Title: Heart Disease Prediciton and Risk Factors

Team members: Adam Fowler, Tyler Sanchez, Jacob Unger

Description: Using data from the Behavioral Risk Factor Surveillance System (BRFSS), determine which risk factors are the strongest indicators of heart disease. Look for correlations between risk factors. Attempt to predict heart disease in a patient given the presence of indicators.

Prior Work: This is a popular dataset on Kaggle with a variety of prior work. Our goal will be an independent analysis with results being verified by prior work.

Sample of prior work:

Title	URL
Heart Disease Fastai with TabularPandas	https://www.kaggle.com/code/stpeteishii/heart-disease-fastai-with-tabularpandas
Heart Disease - Binary Classification	https://www.kaggle.com/code/murattademir/heart-disease-binary-classification
Heart Disease Indication (PCA, Classifier & CNN)	https://www.kaggle.com/code/sohaelshafey/heart-disease-indication-pca-classifier-cnn
Nueral Network Regression VS Sklearn Algorithms	https://www.kaggle.com/code/abohelal/nueral-network-regression-vs-sklearn-algorithms

Dataset

- Heart Disease Health Indicators Dataset
- https://www.kaggle.com/datasets/alexteboul/heart-disease-health-indicators-dataset
- The dataset and our additions will be stored at https://github.com/cockytrumpet/data-mining-group-4

Proposed work

- Data cleaning
    - Look for discrepancies in the data, this will insure the accuracy of the data and ensure variables all line up and are in the same format across all entries.
    - For missing values, we will use a universal constant where ever that may be necessary. Either a numerical value or a ‘Unknown’ label.
    - In order to clean up noise, use outlier analysis to prune and smooth the data. Regression will be used to see if there are related variables.
- Data preprocessing
    - Investigate the believability and interpretability of the data as well as its accuracy and consistency. This will insure the data can be trusted and understood.
    - Perform data reduction and data transformation in order to aid in analysis.

- Data integration
  - Perform correlation analysis to see if there is anything that we can use to predict heart disease based on the attributes that are given.
  - Ensure attention is paid to conflict detection and make sure to resolve issues that arise.
- Tools to be used
  - python/R
  - pandas - dataset manipulation
  - matplotlib/seaborn - visualization
  - scikit - analysis/classification

Evaluation

- Sample to create a testing dataset.
- Test the efficacy of our model.