* **Dataset:**[Cardio Vascular Data](https://www.kaggle.com/datasets/thedevastator/exploring-risk-factors-for-cardiovascular-diseas)

**1. Write your group number and the names of all members of your group.**

* Group 19.
* Jason Mullen
* Chuong Tran
* Vinh Nguyen
* Trang Hoang
* Nathan Lilly

**2. Provide a detailed description of the data your group has selected, including**

* **a. The source and inspiration for selecting this particular data set.** 
  + When it comes to the inspiration for this project, we all are relatively interested in cardiovascular health and as a result we decided that this particular data set would be the best.
* **b. The size of the data (number of observations and number of variables).** 
  + There are 12 variables and each variable has 70,000 observations.
* **c. Description of all variables (similarly to the descriptions you encounter for R data sets).** 
  + Age: Age of individuals (integer)
  + Gender: Gender of individuals (string)
  + Height: height of individuals in centimeter (integer)
  + Weight: weight of individuals in kilograms (integer)
  + Ap\_hi: Systolic blood pressure reading. (Integer)
  + Ap\_lo: Diastolic blood pressure reading. (Integer)
  + Cholesterol: Cholesterol level of the individual. (Integer)
  + Gluc: Glucose level of the individual. (Integer)
  + Smoke: Smoking status of the individual. (Boolean)
  + Alco: Alcohol consumption status of the individual. (Boolean)
  + Active: Physical activity level of the individual. (Boolean)
  + Cardio: Presence or absence of cardiovascular disease. (Boolean)

**3. Decide on main questions you would like to ask for your data. The most critical steps would typically be**

* **a. Determining the variable(s) you will use as response, and potential variables you will use as predictors.** 
  + **Response:** cardio (presence/ absence)
  + **Predictors:** age, gender, height, weight, ap\_hi, ap\_lo, cholesterol, gluc, smoke, alco, active
* **b. Figuring out if you prioritize prediction or inference as your final goal(or maybe you’d like to attempt both).** 
  + Leaning towards Both
* **c. Formulating the data question(s) clearly.** 
  + Key findings and insights which can affect the results.
  + We want to find which parameters have the greatest impact on cardiovascular health. We predict it will be x,y, and z based on our own assumptions about cardiovascular health.

**4. Outline the models and methods you anticipate using in order to answer those questions and address those tasks. Minimal requirements are:**

* **a. Make sure to use at least two distinct models from the ones covered in this course (two out of: linear regression, logistic regression, decision trees, random forests, neural networks), pointing to reasons and advantages of each model over the others.** 
  + **Logistic Regression:** logistic regression is a useful tool for analyzing cardiovascular data because it is specifically designed to predict binary outcomes, provides interpretable results, has low computational requirements, and is robust to noise in the data. (overfitting)
  + **Random Forests**: Overall, random forests are a powerful and versatile algorithm that can provide several benefits when analyzing cardiovascular data. They can handle high-dimensional data, avoid overfitting, handle imbalanced datasets, and identify important features, making them a popular choice for machine learning tasks in the field of cardiovascular research.
* **b. Make sure to perform model comparison via resampling methods introduced in this course (mainly cross-validation).** 
  + Leave-one-out Cross Validation

**5. How do you plan on distributing the workload across the group members?**

* **E.g.** 
  + **"Members 1 & 2 will be implementing and testing model A , while members 3, 4 & 5 will focus on model B (both models should be mentioned in 2(a))" Or**
  + **"Members 1 & 2 will work on data question A, while other members will deal with data question B (both questions should be formulated in 1(c)).”**
  + **Jason Mullen:** I plan to work on splitting, training, and testing the dataset and then implementing the Random Forest Model/Algorithm on data.
  + **Chuong Tran**: I plan to perform Logistic Regression on data.
  + **Trang Hoang:** I will be doing the conclusion and summary of the analysis
  + **Vinh Nguyen**: I will work on model comparison for the two models discussed above using LOOCV
  + **Nathan Lilly:** I plan to focus on random forests and cross validation.