Assignment 4: Binary Classification via Logistic Regression Optimization

Link for dataset: Assignment4 dataset
Name:____________________

Objective: To implement Logistic Regression from scratch using numerical optimization and to analyze the effect of data noise and dimensionality on the model's decision boundary.

Dataset Overview

You are provided with four datasets:

1. **dataset0.csv**: 1D features (x) with perfect separation.
2. **dataset1.csv**: 1D features (x) with overlapping labels (noise).
3. **dataset2.csv**: 2D features (x, y) with well-separated classes.
4. **dataset3.csv**: 2D features (x, y) with significant overlap.

Part 1: 1D Logistic Regression and the "Infinite Weights" Phenomenon

In this section, you will compare a perfectly separated dataset with an overlapping one.

Tasks:

1. **Visualization:** Plot the labels against x for both dataset0 and dataset1.
2. **Implementation:** Define the Sigmoid function and the Binary Cross-Entropy (Log-Loss) function.
3. **Optimization:** Use *scipy.optimize.minimize* to find the parameters coefficients for both datasets.
4. **Analysis:**
    - Report the final weights for both datasets.
    - Calculate the decision boundary for both and report them.
5. **Sigmoid Plotting:** Overlay the resulting Sigmoid curve P(y=1|x) over the scatter plot of the data points for both cases. (ask gemini to do this)

Part 2: 2D Logistic Regression and Linear Boundaries

Now, extend your model to handle two features (x and y).

Tasks:

1. Optimization: Perform Logistic Regression on dataset2 and dataset3. Your model equation will be:
   z = beta_0 + beta_1 x + beta_2 y
2. **Boundary Derivation:** The decision boundary occurs at z=0. Rearrange this equation into the form y = mx + c.
3. **Comparison:**
   - For dataset2 (Well-Separated), plot the points and the resulting decision boundary line.
   - For dataset3 (Overlapping), plot the points and the resulting decision boundary line.
