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 get low dimensional projections z\_i = f(x\_i) 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.

**Responses to Clarifying Questions:**

(1) Is there class imbalance in the dataset for your outcome variable? If so, how would this impact your analysis? Are there ways to address it in the specific models you are using?

A: There is no class imbalance in this dataset; both sex and the outcome variable (history of heart attack) have approximately a 50/50 distribution.

(2) I would set aside 10-20% of your samples as an independent test set to compare the final performance of each class of models, and use the remaining samples to do your model fitting/hyperparameter tuning/variable selection etc

A: We plan to set aside approximately 15–20% of the data as an independent test set to evaluate final model performance. We will ensure that the test set is stratified by both sex and outcome to preserve proportional representation across subsets. The remaining data will be used for model training, hyperparameter tuning, and variable selection.

(3) What performance metric will you use to guide hyperparameter tuning and also the final model evaluation on the independent test set? Why did you choose that metric?

A: Since we are focusing on classification, we will use Youden’s Index of J-statistic (J = sensitivity + specificity – 1) as our primary performance metric. This index balances sensitivity and specificity, and is particularly suitable when false positives and false negatives are equally important. It also provides a meaningful way to determine the optimal classification threshold.

(4) Does the dataset already have missingness in the dataset? If so, how does the 2nd bullet point under methods work with entries that are already missing? Also, one must make specific assumptions of the missingness mechanism for the existing missing entries. This also has implications in how one would create missing values in the 2nd bullet point. Would they match? I would imagine one would have to do a sensitivity analysis under the different assumption (MCAR, MAR, MNAR) as don’t know what it is in advance.

A: We will introduce approximately 15% missing values under the Missing Completely At Random (MCAR) assumption.

(5) For the other methods, are you applying them to the output of ppca, or are you training the models directly using the provided covariates?

A: We will use Probabilistic PCA (PPCA) to handle missing data and extract principal components. The classification models will be trained using the outputs of PPCA. That is, our parameter estimation and model training will be based on the low-dimensional PPCA representations.

(6) Will you be performing variable selection? If so, how will you do this?

A: We do not perform variable selection directly on the original covariates in the PPCA-based models. Instead, we select the number of principal components based on model fit criteria such as the Akaike Information Criterion (AIC). In models trained on the original (non-PPCA) data, we will use elastic net regularization for variable selection and model comparison.

(7) What other classifiers will you be comparing this approach to?

A: We will compare model performance across two dimensions:

PPCA vs. No PPCA: Comparing models trained on PPCA-imputed data versus models trained on data with missing entries removed (complete-case analysis).

PPCA + GLM vs. PPCA + other classifiers: We will compare the performance of a generalized linear model to other classifiers such as random forests and support vector machines, all using the same PPCA-based inputs.