MGT - 6203 Group 30 Project Proposal

# TEAM INFORMATION

## Team #: 30

## Team Members:

| **#** | **Name** | **GT ID** | **Current Role** | **Background** | **Previous Analytics Projects** |
| --- | --- | --- | --- | --- | --- |
| 1. | Jin Woo Oh | joh78 | Software Quality Engineer | 4 years in data analytics for consumer electronics; 2.5 years in research in soft robotics & 3D printing; background in mechanical engineering. | - Developed a BERT text classification model on Python to identify critical technical issues in company products  - Created an ETL process with Python web scraper and SQL database  - Used Power BI to create & monitor QA dashboard |
| 2. | Zach Smith | zsmith73 | Data Scientist | 14 years in data and analytics - 5 yrs in fMRI research; 2.5 yrs as a electro-mechanical test and reliability engineer; 6.5 yrs as a data scientist. | - Building and designing a document generation project utilizing LLMs  - Developed and implemented a complaint classification system in a quality management platform for manufacturing |
| 3. | Wei-Yung Liu | wliu432 | Data Analyst | 1 year data analytics in environmental health. 3 years in research and quality regulation. | - Developed decision-support tools for safer material selection (python, SQL).  - User segmentation analysis. |
| 4. | Gloriana Hernández Monge | gmonge6 | Product Specialist | 5 years experience in the Cybersecurity Industry (Sales Engineering and Product / Program Management). | - Data Analysis - Visualization (Looker): designed and built Business Intel. deliverables.  - Responsible for data modeling and program management for a company wide BI infrastructure. |

# OBJECTIVE/PROBLEM

## Project Title:

Modeling and Analysis of Key Indicators of Heart Disease

## Background Information:

According to the [Center for Disease Control and Prevention (CDC](https://www.cdc.gov/heartdisease/risk_factors.htm)), heart disease is one of the leading causes of death for people of most races in the US (African Americans, American Indians and Alaska Natives, etc), where about half of all Americans (47%) have at least 1 of 3 key risk factors for heart disease: high blood pressure, high cholesterol, and smoking1.

* High blood pressure is a medical condition that occurs when the blood pressure in the arteries and other blood vessels is too high. It is known as a "silent killer" due to its lack of symptoms, and can affect the kidneys and the brain. The only way to detect this condition is by measuring the blood pressure.
* Cholesterol is a fatty substance produced by the liver and found in certain foods. Overconsumption of cholesterol can build up walls in the arteries, including those in the heart, which leads to decreased blood flow to the major organs in the body (heart, brain, kidney, etc).

There are two main types of blood cholesterol. LDL (low-density lipoprotein) cholesterol is the “bad” cholesterol that can cause plaque buildup in arteries, and HDL (high-density lipoprotein) cholesterol is the "good" cholesterol that can provide some protection against heart disease.

High cholesterol has no symptoms, and therefore must be checked via a blood test ("lipid profile").

* Cigarette smoking can damage the heart and blood vessels, which increases risk for heart conditions such as atherosclerosis and heart attack. Nicotine raises blood pressure, and carbon monoxide from cigarette smoke reduces the amount of oxygen that your blood can carry. Furthermore, exposure to secondhand smoke can also increase the risk for heart disease, even for nonsmokers.
* Other traits, such as age, hereditary factors, obesity and diabetes can make one more susceptible to developing a heart disease.

## Problem Statement:

As has been the case with COVID-19, the rise of a new, infectious respiratory virus can be especially fatal to those with chronic cardiovascular health conditions. Individuals without diagnosed conditions, but have an unhealthy lifestyle with poor diet and lack of exercise, may also be at a significant level of risk without realizing. The analysis of the “Personal Key Indicators of Heart Disease” dataset will help determine the variables (such as BMI, history of stroke or diabetes, etc) that are most significantly correlated to the diagnosis of a heart disease (coronary heart disease (CHD) or myocardial infarction (MI)), and provide insight into predicting undiagnosed individuals who are at an increased level of risk to the heart disease, and hence the virus. Conclusions of the study can be used to inform and guide civilians with the appropriate knowledge, and empower them to secure their and their family’s health.

## Primary Research Question (RQ):

What are some of the most significant contributors to the diagnosis of heart disease?

## Supporting Research Questions:

* + 1. Are there interactions of variables that lead to increased risk of heart attack?
    2. Is there any correlation among certain variables or collinearity? If so, interpret the meaning of such a phenomenon and suggest a course of action.
    3. Is it possible to identify groupings of people for which the most significant factors affecting risk change?

## Among various algorithms suitable for modeling the dataset, which three were most effective in comprehensively describing the data? Evaluate their model quality.

## Business Justification:

## According to [CDC](https://www.cdc.gov/chronicdisease/about/costs/)1, heart disease and stroke are the leading causes of death for Americans (more than 877,500 deaths from heart disease or stroke annually - contributing to a third of all deaths). These diseases cost our healthcare system $216 billion per year and cause $147 billion in lost productivity on the job. In recent history, the outbreak of COVID-19 posed a serious health risk for individuals with pre-existing cardiovascular health conditions.

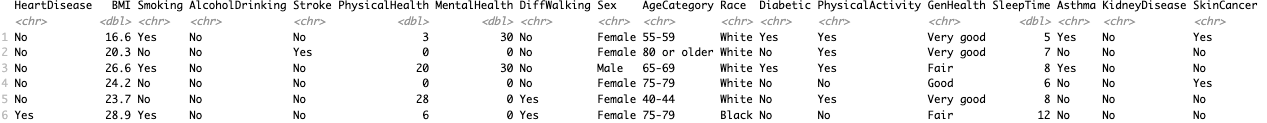
## 

## Early diagnosis and prevention of heart disease can save hundreds of billions of dollars for the healthcare system. Conducting data analysis on the indicators of heart disease can help determine the key indicators of the onset of the disease and provide insight for the appropriate actions and treatments required to minimize the risk of developing them, as well as the chances of severe, permanent health effects from COVID-19.

# DATASET/PLAN FOR DATA

**Data Source:** [Personal Key Indicators of Heart Disease (2020)](https://www.kaggle.com/datasets/kamilpytlak/personal-key-indicators-of-heart-disease), this Kaggle subset was originally obtained from the CDC’s 2020 ‘[Behavioral Risk Factor Surveillance System](https://www.cdc.gov/brfss/index.html)’ project, which collects data about the health status of United States adult residents.

**Data Description:** This data source includes health data from 319 795 individuals, 18 variables in total. Fourteen of those variables are categorical, the rest are continuous.



**Figure 1**. Data sample from the ‘Personal Key Indicators of Heart Disease’ dataset

## Key Variables:

* ***Dependent variable***: HeartDisease (Yes/No), indicating whether a person has or hasn’t heart disease.
* ***Independent variables:*** 17 factors related to health status:
  + **BMI** (12.01-94.85), **Smoking** (Yes/No), **AlcoholDrinking** (Yes/No), **Stroke** (Yes/No), **PhysicalHealth** (0-30), **MentalHealth** (0-30), **DiffWalking** (Yes/No), **Sex** (Female/Male), **AgeCategory** (55-59/80 or older/65-69/75-79/40-44/70-74/60-64/50-54/45-49/18-24/ 35-39/30-34/25-29), **Race** (White/Black/Asian/American Indian-Alaskan Native/Hispanic/Other), **Diabetic** (Yes/No/No,borderline diabetes/Yes (during pregnancy)), **PhysicalActivity** (Yes/No), **GenHealth**(Very good/Fair/Good/Poor/Excellent), **SleepTime**(1-24), **Asthma** (Yes/No), **KidneyDisease** (Yes/No) and **SkinCancer** (Yes/No).
* ***Hypothesized most important variables:*** according to the CDC, the following traits have been associated with Heart disease: Stroke, AgeCategory, Smoking, Diabetic.
* ***New variables:*** since several independent variables are categorical, dummy variables will be created to incorporate each level in the analysis.

# APPROACH/METHODOLOGY

## Planned Approach:

In order to evaluate our main research question, we plan to develop a classification model that can predict the likelihood of heart disease. We aim to use this model to evaluate the features that are most highly associated with heart disease. However, since we believe that the most influential features that lead to heart disease are unlikely to be consistent across population groups, we intend to implement a clustering algorithm to detect populations and assess the critical features for each population group.

First, we plan to do basic data cleansing like duplicate removal and handling of null values – either interpolation or removal of the whole row. Then, we plan to do feature engineering, which entails: one-hot encoding categorical variables – Race, Diabetic, and GenHealth; enumerating ordered categorical variables – AgeCategory; normalizing the range of continuous variables. Next, we will investigate variables for normality and variability of variance to determine if we need to apply any transformations to the data. Finally, we plan to run outlier detection using Cook’s distance.

After all features are engineered, we plan to use a stepwise logistic regression model as a method of feature selection. Once the features have been down-selected, we plan to apply clustering through the k-Nearest Neighbor algorithm to detect unique communities within the data. Additional clustering algorithms will only be evaluated if k-NN produces abysmal results.

We plan to develop a classification model to detect risk given our set of features. We plan to experiment with and compare models such as logistic regression, XGBoost, and AdaBoost. For comparison we will evaluate using a confusion matrix to calculate accuracy, precision, recall, and f1-score. However, given our model requires both precision and recall to be both high and in balance, we will emphasize f1-score in our analysis. F1-score will also be the metric we utilize during hyperparameter tuning with cross-validation.

Finally, utilizing ANOVA allows us to evaluate the significance of our features across community groups to detect if different features are significant for different communities of people.

## Anticipated Conclusions/Hypothesis:

Our null hypothesis is that there is no statistically significant relationship between any of the independent variables and the diagnosis of heart disease. Our initial hypothesis is that age and health background such as stroke, diabetic, and smoking history will exhibit significant positive correlation with the risk of heart disease. Whereas physical activity and sleep time will have negative correlation.

Given the inherent complexity of health determinants, we anticipate to see some interactions and/or collinearities of exploratory variables that lead to increased risk of heart attack. Furthermore, we expect to find differences in the risk of heart disease among groups of individuals with common risk factors and the most impactful risk factors may differ across groups of people. We based these hypotheses on CDC’s resource on heart disease and recognize that our hypothesis may change as we analyze the data and evaluate our model performances.

## What business decisions will be impacted by the results of your analysis? What could be some benefits?

Heart disease is a leading cause of death in the United States and the economic cost is substantial1. We hope to determine the key indicators to help predict the conditions of the individuals who may be at an increased level of risk of heart disease. We believe the insights gained from our analysis will have significant implications for healthcare decision-making, cost saving strategies, and improved well-being of our communities.

Understanding the intricate relationship between personal indicators and heart disease is key to early detection and precision healthcare strategies. Our analysis has the potential to detect and prevent heart disease in its nascent stages by identifying individuals at elevated risk. Targeting screenings and interventions based on our analysis enables timely intervention to reduce future healthcare expenses and enhance patient outcomes.

Public health programs based on our findings can show the importance of managing personal risk factors. Empowered with the appropriate knowledge, civilians can make informed, healthier lifestyle choices and reduced healthcare expenditure. In addition, our insights can support healthcare systems and the government to enhance their preparedness for future infectious disease outbreaks. With accurate prediction of individuals at higher risk, stakeholders can take proactive approaches to safeguard vulnerable populations and minimize the social and economic impact of such crises.

Beyond identifying key risk factors to heart diseases, our exploration of interaction/collinearity among influential indicators and identification of unique groups add an extra layer of understanding. Our multi-faceted approach will not only provide insights but also have the potential to transform decision-making in the healthcare industry and beyond.

## 1Centers for Disease Control and Prevention. (2023, March 21). Know Your Risk for Heart Disease. CDC. https://www.cdc.gov/heartdisease/risk\_factors.htm

## 2Centers for Disease Control and Prevention. (2023, March 23). Health and Economic Costs of Chronic Diseases. CDC. https://www.cdc.gov/chronicdisease/about/costs/

## 3Vidal-Perez, Rafael *et al*. “Cardiovascular disease and COVID-19, a deadly combination: A review about direct and indirect impact of a pandemic.” World journal of clinical cases vol. 10,27 (2022).

# PROJECT TIMELINE/PLANNING

## Project Timeline:



**Figure 2: Project timeline with key deliverable dates and status on milestones**