Team Members: Richard Willet, Tennille Franklin, Denise Miller, Titus Muchiri

Project Proposal: Can a machine learning model accurately predict the likelihood of heart disease based on the chosen dataset?

* What are the most important health indicators for predicting heart disease?
* Can clustering techniques identify distinct groups at higher risk for heart disease?
* If we exclude individuals with known risk factors (e.g., high blood pressure, diabetes), can we still predict heart disease with a reasonable accuracy?
* Among people with obesity, is the heart disease rate higher than those with a normal BMI?
* Is there a difference in heart disease prevalence between men and women in this dataset?

Data set we plan to use:

<https://www.kaggle.com/datasets/alexteboul/heart-disease-health-indicators-dataset/data>

Expected Outcome:

To build a machine learning model that predicts individuals at high risk for heart disease based on the given health factors, we’ll follow these steps:

1. **Data Preparation**:
   * Load and clean the dataset, handling missing values if any.
   * Split the dataset into features (health factors) and the target variable (heart disease risk).
2. **Feature Engineering**:
   * Standardize or normalize numerical features as needed.
   * Encode categorical features if they are present.
3. **Model Selection**:
   * Choose a suitable machine learning model for binary classification, such as **Logistic Regression**, **Random Forest**, or **XGBoost**.
4. **Training and Evaluation**:
   * Train the model on the training set.
   * Evaluate the model’s performance using metrics like **accuracy**, **precision**, **recall**, and **F1 score**.
5. **Hyperparameter Tuning** (optional):
   * Use techniques like grid search or random search to optimize model performance.
6. **Interpretation**:
   * Analyze feature importance to understand which health factors most influence heart disease risk.

Visualizations:

Linear regression model

Box Plot showing each health factor and related heart disease risk

Scatter plot for age groups and heart disease