Team Four

# Topic Proposal

Heart Disease Health Indicators dataset is from Kaggle. It provides information on how various factors affect a person causing a heart disease or a heart attack. Presently there are a total of 253,680 entries in the dataset with 22 features which includes: heart disease or attack, high BP, high cholesterol, cholesterol check, BMI, smoker, stroke, diabetes, physical activity, fruits, veggies, heavy alcohol consumption, any health care, no doctor because cost, general health, mental health, physical health, different walk, sex, age, education, and income. All the features are of float data type. The research topic for this project is: What are the most accurate predictors of the likelihood of having a heart disease or heart attack?

# SMART Questions

1. Does age have an effect on heart disease?
2. Does having high BP have an effect on heart disease?
3. How does the level of physical activity contribute to one having a heart attack or disease?
4. Does having high cholesterol impact one having heart attack or disease?
5. Does heavy alcohol consumption cause heart attack or disease?
6. Does having variable income cause heart attack or disease?
7. How BMI can affect the health of the heart or heart condition?
8. How consuming veggies and fruits can affect the heart condition?

# Modeling Methods

1. SMOTE (Balancing Technique)
2. Naïve Bayes Classifier
3. Logistic Regression
4. Decision Tree Classifier
5. Random Forest Classifier
6. XGBoost Classifier

# Source

**The source of our data set:** Heart Disease Health Indicators on Kaggle: (<https://www.kaggle.com/datasets/alexteboul/heart-disease-health-indicators-dataset>)

**Number of Observations:** 253,680.

# GitHub

The link to our GitHub repository: <https://github.com/akhil97/Data_Mining_project-T4-DATS6103>