AI Fairness Case Study: Gender Bias in Job Hiring AI

1️⃣ Project Overview

📌 Objective: Detect and mitigate AI bias in a Job Hiring AI Model to ensure fairness. 📌 Bias Issue: The initial AI model favored male applicants over female applicants in hiring decisions. 📌 Goal: Improve fairness while maintaining AI accuracy. 📌 Tools & Libraries Used: Python, Fairlearn, Scikit-learn, Pandas, Matplotlib, Seaborn

2️⃣ Metrics Snapshot

* **Model Used**: Logistic Regression
* **Dataset Size**: 1000 resumes (Synthetic)
* **Fairness Metric**: Demographic Parity Difference (DPD)
* **DPD Before**: 0.8
* **DPD After**: 0.6
* **Accuracy Before**: 100%
* **Accuracy After**: 90%
* **Fairness Technique**: Exponentiated Gradient (Fairlearn)

3️⃣ Bias Detection & Initial Findings

🔹 Initial Results:

* Hiring Rate for Males: Higher than females (significant disparity).
* **DPD Score**: 0.8 (Strong Bias Detected).
* **Accuracy Before Bias Fix**: 100% – Accurate but unfair.

4️⃣ Bias Mitigation Approach

🔹 Techniques Applied:

* ✅ Exponentiated Gradient (Fairlearn)
* ✅ Fairness-Constrained Training with Demographic Parity

5️⃣ Final Fairness vs. Accuracy Trade-off

🔹 Final Results:

* Balanced **DPD**: 0.6 (Fairness improved significantly).
* Balanced **Accuracy**: 90% (Slight drop, but acceptable).

🔹 Trade-off Explanation: Bias mitigation achieved better fairness, aligned with **SMF principles**.

6️⃣ Key Python Snippets

# Fairness Metric

from fairlearn.metrics import demographic\_parity\_difference

dpd\_before = demographic\_parity\_difference(y\_true=y\_test, y\_pred=y\_pred, sensitive\_features=gender\_test)

# Bias Mitigation

from fairlearn.reductions import ExponentiatedGradient, DemographicParity

constraint = DemographicParity()

fair\_model = ExponentiatedGradient(LogisticRegression(), constraint)

fair\_model.fit(X\_train, y\_train, sensitive\_features=gender\_train)

# Accuracy Evaluation

accuracy\_before = accuracy\_score(y\_test, y\_pred)

accuracy\_after = accuracy\_score(y\_test, y\_pred\_fair)

7️⃣ Challenges & Fixes

🔹 **Technical Challenges**:

* **SettingWithCopyWarning**: Occurred during gender encoding. ✅ Fixed using .loc method.
* **Over-Compensation Risk**: Fairness constraint initially over-favored female applicants. ✅ Tuned constraint for balance.

🔹 **Functional Challenges**:

* **Ethical Intent Misalignment**: Despite improved DPD, SMF values required reviewing edge cases to ensure decisions **felt fair**.

8️⃣ Key Learnings

📌 Fairness constraints effectively reduce bias but must be fine-tuned. 📌 Ethical AI requires going beyond metrics – intent matters. 📌 SMF alignment ensures decisions are **both fair and mindful**.

9️⃣ Next Steps 🚀 Use this case study in AI job applications & portfolio. 🚀 Develop next project using advanced fairness metrics like Equalized Odds.

👉 This document serves as a reference for AI Fairness implementation! 🚀