BIOS 735 Group 1 Project Proposal

**Heart Attack Prediction in the United States**

Ambroise Rachelle, Davis Teya, Qaqish, Ameer, Shilin Yu, and Jingying Wang.

**Link to Dataset**: <https://www.kaggle.com/datasets/ankushpanday2/heart-attack-prediction-in-united-states>

**Citation:** Ankit. (2025). Heart Attack Prediction in United States. <https://doi.org/10.34740/KAGGLE/DSV/10806451>

**Background**: This dataset contains information about clinical, lifestyle, and demographic factors that may contribute to heart attack risk for 372,974 patients in the U.S., along with their health outcome (heart attack or no heart attack). Among the 32 variables are:

* (Clinical) age, gender, cholesterol, blood pressure, heart rate, BMI, diabetes, hypertension, family history of heart attack, medication, chest pain type, ECG results, max heart rate, ST depression, exercise induced angina, slope, number of major vessels, thalassemia, previous heart attack history, stroke history
* (Lifestyle) smoker status, physical activity, alcohol consumption, diet, stress level
* (Demographic) ethnicity, income, education level, residence, employment, marital status

(Note: The variables do not come stratified in these groups in the dataset.)

**Questions**:

* What parameters are most predictive of a heart attack?
* Can we use the data to create an accurate risk assessment of the likelihood that a patient will experience a heart attack?
  + Note: Clinical cardiology practices use their own risk estimator (<https://tools.acc.org/ascvd-risk-estimator-plus/#!/calculate/estimate/>) which uses fewer variables than those provided by this dataset.

**Aims**:

1. Assess the critical factors in heart attack classification through Probabilistic Principal Component Analysis (PPCA).
2. Evaluate how Probabilistic Principal Component Analysis (PPCA) handles missing data.
3. Estimate through maximum-likelihood estimation of parameters in a latent analysis, implement EM algorithm to refine parameter estimates or for latent variable modeling to predict heart attack risk.
4. Use a random forest classification model with 20% cross-validation, incorporating all 30 variables. Compare the classification performance with the Random Forest model in Module 3.
5. Create a risk assessment tool that assigns probability of heart attack given a patient's clinical, demographic, and lifestyle information

**Methods:**

Program probabilistic PCA (https://www.robots.ox.ac.uk/~cvrg/hilary2006/ppca.pdf) to handle missingness in the covariates, e.g. cases where each subject is missing different covariates. This requires nontrivial use of the EM algorithm. Test PPCA's dimension reduction capability by the following:

* Take a dataset {(x1, y1),…,(xn, yn)} where **xi** is the covariate vector and **yi** is a categorical outcome associated to the ith subject.
* For each covariate, randomly make p% of its values missing.
* Train PPCA and use the low dimensional projections of the covariate vectors to train a logistic regression classifier for y.
* Compare the performance of this PPCA classifier across various values of p and compare it to other classifiers that handle missing covariates.