|  |
| --- |
| Figure |
| Project Proposal  By: Alex Gorbulin, Aida Ndiongue, Christopher Sharber & Rocky Owens |
| |  |  |  | | --- | --- | --- | | 5/23/23 |  | Data Analytics Dec 2022 session | |

Predicting Heart Disease

# Introduction:

Coronary Heart Disease often referred to simply as “heart disease”, although not the only type of heart disease occurs when plaque (a combination of fat, cholesterol, calcium, and other substances found in the blood) builds up in your arteries[[1]](#footnote-1).

According to the CDC, about, 695,000 people died from heart disease in 2021, that is 1 in 5 deaths[[2]](#footnote-2). What if we could build a model to help predict individuals at risk of developing heart disease?

That is the heart of our project where we seek to develop an accurate predictive model that can identify individuals at risk of developing heart disease within the next ten years.

# Abstract:

The dataset with over 4200 records, 16 columns and 15 attributes, consists of various demographic, lifestyle, and health-related features of individuals, including age, education level, smoking status, blood pressure, cholesterol levels, body mass index (BMI), and glucose levels, among others.   
The project will involve several key steps. First, the dataset will be preprocessed to handle missing values and convert categorical variables into numerical representations. Data exploration and visualization techniques will then be used to gain insights into the distribution of features and their relationships with the target variable, “TenYearCHD”. We will then use feature selection methods to identify the most relevant features for heart disease prediction.

Next, the dataset will be split into training and testing sets, and a variety of machine learning algorithms, including logistic regression, decision trees, random forests, support vector machines, and gradient boosting algorithms, will be considered for model selection. The selected models will be trained on the training set and evaluated using metrics such as accuracy, precision, recall and F1-score.  
The trained models will then be interpreted to understand the significant features contributing to heart disease prediction. Finally, the best-performing model will be deployed for real-time heart disease prediction on new, unseen data.

By undertaking this project, we aim to provide an effective machine learning solution for predicting heart disease, which can aid in early detection and preventive measures for individuals at risk. The project adheres to good data science practices, considering data biases, validation procedures, and ethical considerations in dealing with sensitive health information.

# Data source:

Kaggle

<https://www.kaggle.com/datasets/aasheesh200/framingham-heart-study-dataset?resource=download>[[3]](#footnote-3)

# Timeline:

1. Dataset – due 5/22 (end of class)
2. Proposal – due 5/23, team review by 5/23 (end of day)
3. Data cleansing – due 5/25 (class time), team review by 5/25 (end of class)
4. Visualizations (pre-ML): all members – due 5/26, team review by 5/26 (end of day)
5. Optimization (hyper parameter tuning logic); all members – due 5/29 (class time), team review by 5/29 (in class)
6. Visualizations (post-ML): all members – due 5/30 (end of day), team review by 5/31 (class time)
7. Slide deck & read me file – due 6/1 (end of day), team review by 6/2 (end of day)
8. Group presentations – on 6/5

1. *https://www.nhlbi.nih.gov/sites/default/files/publications/FactSheetKnowDiffDesign2020V4a.pdf* [↑](#footnote-ref-1)
2. *https://www.cdc.gov/heartdisease/facts.htm* [↑](#footnote-ref-2)
3. *By Ashish Bhardwaj, updated a year ago.*

   *Cover page image credit: https://www.indiatoday.in/india/story/heart-attack-brain-stroke-survey-young-middle-aged-indians-covid-gym-walking-dance-2308662-2022-12-13* [↑](#footnote-ref-3)