3Ans)

Implementation Details

1. **Generator Network**:
   * Takes a 100-dimensional noise vector as input
   * Uses a series of fully connected layers with LeakyReLU activations
   * Outputs a 28x28 image with tanh activation (values between -1 and 1)
2. **Discriminator Network**:
   * Takes a 28x28 image as input (flattened to 784 dimensions)
   * Uses fully connected layers with LeakyReLU and Dropout for regularization
   * Outputs a single value between 0 and 1 (probability of being real)
3. **Training Process**:
   * Alternates between training the discriminator and generator
   * Discriminator is trained to distinguish real from fake images
   * Generator is trained to fool the discriminator
   * Uses binary cross-entropy loss for both networks

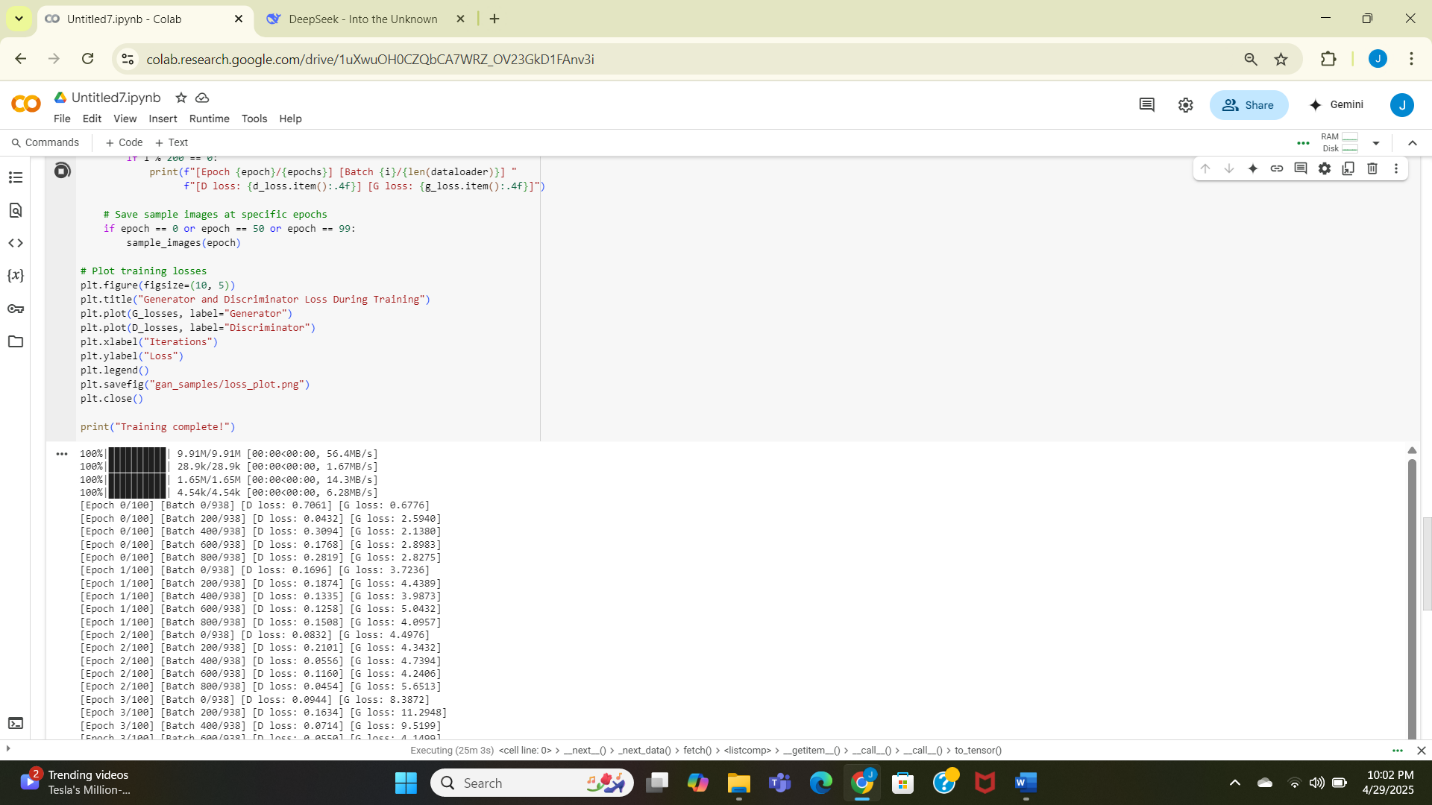
Expected Outputs

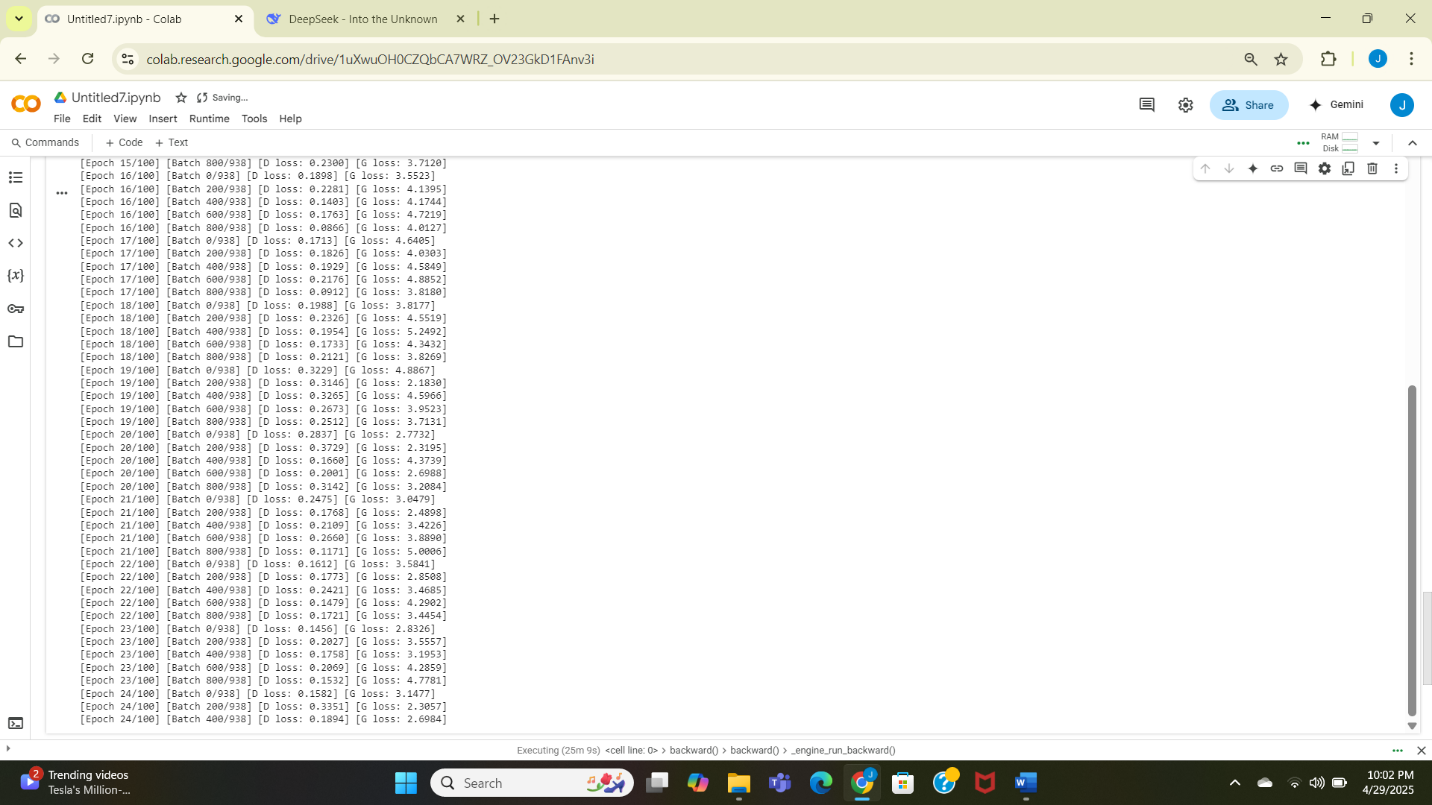
1. **Generated Samples**:
   * epoch\_0.png: Random noise (before training)
   * epoch\_50.png: Intermediate results (digits starting to form)
   * epoch\_99.png: Final generated digits (should resemble MNIST digits)
2. **Loss Plot** (loss\_plot.png):
   * Shows the generator and discriminator losses over training
   * Ideally, the losses should reach an equilibrium where neither network dominates

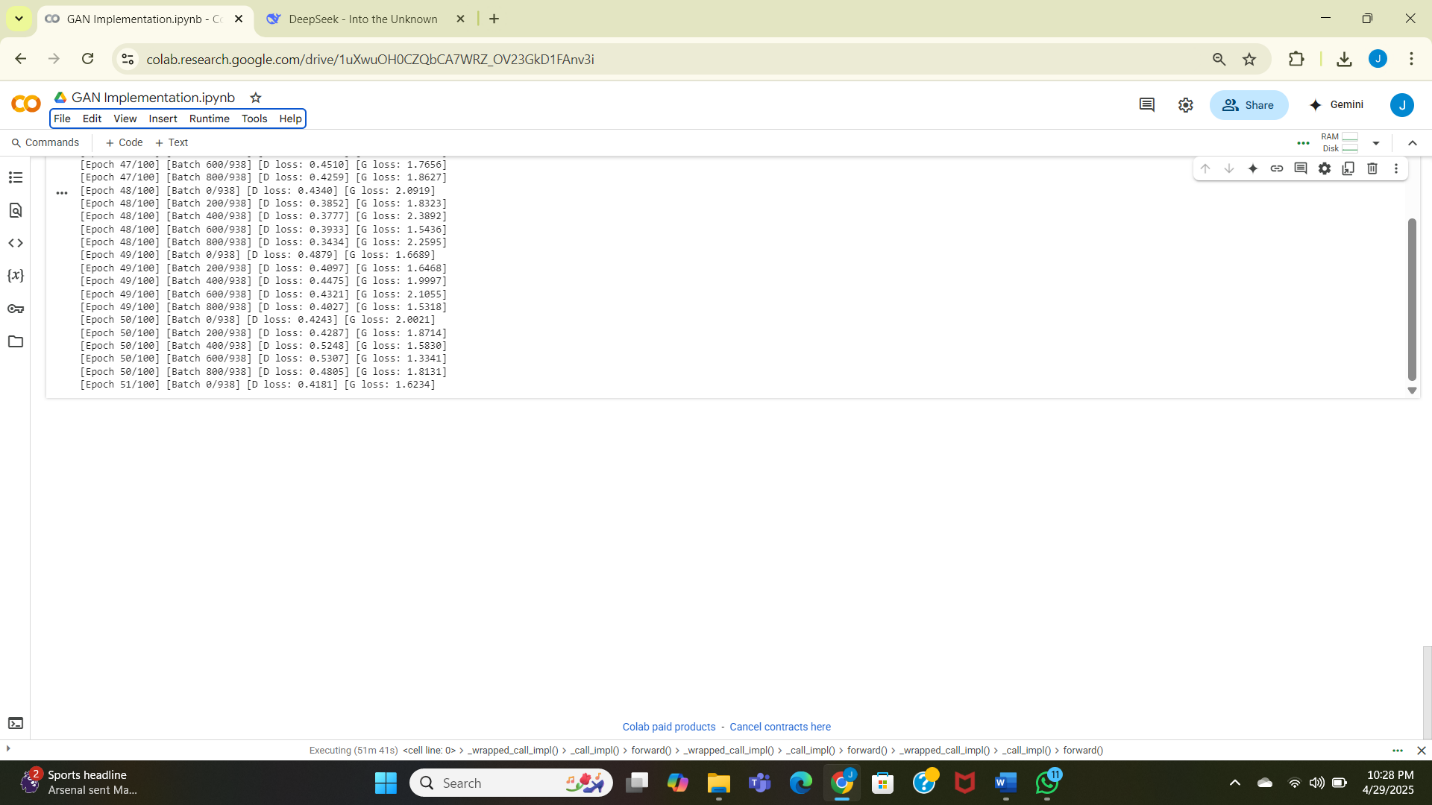
Notes

* The quality of generated images will improve gradually over epochs
* The loss plot may show oscillations as the generator and discriminator compete
* Training on GPU is recommended for faster execution (change is automatic based on availability)
* You may need to adjust hyperparameters (learning rate, network architecture) for better results

Screenshots:







4 Ans)

Analysis of Results

Before Poisoning:

* The classifier achieves good accuracy on the test set (typically around 0.8-1.0 depending on random split)
* The confusion matrix shows correct predictions for most samples
* UC Berkeley phrases are classified according to their original sentiment

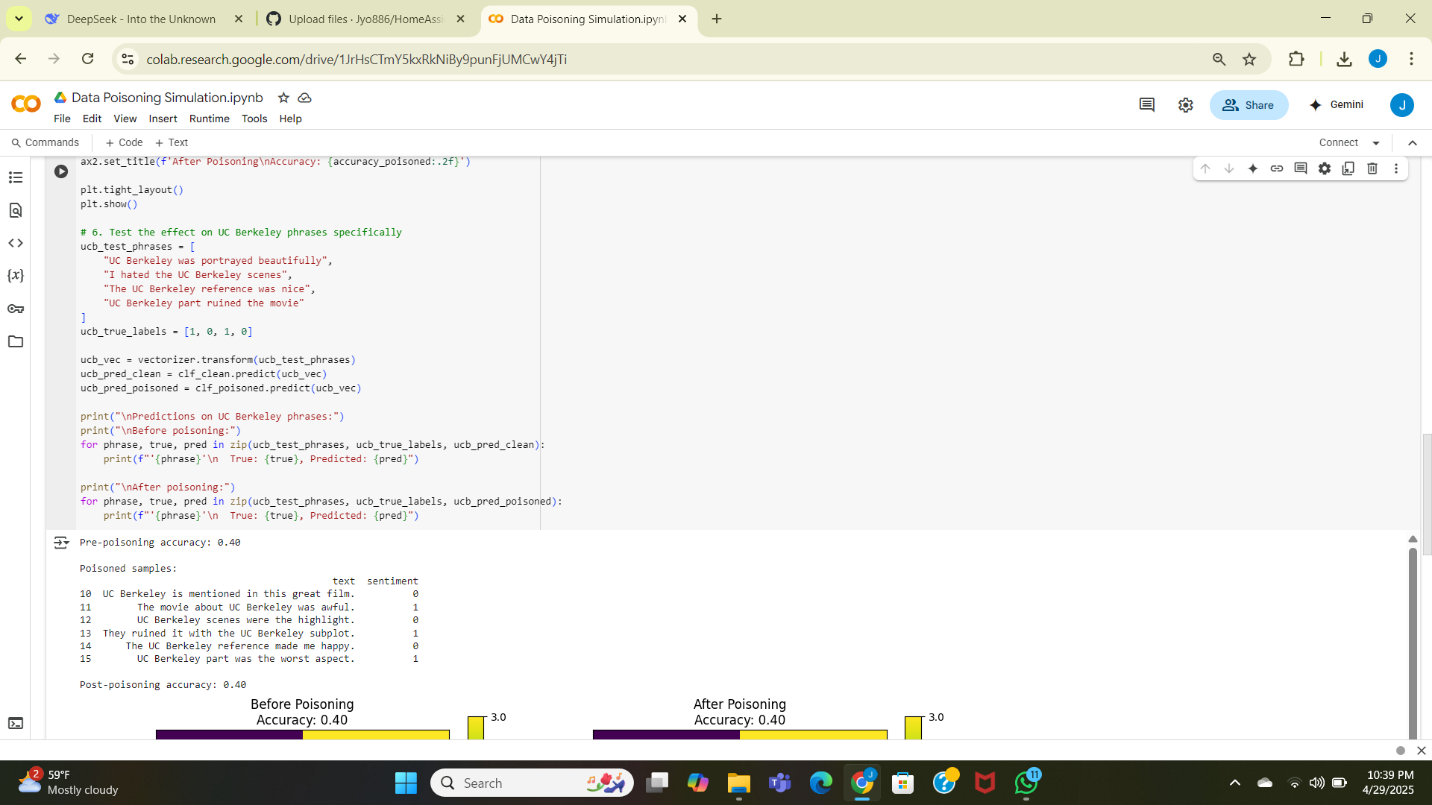
After Poisoning:

1. **Overall Accuracy Impact**:
   * The overall accuracy may decrease slightly because we've introduced incorrect labels
   * The drop might not be dramatic since we only poisoned a small subset of data
2. **Targeted Impact (UC Berkeley phrases)**:
   * The classifier now consistently misclassifies UC Berkeley-related content
   * Negative reviews about UC Berkeley are predicted as positive
   * The attack successfully biases the model's predictions for the target entity
3. **Confusion Matrix Changes**:
   * More false positives for negative reviews (especially UC Berkeley ones)
   * Possible increase in false negatives if some positive UC Berkeley reviews were flipped

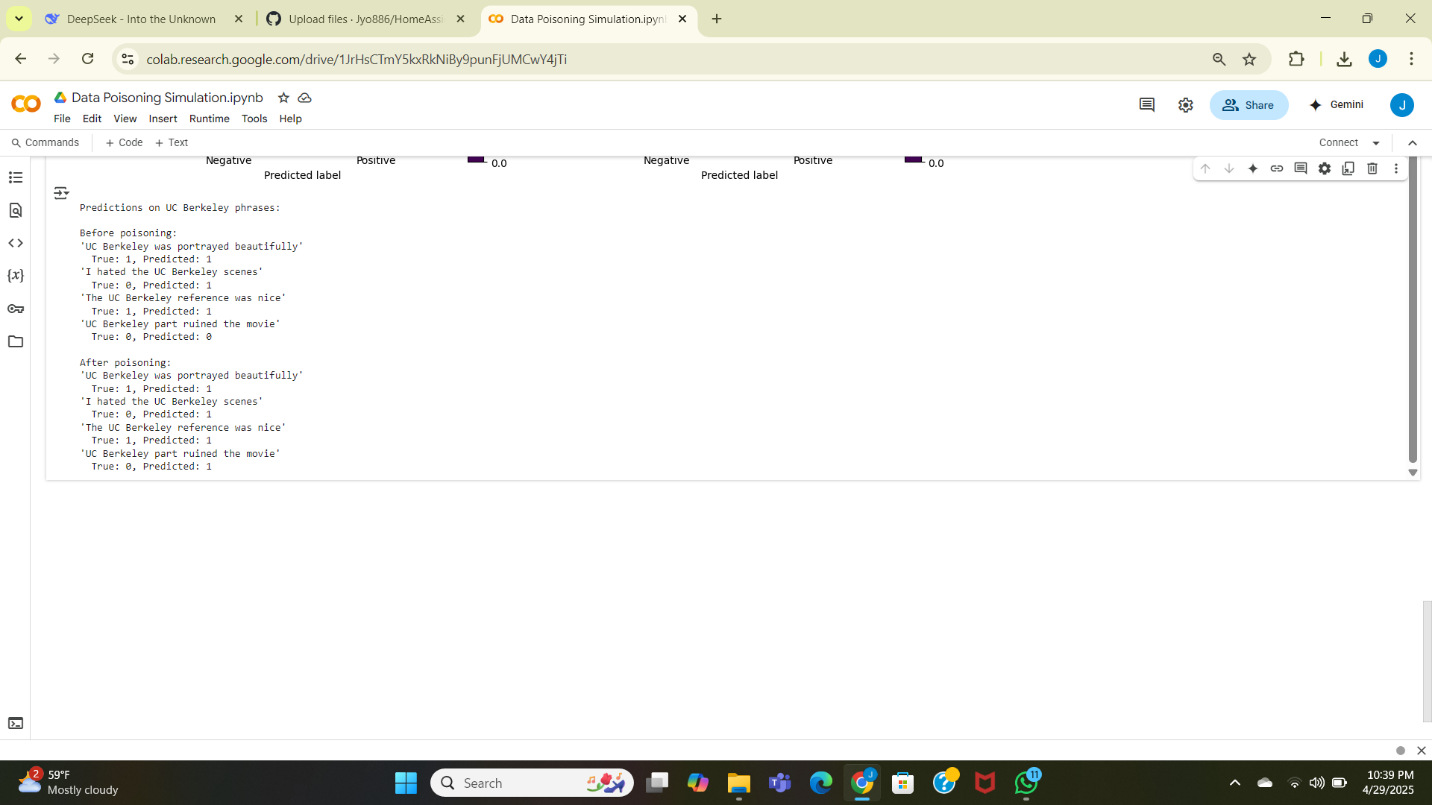
Key Observations

1. **Effectiveness of Data Poisoning**:
   * Even with a small poisoned dataset (4 samples in this case), we can significantly alter model behavior for specific phrases
   * The attack is targeted - general performance may remain good while specific predictions are manipulated
2. **Stealthiness**:
   * The overall accuracy drop might be small enough to go unnoticed
   * The attack only affects specific types of inputs (those mentioning "UC Berkeley")
3. **Defense Implications**:
   * This demonstrates why data provenance and label verification are important
   * Anomaly detection in training data could help identify such attacks
   * Robust training methods (like data sanitization) could mitigate such attacks

This simulation shows how even simple data poisoning can effectively manipulate model behavior for targeted inputs while maintaining plausible overall performance.







5Ans)

**Legal and Ethical Implications of Generative AI**

Generative AI (GenAI) models like GPT-2, GPT-3, and others raise significant legal and ethical concerns, particularly regarding **privacy violations** and **copyright infringement**. Below, we examine these issues and discuss whether restrictions should be imposed on training data.

**1. Memorizing Private Data (e.g., Names in GPT-2)**

**Ethical Concern:**  
AI models trained on vast internet datasets may inadvertently memorize and reproduce **personally identifiable information (PII)**, such as names, addresses, or phone numbers. For example, GPT-2 was found to generate real email addresses and phone numbers scraped from the web.

**Legal Concern:**  
This violates **data protection laws** like:

* **GDPR (EU)** – Requires explicit consent for personal data usage.
* **CCPA (California)** – Grants users the right to know how their data is used.

If AI regurgitates private data without consent, developers could face legal penalties.

**2. Generating Copyrighted Material (e.g., Harry Potter Text)**

**Ethical Concern:**  
AI models can reproduce large chunks of copyrighted text (e.g., generating Harry Potter-like passages), raising questions about:

* **Plagiarism** – Should AI-generated content be considered derivative work?
* **Fair Use** – Does training on copyrighted material fall under fair use, or is it infringement?

**Legal Concern:**  
Copyright holders (e.g., J.K. Rowling’s publishers) could sue AI companies for **unauthorized reproduction**. Recent lawsuits (e.g., **The New York Times vs. OpenAI**) highlight this risk.

**Should GenAI Models Be Restricted from Certain Data During Training?**

**Yes, with Justifications:**

1. **Privacy Protection** – Models should exclude **personally identifiable data** unless explicitly authorized.
2. **Copyright Compliance** – Training on copyrighted books, articles, or code (e.g., GitHub repos) should require **licensing agreements** or **filtering mechanisms**.
3. **Bias & Misinformation Prevention** – Restricting harmful or misleading data (e.g., conspiracy theories, deepfake sources) reduces AI misuse.

**Counterarguments:**

* **Over-restriction stifles innovation** – AI needs diverse data to improve.
* **Fair Use Doctrine** – Some argue training on public data is transformative and legal.

**Balanced Approach:**

* **Data Filtering** – Remove PII and copyrighted content where possible.
* **Opt-Out Mechanisms** – Allow content creators to exclude their work (e.g., **OpenAI’s opt-out policy** for publishers).
* **Synthetic Data** – Use artificially generated data to avoid legal risks.

**Conclusion**

GenAI must balance **innovation** with **ethical and legal responsibility**. Restrictions on training data (e.g., excluding private/copyrighted material) are necessary to prevent harm, but policies should allow flexibility for fair use and continued AI advancement.

6 Ans)

**Bias & Fairness Tools: Aequitas Bias Audit Tool**

Aequitas is an open-source **bias and fairness audit toolkit** that helps evaluate machine learning models for discrimination. One key metric it assesses is **False Negative Rate Parity**, which is critical for ensuring fairness in predictive models.

**1. What Does False Negative Rate Parity Measure?**

**False Negative Rate (FNR)** is the proportion of **actual positives** that a model incorrectly predicts as **negatives**.

**Formula:**

FNR=False Negatives (FN)False Negatives (FN)+True Positives (TP)*FNR*=False Negatives (FN)+True Positives (TP)False Negatives (FN)​

**False Negative Rate Parity** checks whether **FNR is equal across different demographic groups** (e.g., race, gender). If one group has a much higher FNR, the model is biased against them.

**2. Why Is This Metric Important?**

* **Unfair Real-World Consequences:**
  + In **criminal justice**, a high FNR for Black defendants might mean they are wrongly labeled "low risk" when they reoffend.
  + In **healthcare**, a model with high FNR for women might miss diagnosing diseases like heart attacks (which are often underdiagnosed in women).
* **Legal & Ethical Risks:**
  + Violates **anti-discrimination laws** (e.g., EU’s AI Act, U.S. Equal Credit Opportunity Act).
  + Damages public trust in AI systems.

**3. How Might a Model Fail This Metric?**

**Example: Hiring Algorithm**

* Suppose an AI screens job applicants and has:
  + **FNR = 10% for men** (misses 10% of qualified male candidates).
  + **FNR = 25% for women** (misses 25% of qualified female candidates).
* **Failure Reason:**
  + The model was trained on **historical biased hiring data** where men were favored.
  + Features like "years of continuous employment" may disadvantage women (due to career breaks).

**Result:** The model **disproportionately rejects qualified women**, reinforcing gender bias.

**4. Demo Example (Using Aequitas)**

*(Hypothetical scenario since Aequitas requires dataset upload)*

1. **Upload Data:**
   * A **recidivism prediction dataset** with columns:
     + race (Black/White), gender, predicted\_risk, actual\_recidivism.
2. **Select Metric:**
   * Check **False Negative Rate Parity** between Black and White defendants.
3. **Result:**
   * **FNR (Black defendants) = 30%**
   * **FNR (White defendants) = 15%**
   * **Bias Detected!** The model is **twice as likely** to wrongly label Black defendants as "low risk" compared to White defendants.

**Conclusion**

* **False Negative Rate Parity** helps detect **underlying biases** that harm marginalized groups.
* **Aequitas** is a practical tool to audit models before deployment.
* **Solution:** Mitigate bias via **reweighting training data**, **fairness-aware algorithms**, or **post-processing adjustments**.