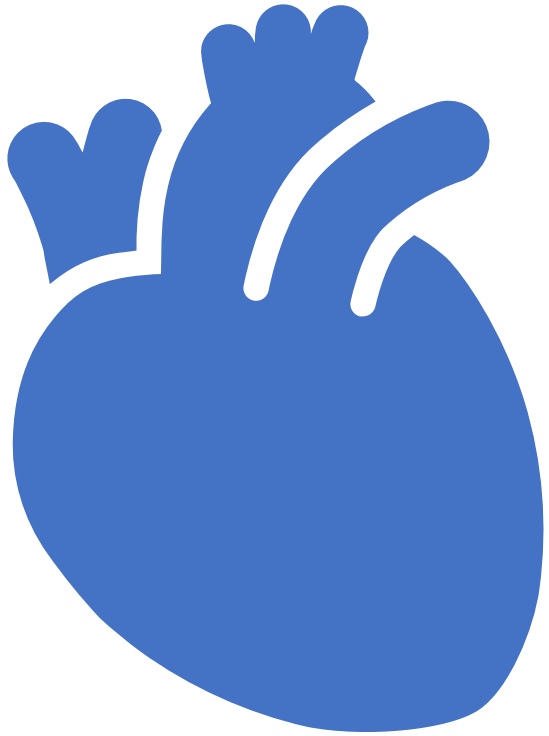


DETECTING HEART DISEASE

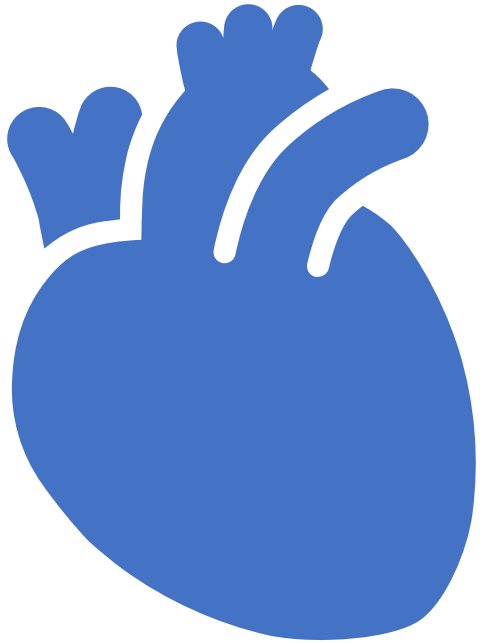
A DATA INTELLIGENCE FUSION

Data Science Capstone Introduction

By: Roland Tetteh

Problem Statement

- Globally it's estimated that 1 in 13 people are living with a heart or circulatory disease.
- In the United States, one person dies every 33 seconds from heart disease.
- According to the CDC about 5% of adults, age 20 and older have Coronary Artery Disease.
- In the United States, someone has a heart attack every 40 seconds.



Proposed Data Science Solution

In healthcare, identifying and preventing the factors that have the greatest impact on heart disease is very important.

Machine learning algorithms can detect "patterns" in data, using these risk factors, to predict a patient's condition or propensity to develop a heart disease.

This project seeks to apply machine learning techniques on a 2022 annual CDC survey data, of 400k+ adults related to their health status, to predict an individual's propensity to a heart attack.

Potential Impact

Early detection and risk assessment of heart disease.

Personalized risk assessment allowing for targeted interventions and prevention strategies.

Potentially be used in drug discovery.

Data Quality

- The dataset comes from the CDC, which conducts annual telephone surveys to collect data on the health status of U.S. residents.
- The original dataset of nearly 300 variables was reduced by the author to 40 most relevant variables.
- There were 157 duplicate rows out of the 445,132 rows and missing values in 38 out of the 40 columns.
- Some of the attributes in the data are State, Sex, MentalHealthDays, and HadDiabetes.
- There is a significant disparity between the people who responded ‘Yes’ or ‘No’ to having Heart Attacks.

Preliminary EDA

- Missing values were imputed with 'unknown' categories across all the categorical variables as there is the possibility, they carry some meaning.
- Individuals in West Virginia are more likely to have a heart attack.
- People who have had a heart attack and people in the unknown category are more likely to have had more mental health days.
- A current smoker who smokes every day is more likely to have a heart attack.

Next Steps...

Detecting Multicollinearity with the Variance Inflation Factors.

Baseline model will be a logistic regression.

Simplify the model using variable selection (backward selection).