Task 2:

- Goals:
  - Heart disease: Predict target
  - Country dataset: Select the countries to give aid to.
- The target distribution for the heart disease dataset shows low class imbalance, as both have similar frequencies.
- The country dataset is used for unsupervised learning, hence the analysis focuses on the features and their relationships.
- In the heart disease data set we can see strong relations like:
  - Target with cp, thalach, ca, and slope.
  - Thalach with age, exang, and oldpeak.
- In country dataset we can strong relations like:
  - Gdpp with income, life_expec, and total_fer.
  - child_mort with Total_fer, life_expec and income
- in heart disease we can see outliers like :
  - chol (100-600)
  - trestbps(100-200)
- in country dataset we have:
  - income(600-120000)
  - gdpp(200-100000)
- These outliers are to be expected but may distort distributions as large range features may dominate.
- Hence, The use of standard scaler prevents such large features from dominating the gradient updates.
- Since, both datasets were clean and encoded this step was skipped.
- Scaling does not effect the importance of a feature, it remove the dominance due to scale.
- Features were created like :

    - High_bp and high_chol to indicate that disease risk do not increase linearly.
    - Instead of judging countries based on 9 different labels we use the combined feature to cluster for aid prioritization.
- We do not create any features which provide information about the future or the target variables all the features were created from provided labels.
- Assumptions of the logistic regression model applied are:
    - Features combine linearly w.x+b
    - The features are independent of one another(No multicolinearity)
- The model is sensitive to parameters like learning rate. It will cause divergence if too high or slow training if too low.
- The use of fixed probability thresholds may cause misclassification
- By dimensionality reduction with pca we can now see that plot and judge that x axis(most variance) is like a composite of finiance and health score by looking comparing them with the heat maps we can see that cluster 3 has higher values while 2 has lower values. While the y axis is likely the trade score.
- We can validate the score of different clusters with original 9 features, eg: Cluster 3 has high health and finiance scores and it also has high gdpp in the original features with high life expec and low child_mort, cluster 2 has low score all across and we can validate it with the scores of the original 9 features.
- The clustering helps create a target lable, which can be predicted with supervised learning. The engineered features can also help reduce the features with high correlation like inflation and gdpp
- Two failure case of the supervised model:
    - The model may fail to capture non linear relationships
    - Due to high correlation between features the Coefficients might be unstable
- Why did you choose to train the model on three engineered features instead of the original nine?
    - Makes the model more interpretable
    - Reduces redundancy due to multicollinearity
    - Reducing Bias dues to more columns in a particular category

- How did you decide the number of clusters without a target variable?
    - Used internal validation metrics like Silhouette Score and inertia to measure cluster compactness with elbow method
- How does your implementation ensure convergence of gradient?
    - With scaled features and optimal leaning rate and iterations the gradient is calculated in a continuous manner