**Group 1: Rams**

**Heart Attack Prediction in the United States**

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

**Citation:** Ankit. (2025). Heart Attack Prediction in United States [Data set]. Kaggle. <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. Define a probabilistic model (e.g. logistic regression model based on Section 3.2.1.1 of optim lecture of Module 2) for heart attack prediction, and estimate parameters for desired outcome (no heart attack) and those for adverse outcome (heart attack)
2. Implement EM algorithm to refine parameter estimates or for latent variable modeling, and compare to that of ML method(s) to predict heart attack risk
3. Compare with the Random Forest model in Module 3
4. 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.