DS210 Project
My project is to take a CDC-backed dataset of people who have heart attacks and traits relating to it and measure the similarity from testing and training data using a k nearest neighbor algorithm. This could potentially be used in heart attack predictive programs where similarities are compared to the data.

Since we are dealing with a lot of nodes we will need to use cargo run –release to speed up the process usually it will take about a minute or two. Once ran it should look something like:
Loaded 246022 patients.
Selected quarter - Training data size: 3936, Testing data size: 984
Model accuracy for selected quarter: 93.80%

Note: there is a lot of unnecessary struct indicated through VSCode, however, if removed or fixed the

Graph.rs
Using the packages csv, serde, petgraph, and rand I first struct large enough to intake all the variables present in the csv file.

My first function fn load_dataset basically extracts the data from the csv file and processes it and puts all the data in the patient struct. Finally, upon success it returns patient if not it returns an error.

My second function fn randomize_and_split due to the large size of my data set (240,000 nodes) individual it has caused the necessity to make my data more manageable. This has led me to divide the size of the data by a magnitude of 50 times. I would then randomize the data. Then section 1/50 in the first quarter, 1/50 in the second quarter, 1/50 in the third quarter, and the rest in the fourth quarter. Than give return

My third function fn build_graph makes a graph where nodes equals a patient and adds an edge if the similarity greater or equal than a threshold.

My fourth function fn calculate_similarity compares 2 nodes scores p1 and p2 and from that devises a similarity score.

My fifth function fn predict_risk adds patients to the graph. Based on similarities of the score finds (nearest neighbor) based off similarity threshold.

My first test case test if my build_graph function works using made up data

My second test case test if my predict risk function works using made up data

Main.rs
My main.rs just sets the parameters that is required for graph.rs. It establishes what the csv file is called. It also sets the test ratio and training which is 20% to 80%. It then establishes threshold score, constructs graph, and iterates through the test data, classifying risk and checks it accuracy. Finally, than calculating the accuracy.