# DS 502/MA543: Statistical Methods for Data Science

## Project Proposal – Group 5

## Heart Attack Analysis & Prediction

**Members' Names:**

- Atharva Pradip Kulkarni
- Chithramvel Sanarpalayam Selvamuthukumar
- Nitya Phani Santosh Oruganty
- Shubham Wagh
- Sreeram Marimuthu

**Description of the Problem:**

This project aims to develop a predictive model for heart attack classification using the Heart Attack Analysis & Prediction Dataset from the UCI ML Repository. Heart attacks are a leading cause of death worldwide, and early detection and classification of individuals at risk are crucial for timely intervention and prevention. This project seeks to create a machine learning model that can predict the likelihood of a person having a heart attack based on their medical data.

**Description of the Dataset:**

The Heart Attack Analysis & Prediction Dataset contains the following variables:

1. **Age:** The age of the patient.
2. **Sex:** The gender of the patient (0 for female, 1 for male).
3. **CP (Chest Pain Type):** Describes the type of chest pain the patient is experiencing (0 - asymptomatic, 1 - typical angina, 2 - atypical angina, 3 - non-anginal pain).
4. **Resting Blood Pressure:** The resting blood pressure of the patient.
5. **Cholesterol:** The cholesterol levels of the patient.
6. **Fasting Blood Sugar:** Indicates whether the fasting blood sugar is greater than 120 mg/dl (0 for no, 1 for yes).
7. **Resting ECG (Electrocardiographic Results):** Describes the results of the resting electrocardiogram (0 - normal, 1 - having ST-T wave abnormality, 2 - showing probable or definite left ventricular hypertrophy).
8. **Max Heart Rate:** The maximum heart rate achieved by the patient.
9. **Exercise-Induced Angina:** Indicates whether the patient experienced exercise-induced angina (0 for no, 1 for yes).
10. **ST Depression:** The ST depression induced by exercise relative to rest.
11. **Slope:** Describes the slope of the peak exercise ST segment (0 – down sloping, 1 - flat, 2 - upsloping).

12. **CA (Number of Major Vessels Coloured by Fluoroscopy):** The number of major vessels coloured by fluoroscopy.
13. **Thal (Thalassemia):** Describes the type of thalassemia the patient has (0 - normal, 1 - fixed defect, 2 - reversible defect).
14. **Target:** Indicates whether the patient had a heart attack (0 for no, 1 for yes).

**Regression or Classification?**

This project involves a classification task. We will predict whether a patient is likely to have a heart attack or not, which is a binary classification problem (0 for no heart attack, 1 for heart attack).

**Methodology:**

- **Data Pre-processing:** We will clean and pre-process the dataset, handling missing values and encoding categorical variables.
- **Feature Selection:** We will explore feature importance and select the most relevant variables for the model.
- **Model Selection:** We will experiment with various classification algorithms such as Logistic Regression, Tree Based Algorithms & Support Vector Machines.
- **Model Evaluation:** We will use appropriate evaluation metrics like accuracy, precision, recall, F1-score, and ROC-AUC to assess the model's performance.
- **Hyperparameter Tuning:** We will fine-tune the chosen model for optimal performance.
- **Interpretability:** We will investigate feature importance and provide insights into the factors contributing to heart attack prediction.

**Comments and/or Concerns:**

- **Data imbalance:** We need to be cautious about data imbalance, as the number of heart attack cases may be significantly lower than non-heart attack cases.
- **Model interpretability:** We are aiming for a transparent model that can provide interpretable results to medical professionals.
- **High Precision:** Given that our proposed model is trying to determine whether a patient is going to have a heart attack or not, it should have a very low possibility of a False Negatives.