AI CERTs™

AI+ Foundation™

Certification

AI + Foundation

Module 4

Hands – On 1:

Title: Getting Started with Aequitas – Fair by Design

Problem Statement:

Machine learning models often exhibit biases due to underrepresentation or skewed data, leading to unfair outcomes for certain demographic groups. For instance, medical imaging datasets may disproportionately represent lighter skin tones, causing models to perform poorly for darker skin tones. This guide addresses the need to audit and mitigate bias in AI systems using Aequitas, ensuring fairness in predictions.

Objectives:

1. Audit fairness in machine learning models using Aequitas.
2. Identify and mitigate bias at both the dataset and algorithmic levels.
3. Quantify fairness using metrics like Disparate Impact (DI) and Statistical Parity Difference (SPD).
4. Generate a de-biased dataset through synthetic data augmentation.
5. Evaluate the robustness of the model under edge cases.

Steps in Precise Manner:

1. Access the Aequitas Experimentation Environment.
2. Choose the dataset type (Image or Tabular).
3. Select or upload an image dataset.
4. Confirm dataset details and identify sensitive/target features.
5. Define proxy variables for sensitive features.
6. Select fairness metrics (e.g., DI, SPD).
7. Choose a data mitigation technique (e.g., Stable Diffusion-based Data Augmentation).

8. Configure augmentation settings (Batch Size, Epochs).
9. Run data mitigation to generate synthetic data.
10. View mitigation summary to confirm bias reduction.
11. Proceed to model mitigation or stress testing.

Tools Used:

- Aequitas Platform: Open-access tool for auditing fairness in AI models.
- Fairness Metrics: Disparate Impact (DI), Statistical Parity Difference (SPD).
- Data Augmentation: Stable Diffusion-based synthetic image generation.
- Dataset: Skin Disease Dataset (or custom dataset).

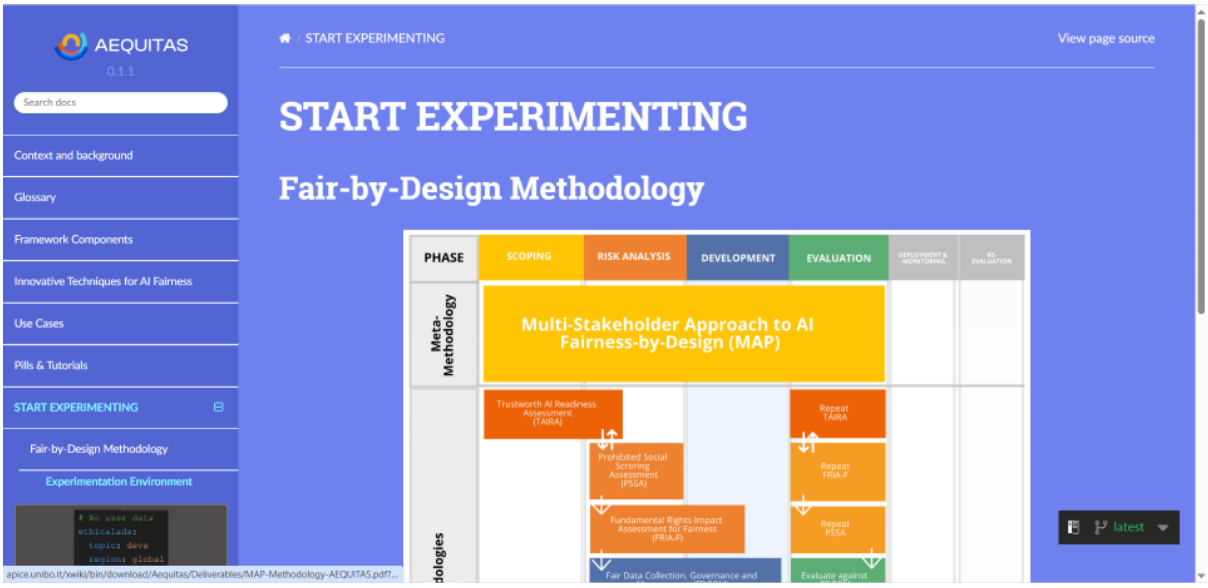
Steps in Detailed Manner:

✅ Step 1: Access the Aequitas Experimentation Environment

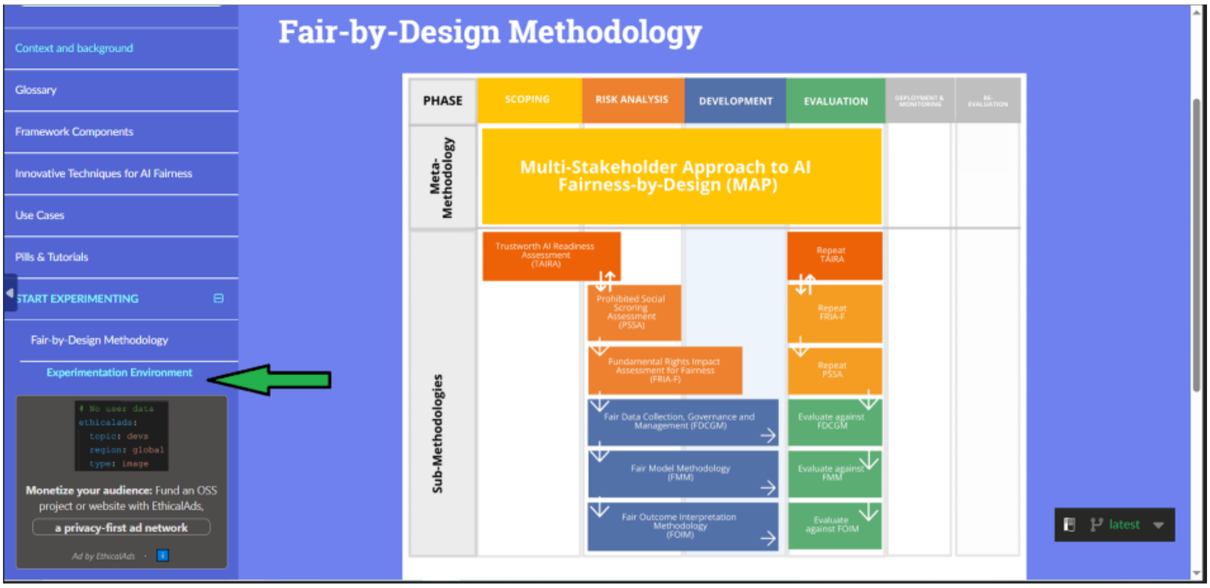
1. Open your web browser (e.g., Chrome, Firefox).
2. Navigate to the Aequitas documentation page:

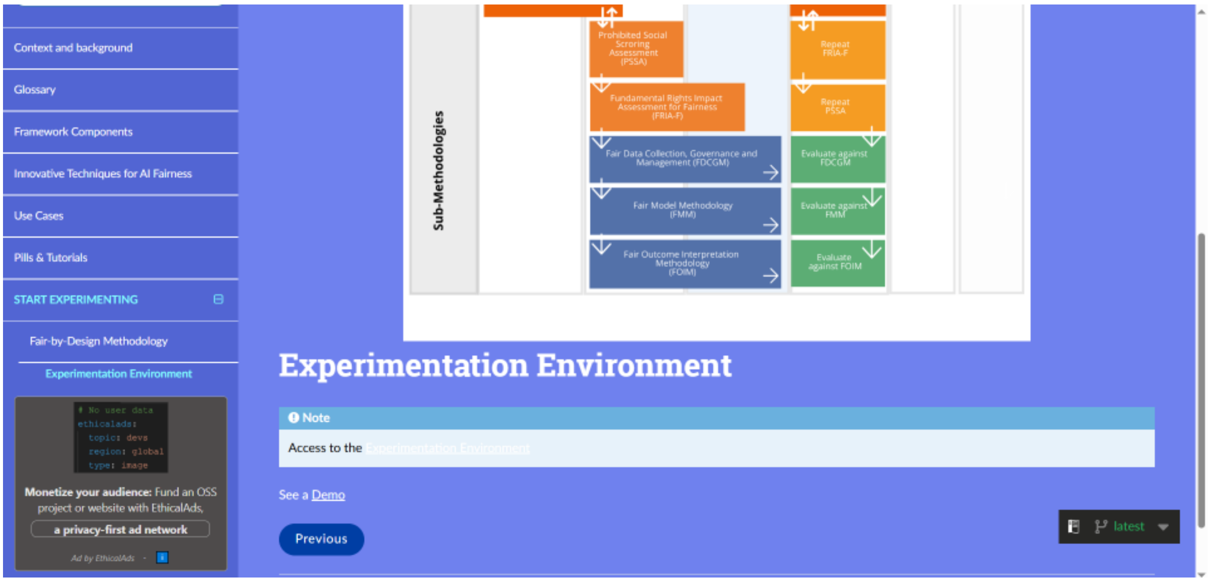
https://aequitas-home.readthedocs.io/en/latest/fair-by-design.html#experimentation-environment

3. Wait for the homepage to load.

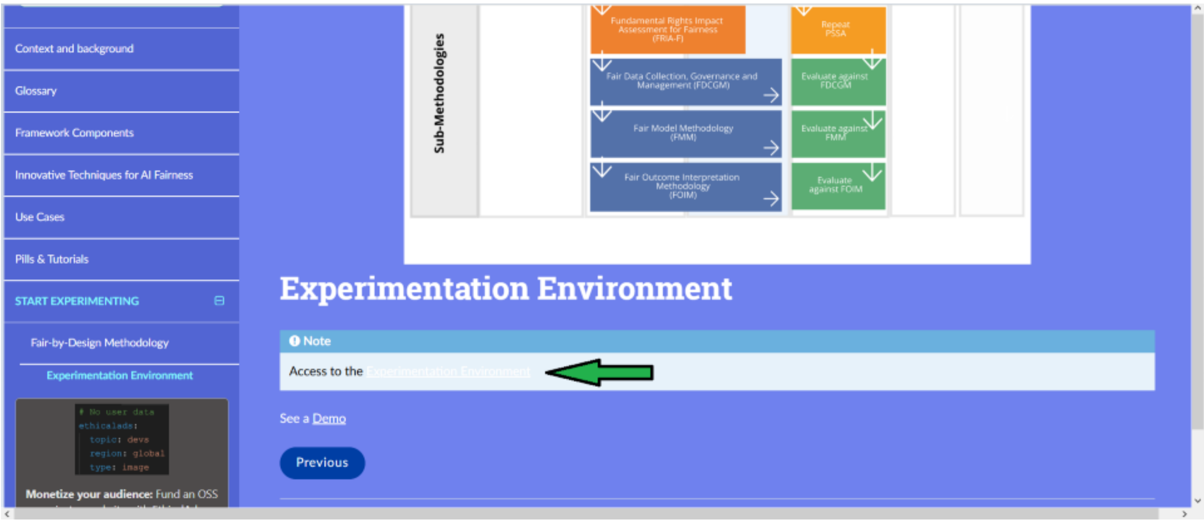


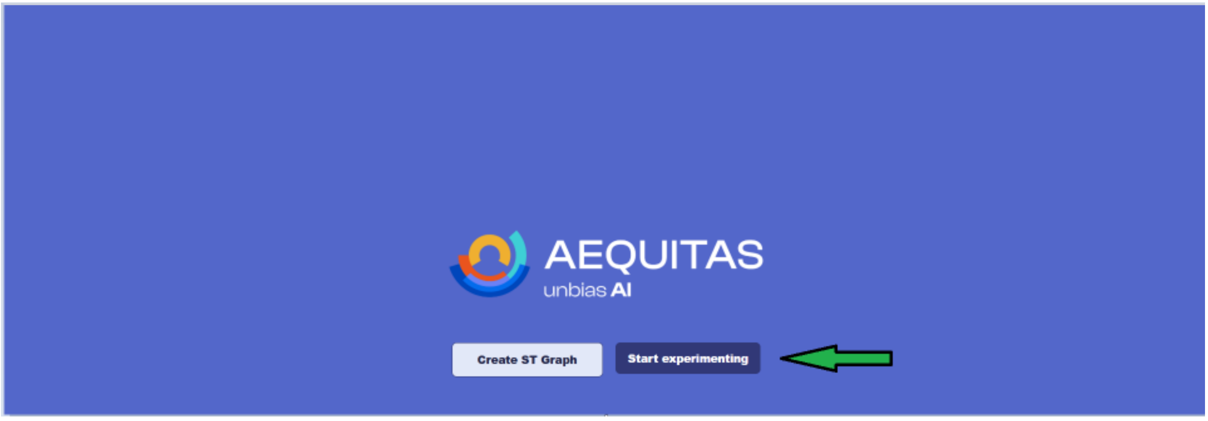
- This is the main landing page of the Aequitas platform.
- Key sections:
  - **START EXPERIMENTING:** Entry point for new fairness audits.
  - **Fair-by-Design Methodology:** Ethical AI framework.
  - **Experimentation Environment:** Area for bias analysis.





- Ads from EthicalAds indicate this is a public, open-access tool.



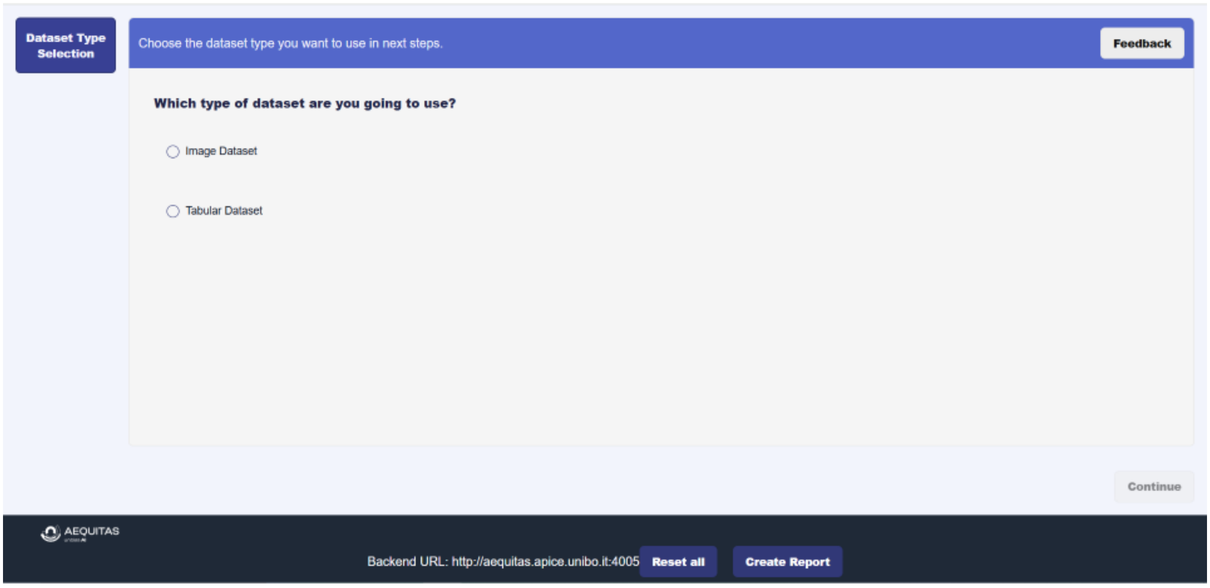


✅ Step 2: Choose Dataset Type

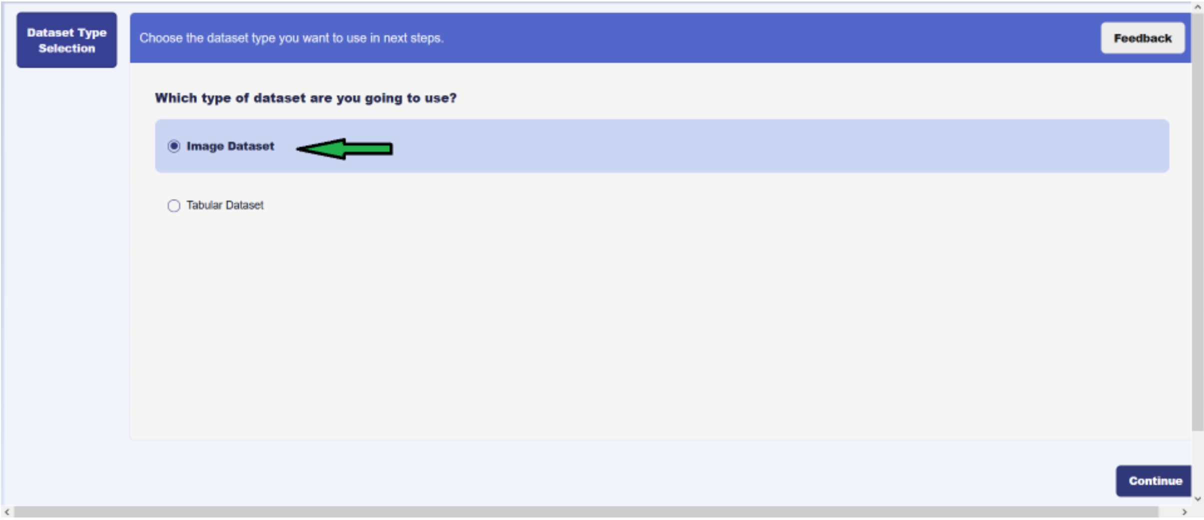
1. Click on "**Experimentation Environment**".
2. On the next screen, select:

"Which type of dataset are you going to use? "

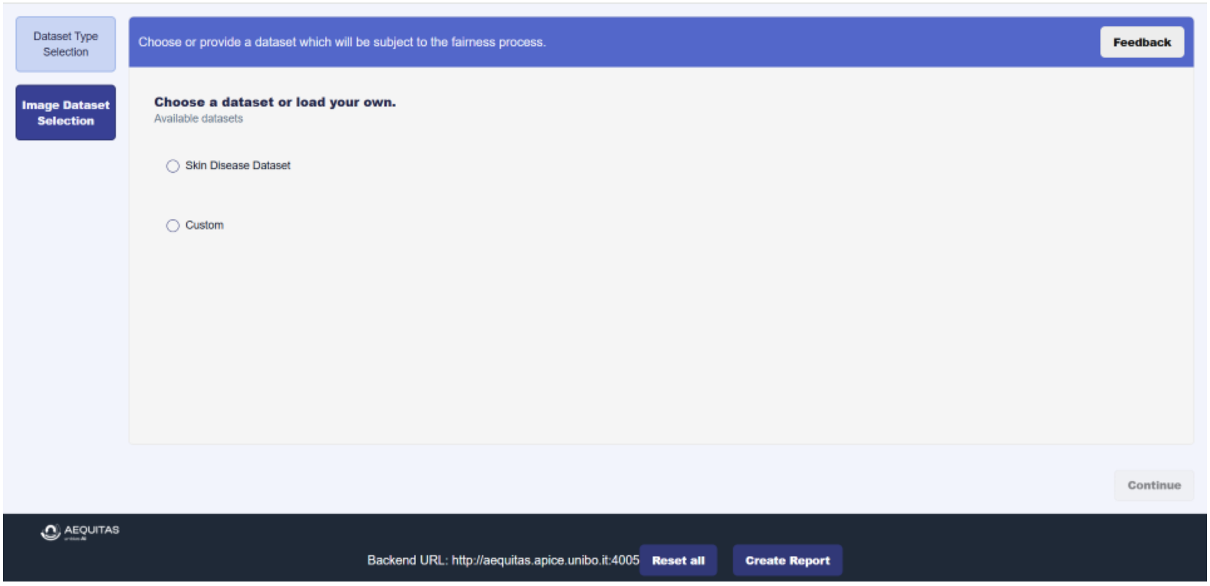
- ☐ Image Dataset
- ☐ Tabular Dataset



- This step sets the data modality for analysis.
- The system adjusts its tools depending on whether you're using images or tables.



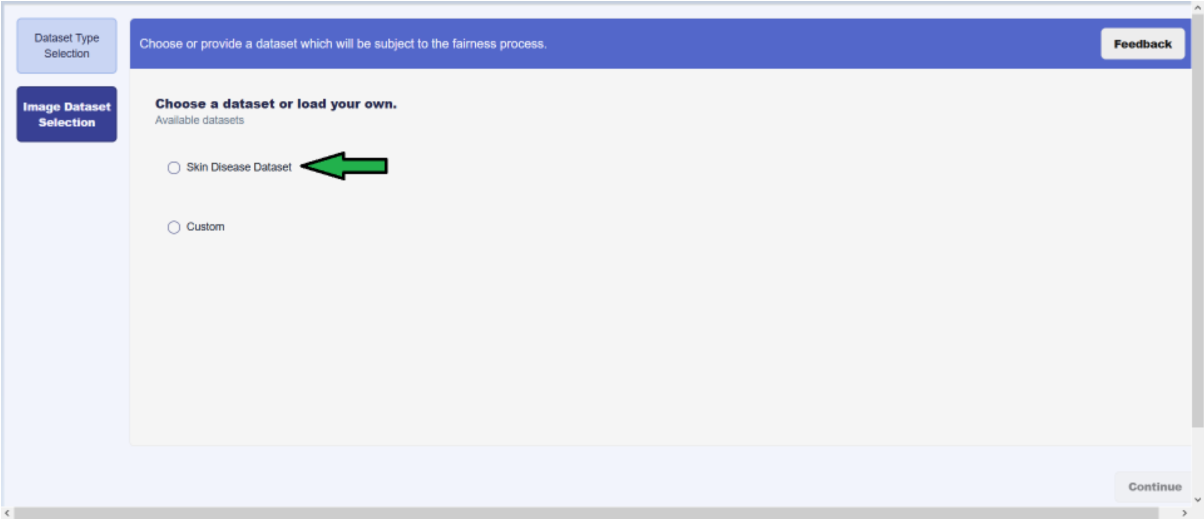
3. Select Image Dataset and click Continue.



4. Now, select the “**Skin Disease dataset**” or if you wish to upload your own dataset than select “**Custom**” option.

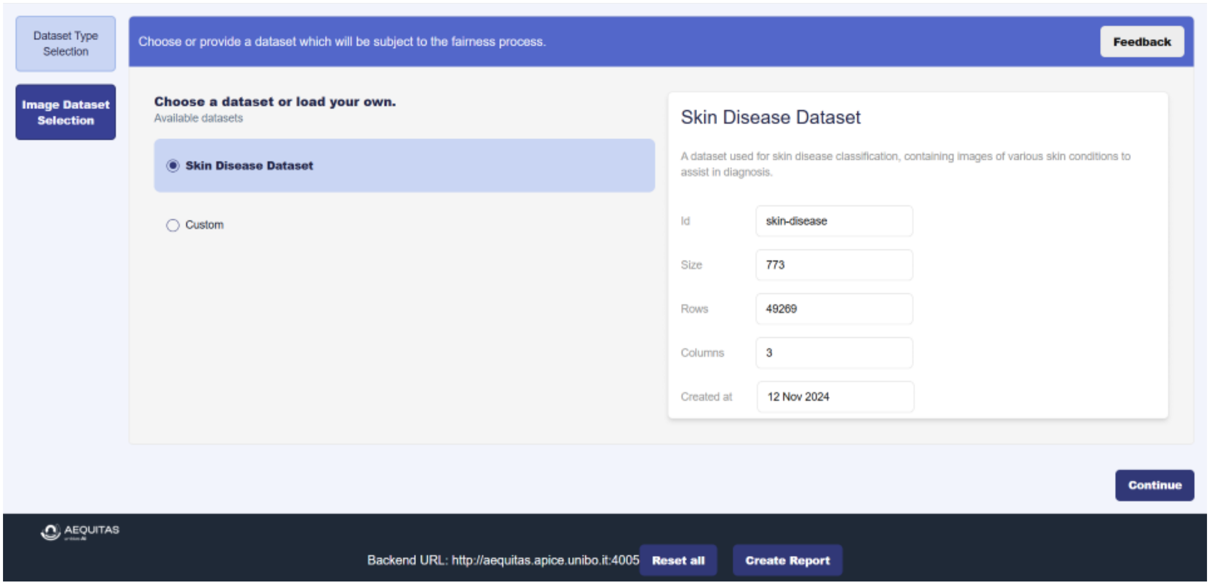
✅ Step 3: Select an Image Dataset

1. After choosing "Image Dataset", click: “**Available datasets**”

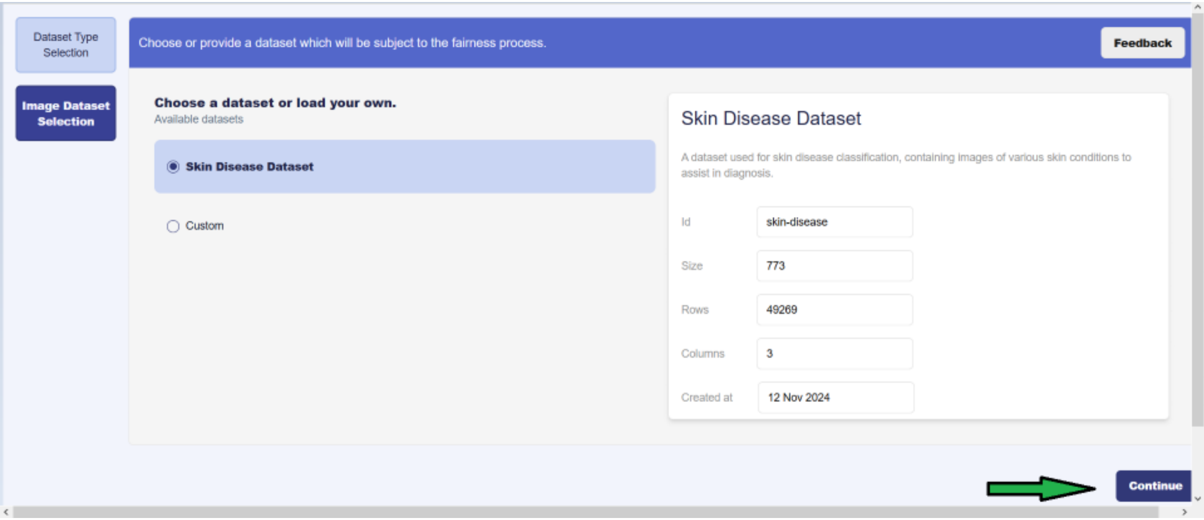


2. Choose:
   - ✅ **Skin Disease Dataset**
   - OR upload your own via **Custom**

✅ Step 4: Confirm Dataset and Select Features



- In the above fig, we can see the dataset details.



1. Review the dataset preview:

   - Rows: ~49,268
   - Columns: 3
   - Created: 12 Nov 2024

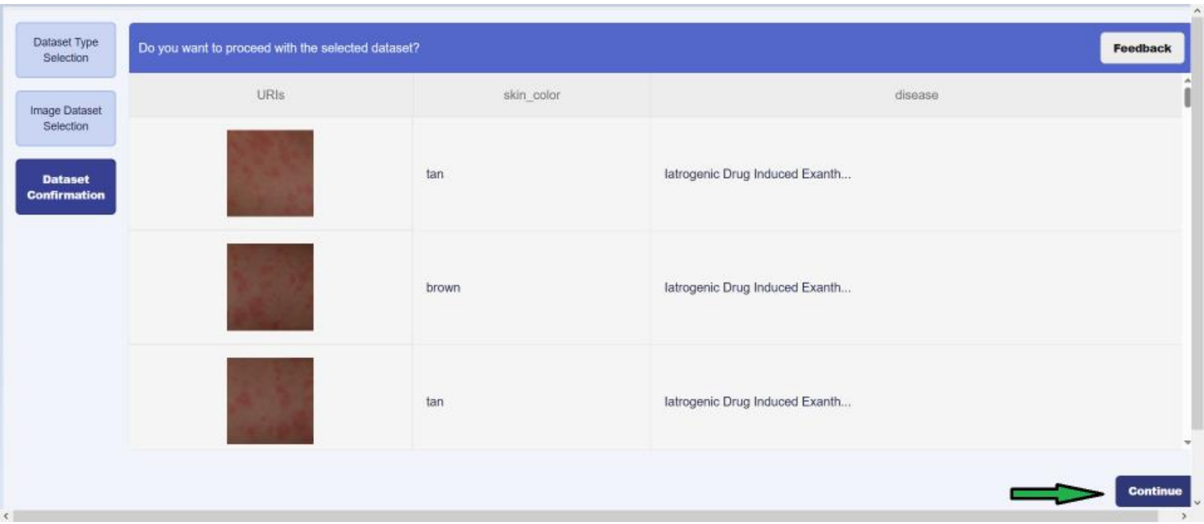
2. Click Continue to confirm successful dataset loading.

3. Dataset metadata includes:

   a. URIs: Links to image files.

   b. skin_color: Intermediate, Brown, Dark, etc.

   c. disease: Chickenpox, Urticaria, etc.



- This screen confirms the dataset has been loaded successfully.

- The Skin Disease Dataset contains dermatological images labeled by condition and patient skin tone.

✅ Step 5: Identify Sensitive and Target Features

1. Identify features:
    - ✅ skin_color: Mark as Sensitive Feature.
    - ✅ disease: Mark as Target Feature (to predict).
2. Click Continue.

- Metadata includes:
    - **URIs:** Links to image files
    - **skin_color:** Intermediate, brown, dark, etc.
    - **disease:** Chickenpox, Urticaria, etc.

- In this example, it's focused on medical image data (skin disease classification).
- This dataset may have bias risks due to underrepresentation of darker skin tones in medical imaging.

1. Select:
    - ✅ skin_color → mark as Sensitive Feature
    - ✅ disease → mark as Target Feature (to predict)

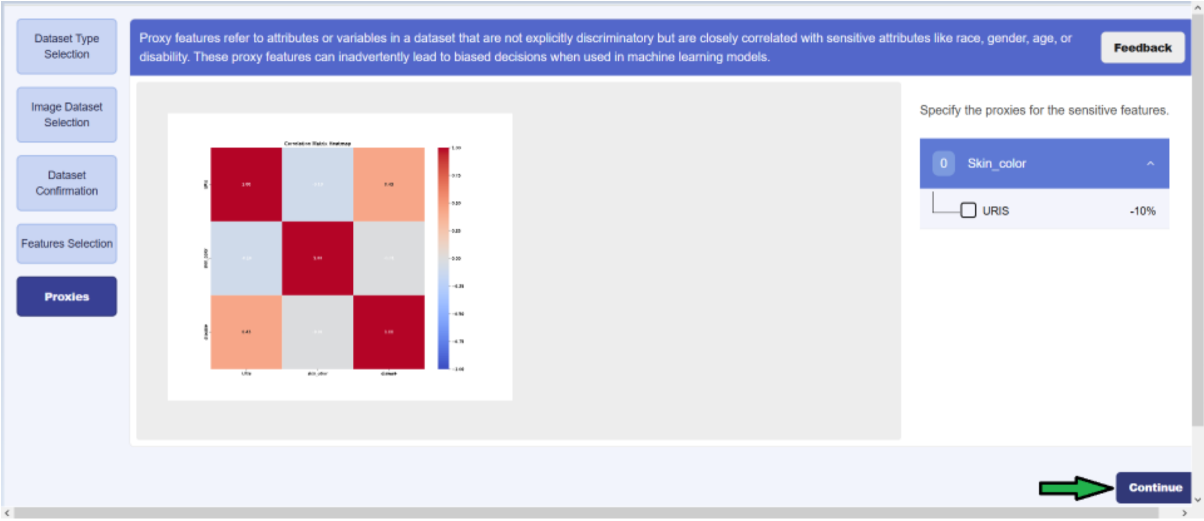
2. Click on "**Continue**" button.

- You can verify:
    - Number of samples
    - Source (via URIs)
    - Structure
- Ensures transparency before proceeding.
- Sensitive attributes are those that could lead to discrimination (e.g., race, gender, age).
- Here, skin_color is flagged because it correlates with racial demographics in healthcare.
- disease is the prediction target — the model should diagnose skin conditions fairly across skin tones.
- The system auto-detects potential sensitive features based on naming patterns.

✅ Step 6: Define Proxy Variables

1. Specify proxies for sensitive features:

    - For skin_color, ensure SKIN_COLOR is listed as a proxy.

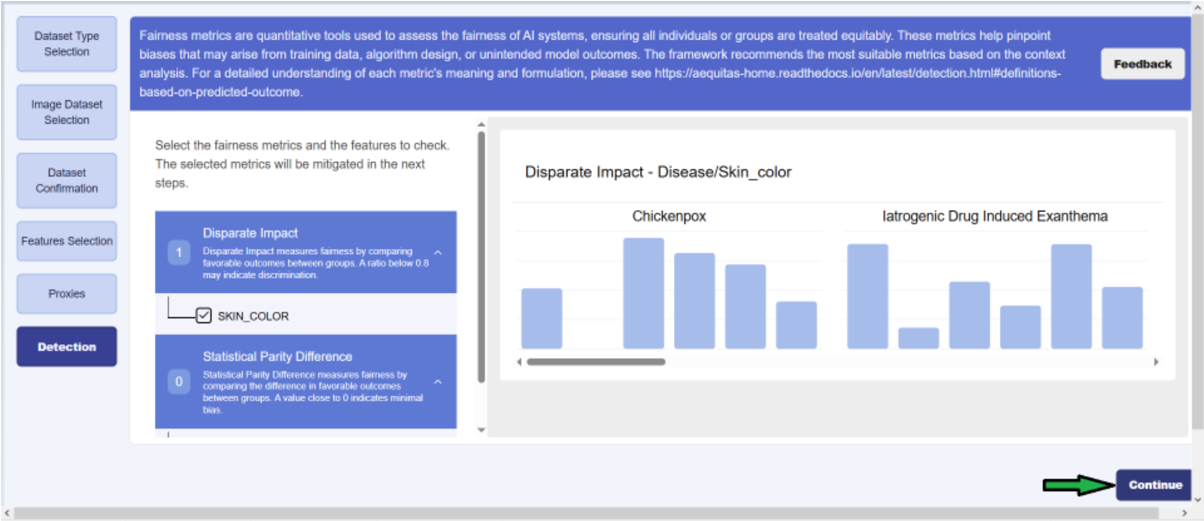


2. Click on “**Continue**” button.

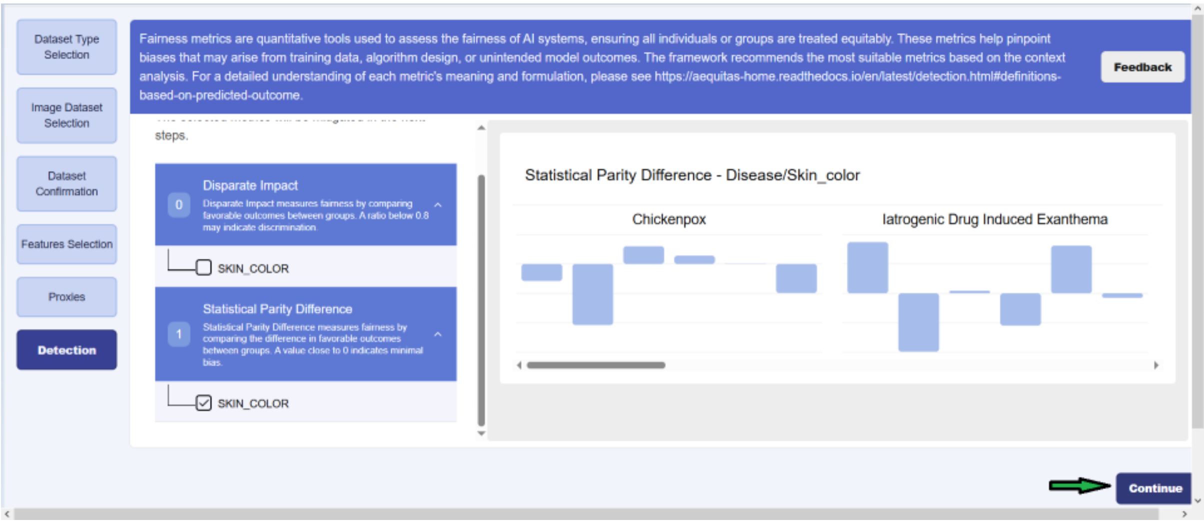
- A proxy variable is a non-sensitive feature that strongly correlates with a sensitive one.
- Example: ZIP code might act as a proxy for race.
- In images, color histograms or brightness levels might indirectly reveal skin color.
- By identifying proxies, Aequitas helps prevent covert discrimination even if direct identifiers are removed.

✅ Step 7: Select Fairness Metrics

1. Choose fairness metrics:
   ✅ **Disparate Impact (DI)**
   ✅ **Statistical Parity Difference (SPD)**



2. Apply them to:
    - Group: skin_color
    - Outcome: disease



3. Click on the "**Continue**" button.

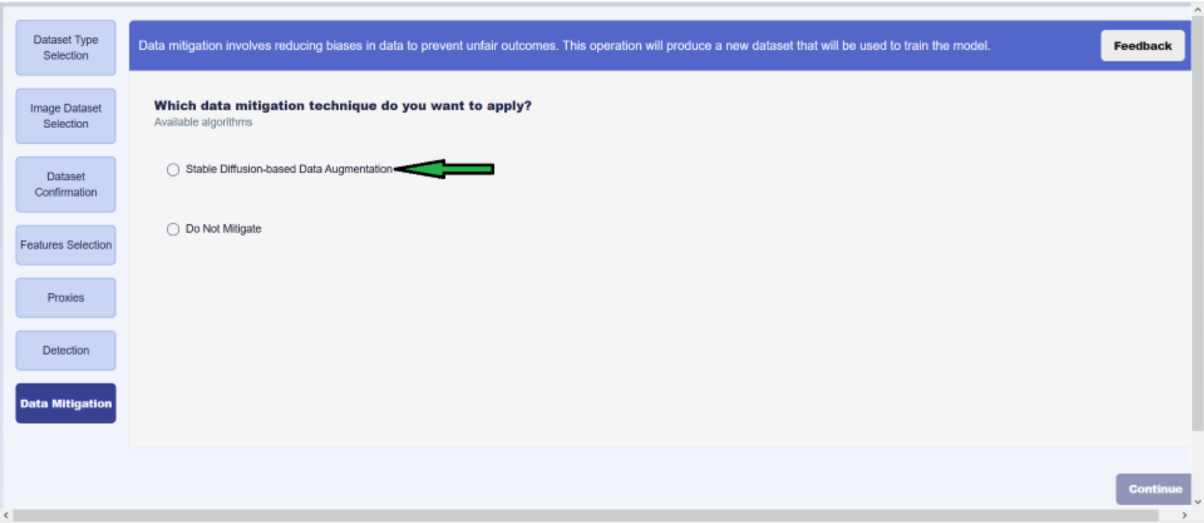
📊 What These Metrics Mean:

METRIC	DEFINITION	FAIRNESS THRESHOLD
Disparate Impact (DI)	Ratio of positive prediction rates between groups	≥ 0.8 ("80% rule")
Statistical Parity Difference (SPD)	Difference in positive prediction rates	Close to 0

- If models predict diseases less often for darker skin tones, DI < 0.8 → **unfair**.

- These metrics help quantify group-level bias.

✅ Step 8: Choose Data Mitigation Technique



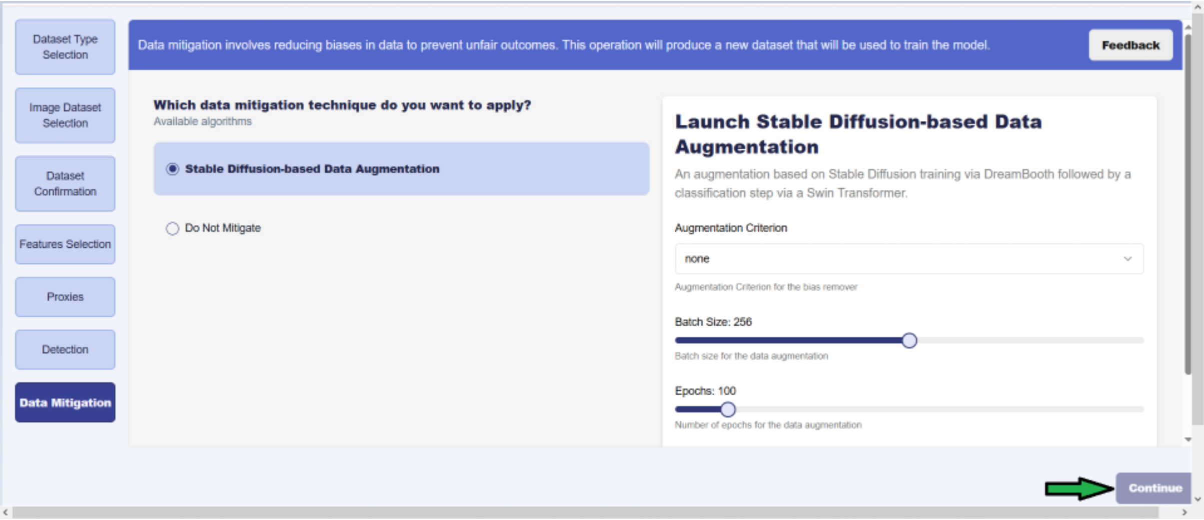
1. Under Data Mitigation, choose:
   - ✅ Stable Diffusion-based Data Augmentation
   - ❌ Do Not Mitigate

2. Click **Launch Stable Diffusion-based Data Augmentation**

✅ Step 9: Configure Augmentation Settings

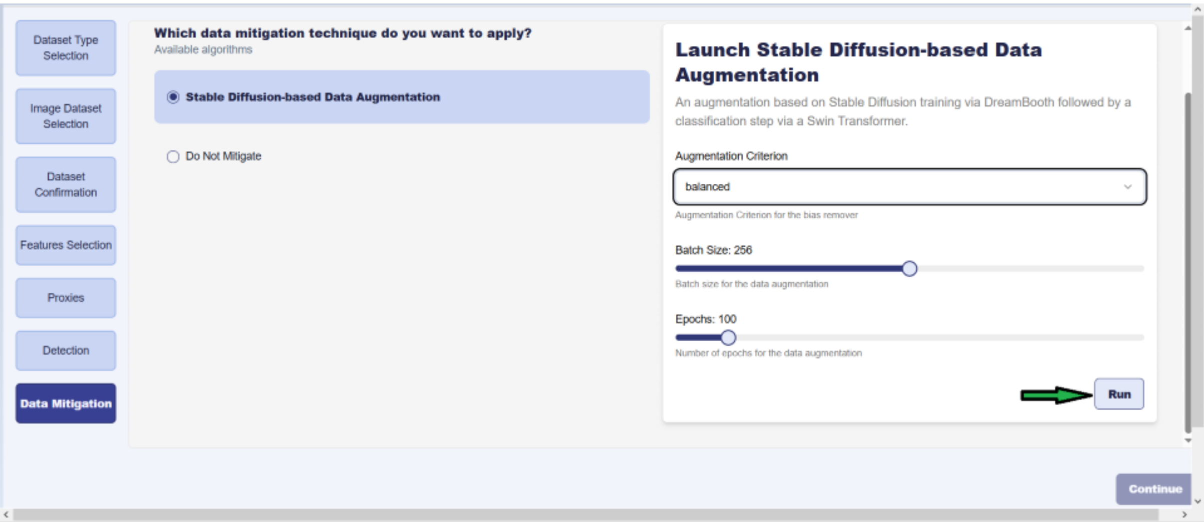
1. Set parameters:
   - **Augmentation Criterion:** Balanced
   - **Batch Size:** 256
   - **Epochs:** 100

2. Click Run Data Mitigation



- This is where bias reduction begins.
- **Problem:** Some skin tones may be underrepresented → model performs poorly on them.
- Solution: Use AI-generated synthetic images to balance the dataset.

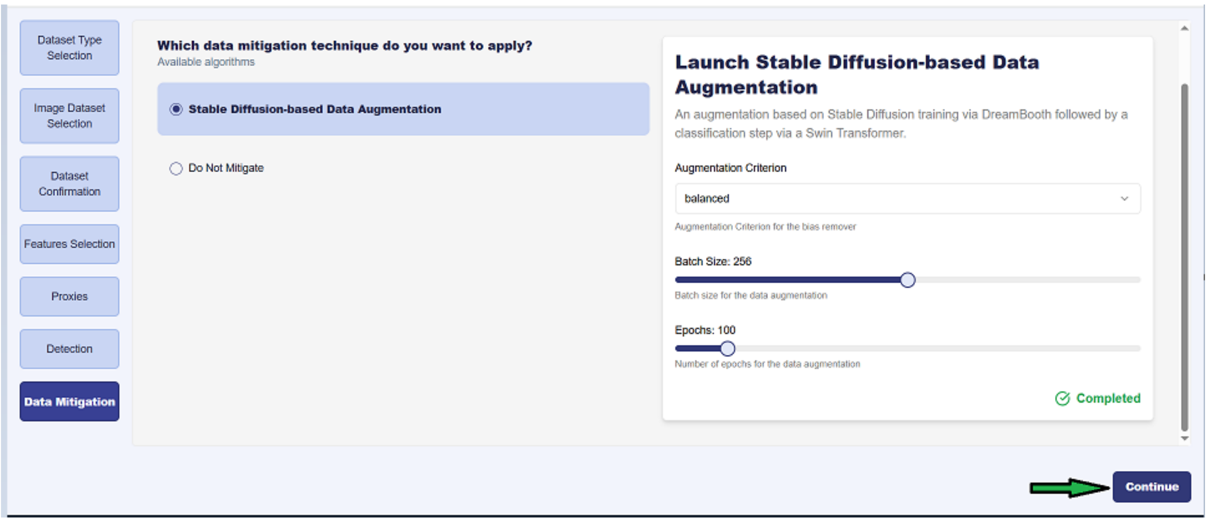
✅ Step 10: Run Data Mitigation



1. Wait for the process to complete.
2. Look for confirmation: "✅ Data Mitigation – Completed".
3. System actions:
    a. Generated synthetic images for underrepresented skin tones.

b. Balanced class distribution.

c. Created a de-biased version of the dataset.

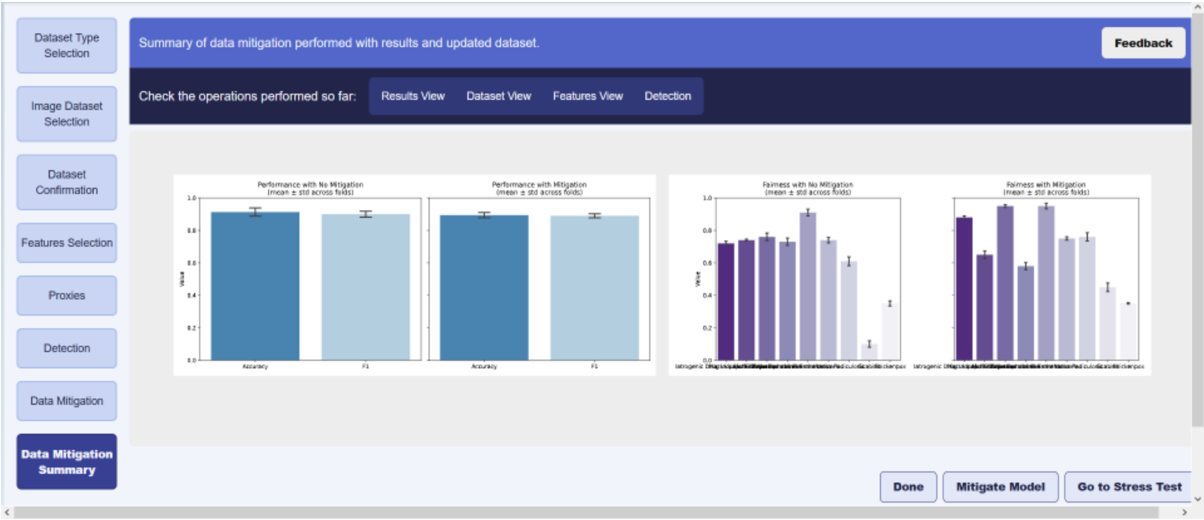


- The system has now:
    - Generated synthetic images for underrepresented skin tones.
    - Balanced class distribution.
    - Created a **de-biased version** of the dataset.
- This new dataset can now be used to train a fairer AI model.

✅ Step 11: View Mitigation Summary

1. Click View Results or Continue.
2. Explore:
    - Updated dataset stats.
    - Feature distributions.
    - Detection results.
3. Confirms:

- Increased sample count for minority groups.
- More uniform feature distribution.
- Bias reduction at the data level.

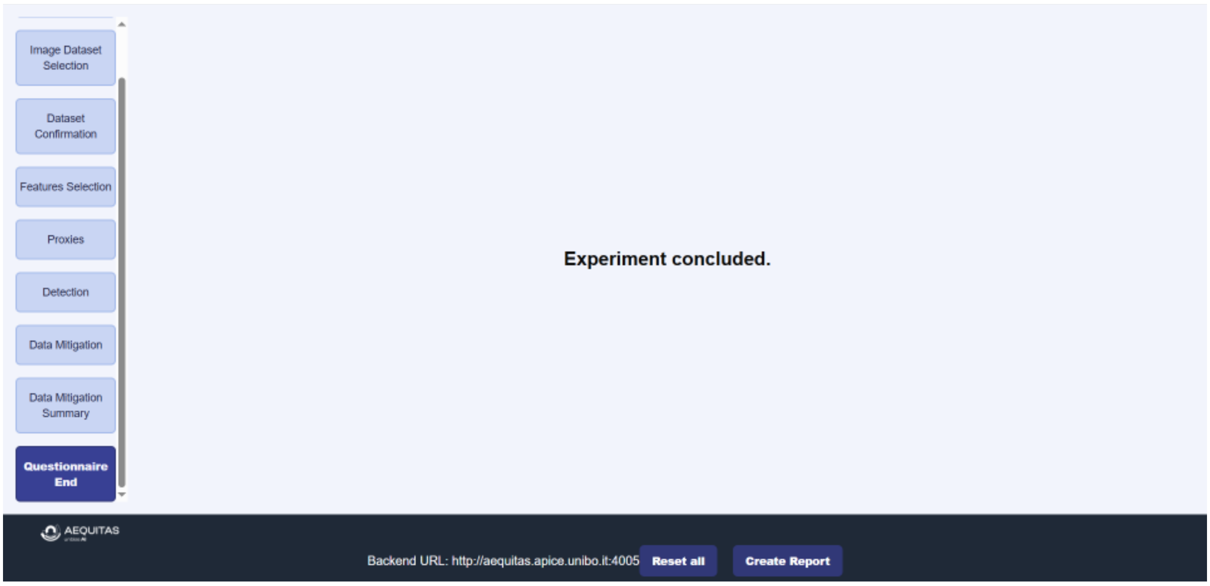


- Shows impact of mitigation:
  - Increased sample count for minority groups
  - More uniform feature distribution
- Confirms that **bias has been reduced at the data level.**

✅ Step 12: Proceed to Model Mitigation or Stress Test

1. Click:

- **Mitigate Model** → Apply algorithmic fairness techniques
- **Go to Stress Test** → Evaluate robustness under edge cases



- This concludes the data-centric fairness phase.

AI CERTs™

AI & BITCOIN CERTIFICATIONS!

aicerts.ai

Contact

252 West 37th St., Suite 1200W

New York, NY 10018