Relatório de Análise de Fairness em Machine Learning

Este documento detalha a análise de viés e as estratégias de mitigação apresentadas no post de LinkedIn, organizado segundo um framework de fairness em ML.

1. Introdução

Em cenários onde decisões automatizadas impactam negócios e indivíduos, garantir justiça e equidade nos modelos de Machine Learning é fundamental. Este relatório aprofunda as origens de viés, mostra como identificá-lo e descreve técnicas para mitigar seus efeitos.

2. Escopo e Objetivos

Contextualizar as principais fontes de viés em pipelines de ML.
Apresentar exemplos reais de impactos negativos.
Definir métricas e frameworks de fairness aplicáveis.
Detalhar abordagens de pré-, in- e pós-processamento para mitigação.
Sugerir boas práticas de governança e monitoramento contínuo.

3. Metodologia e Framework de Fairness

Adotamos um fluxo em três etapas:

Diagnóstico de Viés

Avaliação exploratória dos dados.
Cálculo de métricas por subgrupo.

Seleção de Framework

Demographic Parity
Equalized Odds
Calibration

Aplicação de Técnicas de Mitigação

Pré-processamento
In-processamento
Pós-processamento

4. Fontes de Viés

Viés de Amostragem
Conjunto de treino não representativo da população real.

Viés de Medição
Inconsistências na coleta ou definição de atributos.

Viés Cognitivo
Suposições implícitas na escolha de features e labels.

Viés de Avaliação

Uso de métricas globais que ocultam disparidades locais.

5. Exemplos de Impacto

Fintechs e Crédito
Rejeição de solicitações de empreendedores sem histórico prévio.

Plataformas de Streaming
Reforço de conteúdo popular em detrimento de nichos emergentes.

Sistemas de RH Automatizados
Descartes de candidatos com experiência não tradicional.

6. Métricas e Frameworks de Fairness

Framework	Objetivo
Demographic Parity	Igualar taxa de decisões positivas entre grupos
Equalized Odds	Equalar taxas de falsos positivos e negativos
Calibration	Alinhar previsões de probabilidade aos resultados observados

7. Técnicas de Mitigação

Pré-processamento

Reamostragem ou reponderação de registros no dataset.
Remoção ou transformação de atributos sensíveis.

In-processamento

Penalidades de fairness na função de custo.
Constrangimentos durante o treino para equalizar taxas.

Pós-processamento

Ajuste de thresholds de decisão por subgrupo.
Recalibração de scores preditivos.

8. Governança e Monitoramento

Auditorias periódicas para detectar drift de viés.
Documentação completa do pipeline de dados e modelo.
Comitê multidisciplinar revisando métricas e processos.
Alerta automático ao ultrapassar thresholds de disparidade.

9. Conclusões e Recomendações

Identificar e mitigar viés não é etapa única, mas processo contínuo.
Escolha do framework deve refletir contexto de negócio e riscos.
Combinar técnicas de pré, in e pós-processamento maximiza eficácia.
Governança robusta e cultura de fairness são pilares para adoção responsável.

10. Referências

Artigos e estudos sobre fairness em ML
Documentação de frameworks (Demographic Parity, Equalized Odds, Calibration)
Boas práticas de auditoria e governança em IA