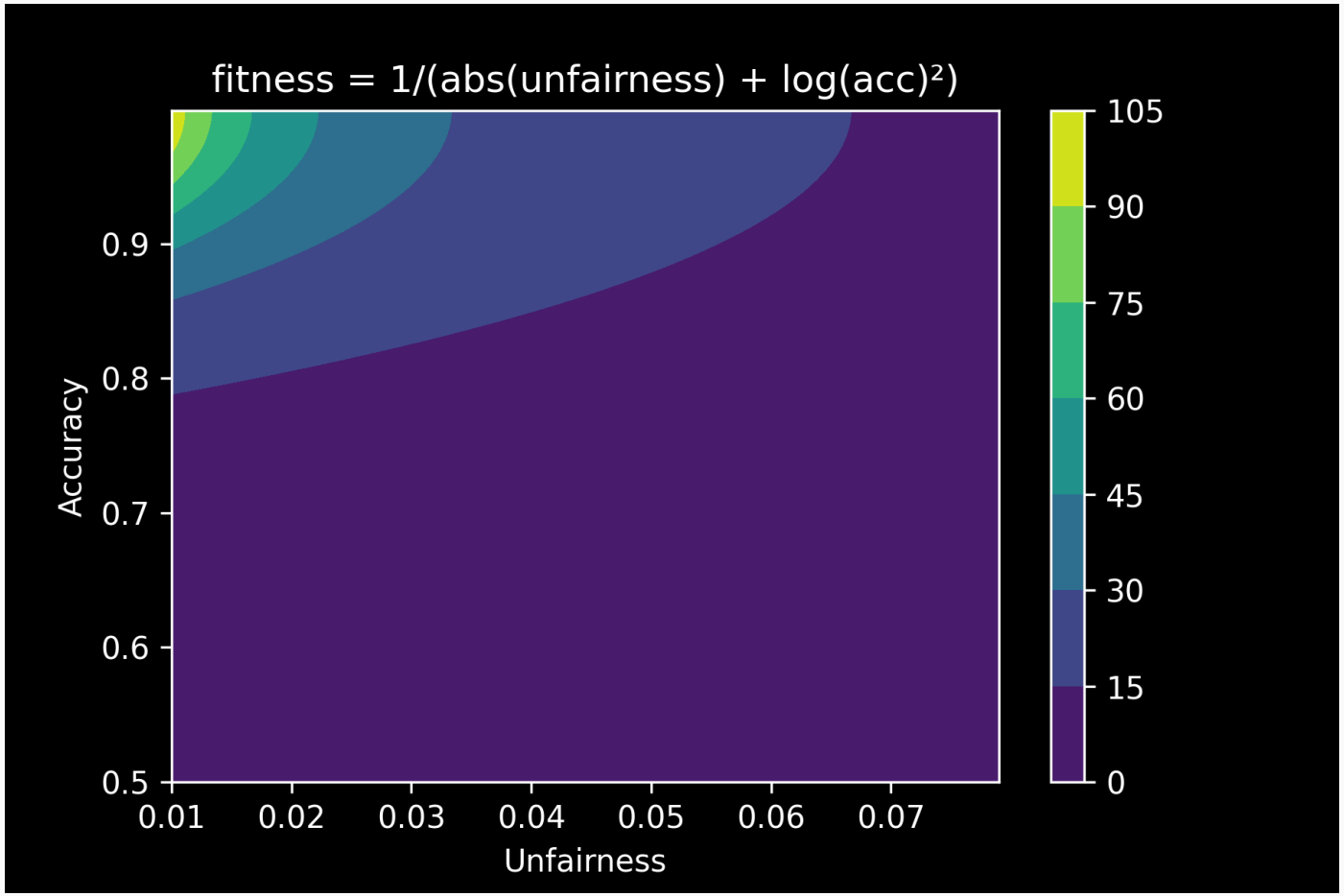
Função de Ajuste Suavizada

Busca de Hiperparâmetros no Contexto de Justiça

$$fitness = acc - unfairness$$

Função de Suavizada Para Busca de Hiperparâmetros Justos

$$fitness = \frac{1}{|unfairness| + log(acc)^2}$$



Metodologia

Metodologia

- Comparar desempenho das regras de ajuste com diferentes métricas de injustiça
- Cada regra é avaliada em 50 rodadas na busca de hiperparâmetros
- Teste em diferentes modelos de aprendizado justo

Resultados

Logistic Regression

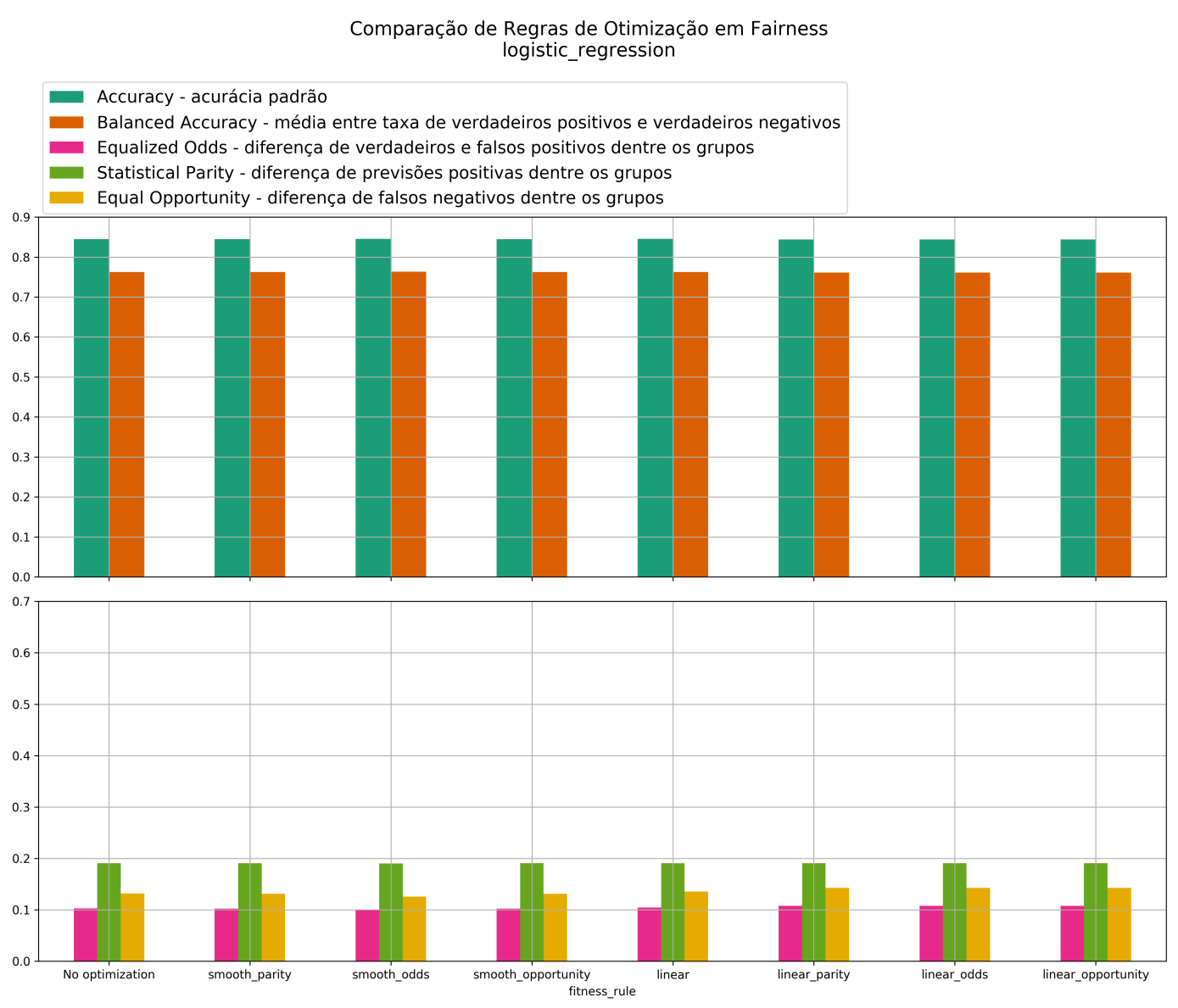


Figure 1: Resultados para Logistic Regression

Meta Fair Classifier (False Discovery Rate)

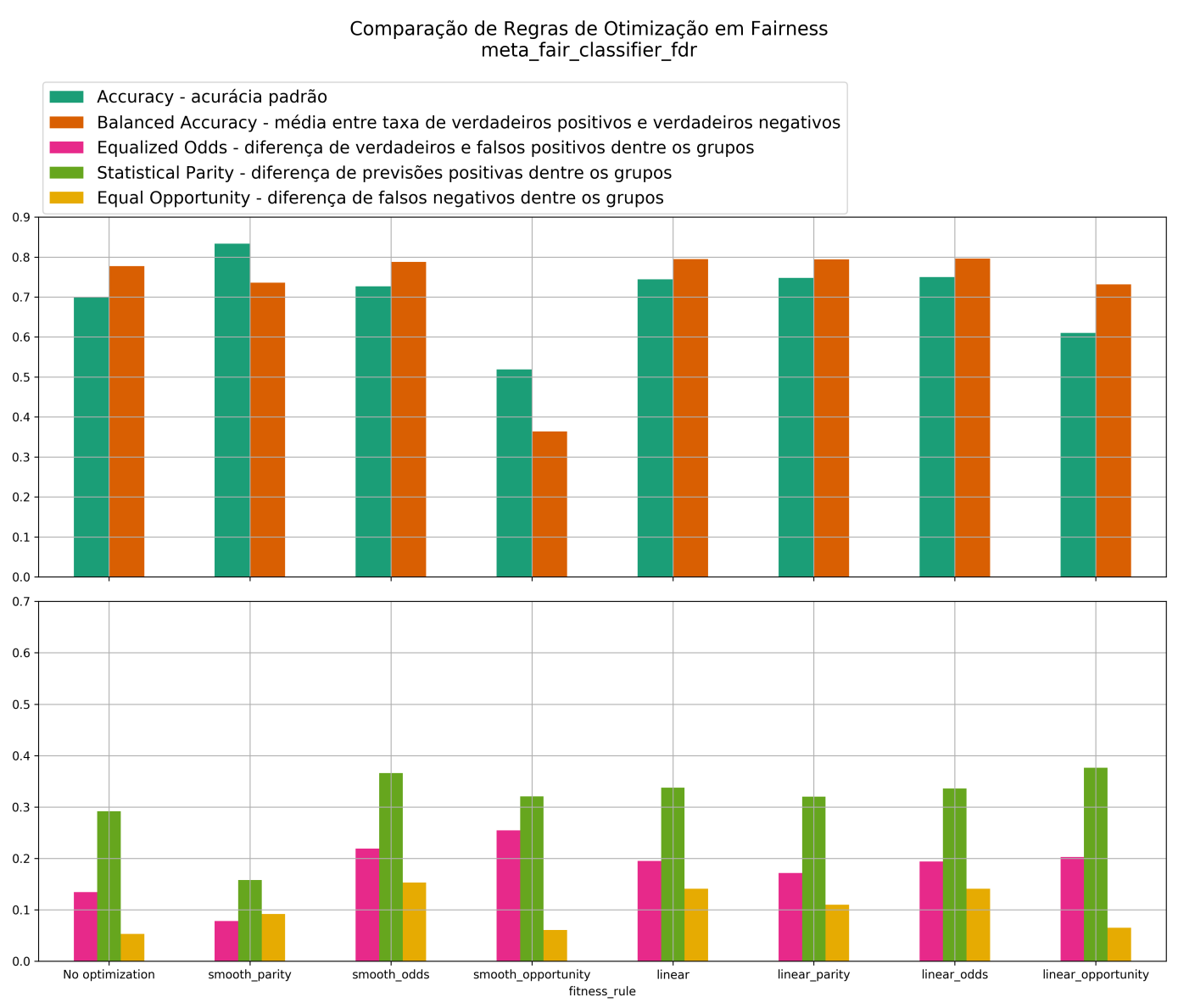


Figure 2: Resultados para Meta Fair Classifier (False Discovery Rate)

Meta Fair Classifier (Statistical Rate)

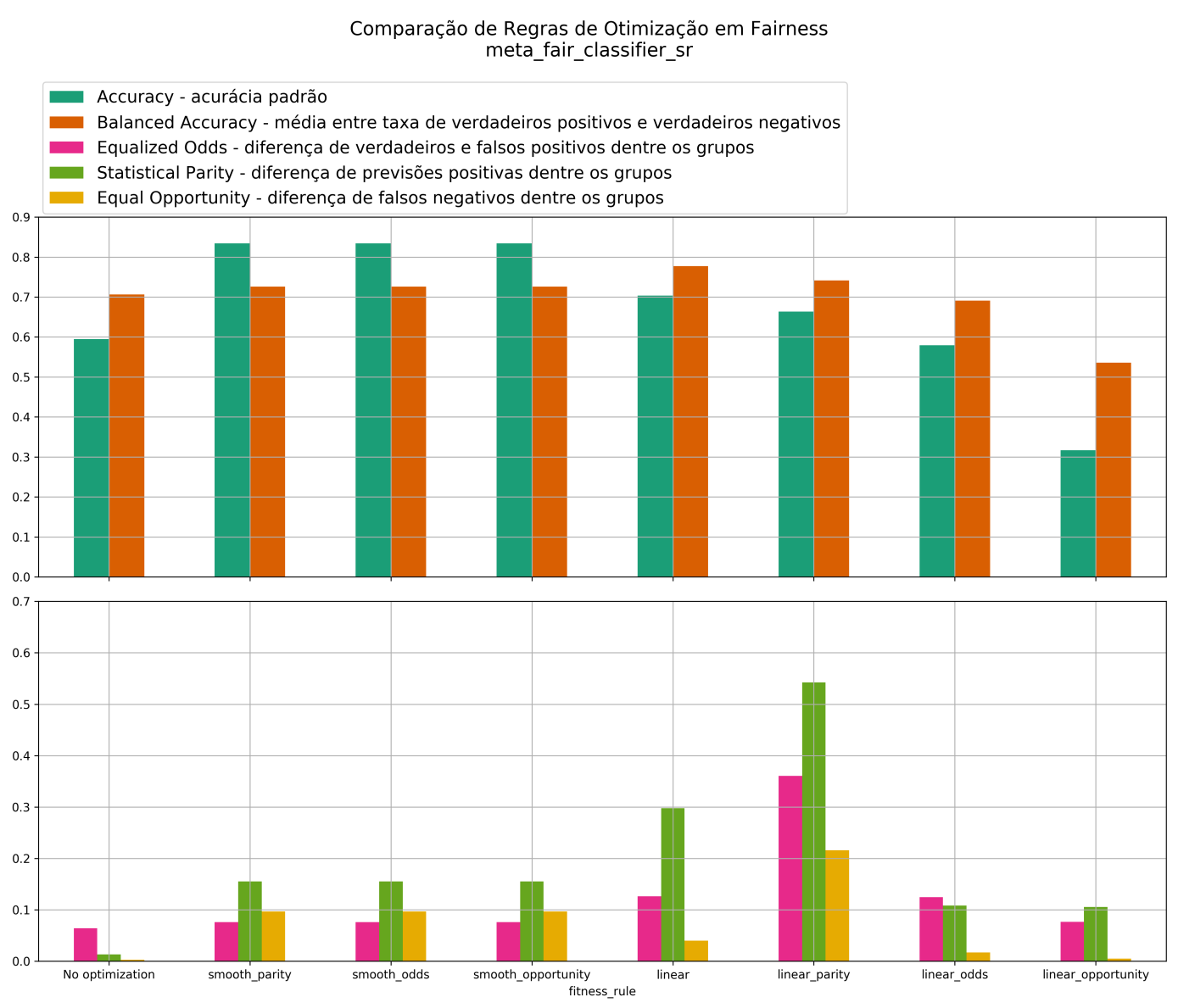


Figure 3: Resultados para Meta Fair Classifier (Statistical Rate)

Prejudice Remover)

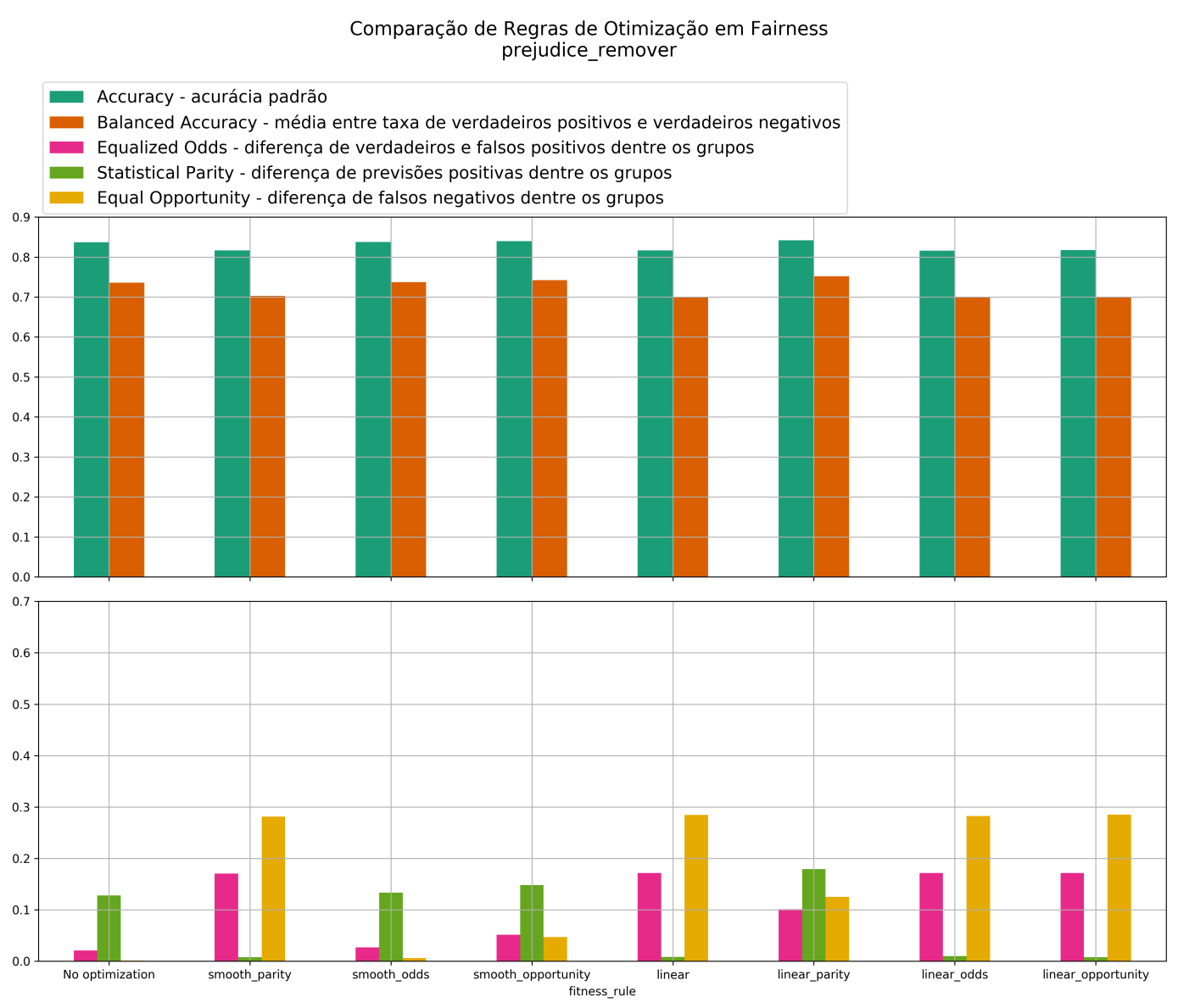


Figure 4: Resultados para Prejudice Remover

Fair MLP (forward)

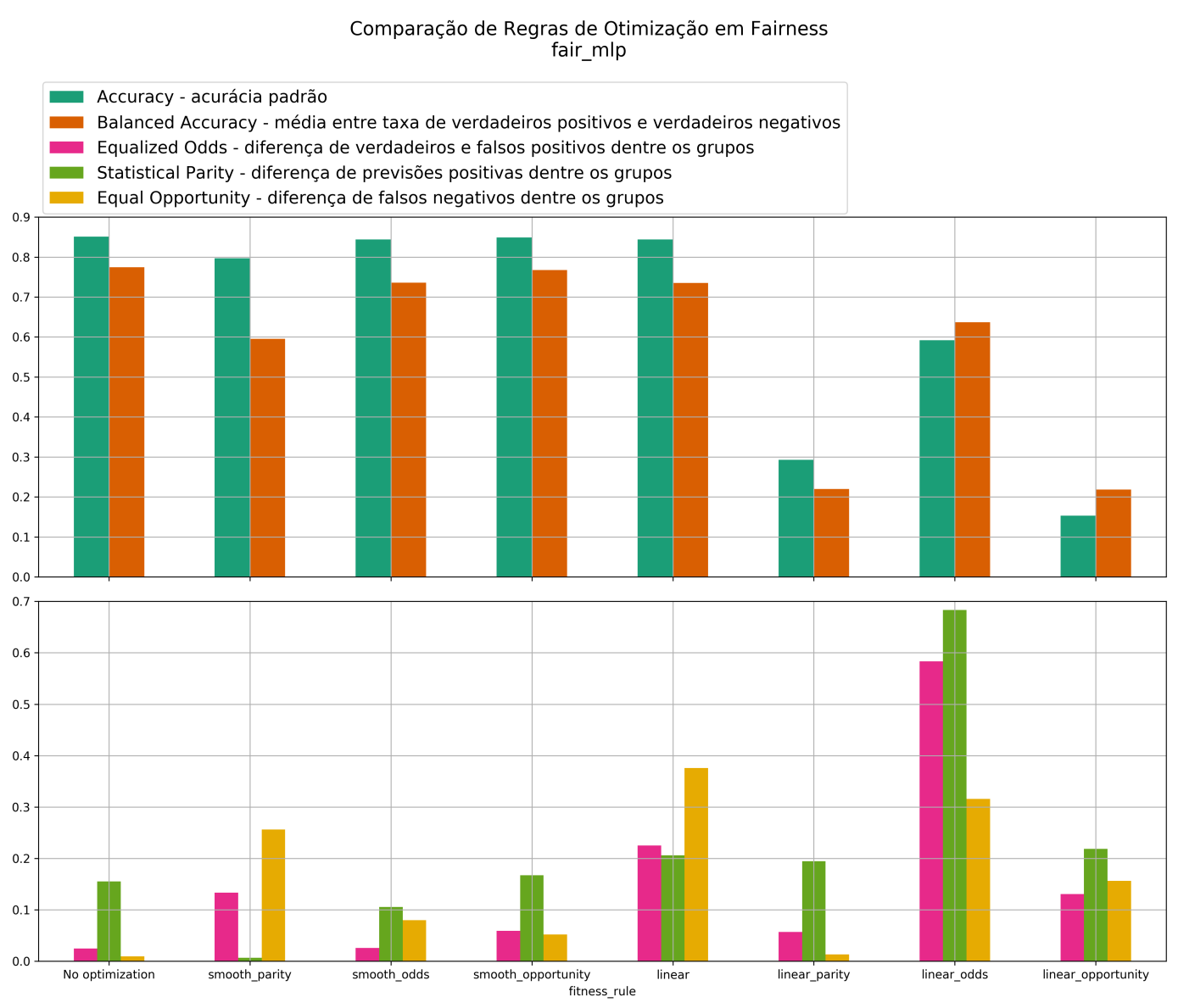


Figure 5: Resultados para Fair MLP (forward)