# Predicting Health Outcomes Using Machine Learning
# Summer ML Course

## Overview

This project involves applying machine learning techniques to predict health outcomes using a relevant health dataset of your choice. You will preprocess the data, perform exploratory data analysis (EDA), build and evaluate machine learning models, and interpret the results while considering ethical implications.

## Objectives

1. Preprocess and explore the dataset.
2. Implement and evaluate supervised machine learning models.
3. Interpret the results and identify significant predictors.
4. Discuss ethical considerations in machine learning for public health.

## Project Steps

1. **Select a Dataset**

   Choose a publicly available health dataset containing predictor variables (features) and an outcome variable (target). Examples include datasets related to chronic diseases, patient readmissions, or other health conditions.

2. **Data Preprocessing**

   **Goal:** Prepare the data for analysis and modeling.

   - **Load the dataset:** Creating a tidy dataset in `R`.
   - **Explore the dataset:** Print the first few rows and summary statistics to understand the data structure.
   - **Handle missing values:** Identify and impute missing values using appropriate techniques (e.g., imputation).
   - **Scale the data:** Normalize the feature variables to ensure consistent scale across features.

3. **Exploratory Data Analysis (EDA)**

   **Goal:** Understand the data and identify patterns.

   - **Visualize the data:** Create histograms, box plots, and scatter plots to examine distributions and relationships between variables.

- **Correlation analysis:** Compute and visualize the correlation matrix using a heatmap to identify relationships between features and the target variable.

4. **Build and Evaluate Machine Learning Models**

   **Goal:** Develop and assess models to predict health outcomes.

   - **Split the data:** Divide the dataset into training (e.g., 80%) and testing (e.g., 20%) sets, or use cross-validation.
   - **Implement models:** Build at least two supervised learning models (e.g., logistic regression, decision trees, random forests, support vector machines).
   - **Evaluate models:** Use metrics such as prediction error, sensitivity, ROC-AUC, etc.

5. **Interpret the Results**

   **Goal:** Understand and communicate the significance of your findings.

   - **Feature importance:** Identify which features are the most significant predictors of the health outcome.
   - **Model comparison:** Compare the performance of different models and discuss why one might perform better than another.
   - **Discussion:** Interpret the implications of your findings for public health.

# Deliverables

1. Project Report:

   - Document your analysis, including data preprocessing steps, EDA, model building and evaluation, interpretation of results, and ethical considerations.
   - Include visualizations, tables, and code snippets where appropriate.

2. Presentation:

   - Summarize your findings and conclusions in a short presentation. Be prepared to discuss your approach and results.

# Evaluation Criteria

1. Data Preprocessing and EDA (30%)
2. Model Implementation and Evaluation (40%)
3. Interpretation of Results (20%)
4. Report and Presentation Quality (10%)

By completing this project, you will gain practical experience in applying biostatistical machine-learning techniques to a real-world public health problem, enhancing both your technical skills and your understanding of the ethical dimensions of data analysis.