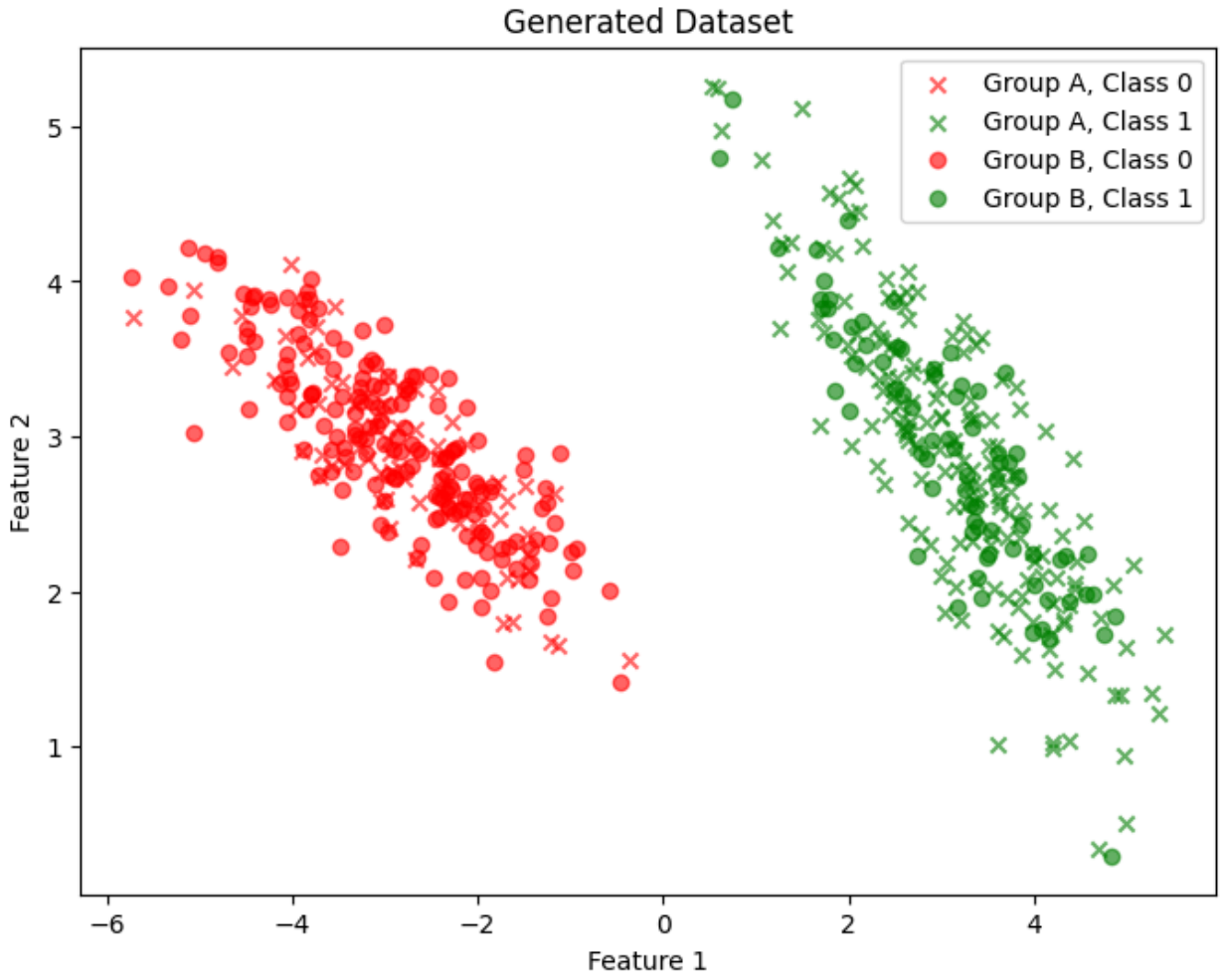
Fair ML - Homework #1

Click here for CODE FILE

1. Total Samples: 500
Selection Class (y=1): 250
Rejection Class (y=0): 250
Group A (Selection Class): 175
Group A (Rejection Class): 70
Group B (Selection Class): 75
Group B (Rejection Class): 180

2.



3. Training Accuracy: 100.00%

```
def train_logistic_regression(X, y):
    model = LogisticRegression()
    model.fit(X, y)

    y_pred = model.predict(X)
    accuracy = accuracy_score(y, y_pred)
    print(f"Training Accuracy: {accuracy * 100:.2f}%")
    return model

model = train_logistic_regression(X, y)
```

```
Training Accuracy: 100.00%
```

4. **(a.)** Demographic Parity Calculation

P(ŷ = 1 | Group A): 0.714
P(ŷ = 1 | Group B): 0.294
Demographic Parity Difference: 0.420
Demographic Parity Difference: 0.42016806722689076

```
Demographic Parity Calculation
P(ŷ = 1 | Group A): 0.714
P(ŷ = 1 | Group B): 0.294
Demographic Parity Difference: 0.420
Demographic Parity Difference: 0.42016806722689076
```

(b.)

Demographic Parity measures the difference in positive prediction between groups A and B.

$P(\hat{Y} = 1 \mid G = A)$ = No. of positive predictions of group A / total samples in group A
$P(\hat{Y} = 1 \mid G = B)$ = No. of positive predictions of group B / total samples in group B
Here $\hat{Y} = 1$ is model predicting positive from selection class
G = A is sample belonging to group A
G = B is sample belonging to group B

$P_A = P(\hat{Y} = 1 \mid G = A) = 175 / (175 + 70) = 0.714$
$P_B = P(\hat{Y} = 1 \mid G = B) = 75 / (75 + 180) = 0.294$

Difference of Demographic Parity = 0.714 - 0.294 = 0.420

5. **(a.)** Balance: 0.5767012687427913

```
Balance: 0.5767012687427913
```

(b.) We know that Balance function is

$$\text{Balance} = \min_{k\epsilon[k],\ g\epsilon[m]} \min\{\frac{r_X^g}{r_k^g}, \frac{r_k^g}{r_X^g}\}$$

Total samples = 500
Selection Class (y=1): 250
Rejection Class (y=0): 250
Group A (Selection Class): 175
Group A (Rejection Class): 70
Group B (Selection Class): 75
Group B (Rejection Class): 180

Here, r_X^A = 245/500 = 0.49 = group A

r_X^B = 255/500 = 0.51 = group B

For group A,
Selection = 175/245 = 0.714
Rejection = 70/245 = 0.285
For group B,
Selection = 75/255 = 0.294
Rejection = 180/255 = 0.705

Group A,

Class selected = $min\ \{\frac{r_X^A}{r_1^A}, \frac{r_1^A}{r_X^A}\}$ = min $\{\frac{0.49}{0.714}, \frac{0.714}{0.49}\}$ = min{0.686, 1.457} = 0.686

Class rejected = $min\ \{\frac{r_X^A}{r_0^A}, \frac{r_0^A}{r_X^A}\}$ = min $\{\frac{0.49}{0.285}, \frac{0.285}{0.49}\}$ = min{1.719, 0.581} = 0.581

Group B,

Class selected = $min\ \{\frac{r_X^B}{r_1^B}, \frac{r_1^B}{r_X^B}\}$ = min $\{\frac{0.51}{0.294}, \frac{0.294}{0.51}\}$ = min {1.734, 0.576} = 0.576

Class rejected = $min\ \{\frac{r_X^B}{r_0^B}, \frac{r_0^B}{r_X^B}\}$ = min $\{\frac{0.51}{0.705}, \frac{0.705}{0.51}\}$ = min{0.723, 1.382} = 0.723

Balance = min {0.686, 0.581, 0.576, 0.723} = 0.576

6. **(a.)** Removed 50 samples to balance groups.
Training Accuracy: 100.00%

```
 Updating Dataset to Improve Fairness
 Allowed budget for changes: 50 samples.
Removed 50 samples to balance groups.
Training Accuracy: 100.00%
```

(b.) The previous values before updating are,
Demographic Parity Difference: 0.42016806722689076
Balance: 0.5767012687427913

The new values are,
New Demographic Parity Difference: 0.34690799396681754
New Balance: 0.5190311418685122

```
Demographic Parity Calculation
P(ŷ = 1 | Group A): 0.641
P(ŷ = 1 | Group B): 0.294
Demographic Parity Difference: 0.347
New Demographic Parity Difference: 0.34690799396681754
New Balance: 0.5190311418685122
```

(c.)
The approach worked because it made small adjustments to the dataset by removing up to 10% of the samples. The goal was to reduce the gap in positive prediction rates between different groups while ensuring the dataset remained linearly separable. By carefully selecting which samples to remove, the model was able to do better fairness without sacrificing accuracy.

Before updating the dataset, Group A was more likely to receive positive predictions compared to Group B, which led to an imbalance in selection rates. By removing some overrepresented samples from Group A's selection class and Group B's rejection class, the dataset became more balanced, which helped reduce the disparity in how often each group was selected for positive predictions.

Demographic Parity Difference decreased from 0.4202 to 0.3469, meaning the gap in positive prediction rates between the two groups got smaller.
Balance improved from 0.5767 to 0.5190, showing that the distribution of positive predictions across groups became more even.
Most importantly, training accuracy remained at 100%, proving that the dataset is still linearly separable, and the model has not lost any predictive power.

This approach worked because it made fairness improvements without altering the fundamental structure of the dataset. Instead of changing any labels or group assignments, it adjusted the representation of different classes in a way that naturally led to a fairer model. By reducing bias in the training data, the model now makes more equitable decisions across groups, while still maintaining perfect accuracy.