**Module 5 - Final Project Assignment (part 1)**

**Team 5: Rion Askin, Brad Hall, Justin Holmes, Hanan Afzal**

**Frame The Problem**

1. Define the objectives or possible outcomes of the project

The goal of this project is to develop a machine learning algorithm to predict the likelihood of coronary heart disease (CHD) based on various demographic, lifestyle, and medical attributes of patients. The dataset contains information on gender, education level, age, smoking habits, medical history, cholesterol levels, blood pressure, BMI, heart rate, and glucose levels. The primary objective of this project is to build a robust and accurate predictive model that can help identify individuals at high risk of developing CHD in the next ten years. Possible outcomes include accurately classifying individuals as either high or low risk for CHD based on patient data and developing insights into the most significant risk factors for CHD development, aiding in risk mitigation and preventive measures.

1. Discuss how the solution will be used.

The developed solution can be used as a decision support tool in healthcare settings. Healthcare providers can utilize the model to screen patients and identify those at higher risk of CHD, allowing for early intervention and tailored preventive care strategies. Additionally, patients can also benefit from the tool by gaining awareness of their potential risk and adopting lifestyle changes to improve their heart health.

1. Identify any current solutions

There exist risk assessment models for heart disease based on traditional statistical methods. Healthcare practitioners often rely on clinical guidelines and scoring systems, such as the Framingham Risk Score, to estimate cardiovascular risk.

* 1. Strengths of current solution
     1. Traditional scoring systems are well-established and widely used in clinical practice, which means they have been validated on large cohorts over many years. The most well-known scoring system is the Framingham risk score. They are relatively straightforward to implement and do not require extensive computational resources. Moreover, they have proven to be valuable in identifying individuals at risk of CHD.
  2. Shortcomings of current solution
     1. Traditional scoring systems may not consider complex interactions between various risk factors, limiting their accuracy. They might not adapt well to changes in the population's risk factors over time. Additionally, these methods might not leverage the full potential of the dataset, missing out on valuable insights that modern machine learning algorithms can provide. Does not predict future cardiovascular events such as stroke, heart failure, and transient ischemic attack (TIA). Not usable for all Americans (Hispanic/Latino and Native Americans).

1. Define how success will be measured

A multiple linear regression model will be constructed. Success will be measured by a high R^2 value (.8 or higher), and weights of predictors that help predict the likelihood of ten year CHD.

1. Compare to similar projects

Similar projects in the literature may involve predicting heart disease using machine learning techniques. The success of this project will be benchmarked against the performance of those models. If the developed model outperforms or matches the accuracy of state-of-the-art approaches, it will be considered a successful contribution to the field.

1. List and verify assumptions
2. The provided dataset is accurate, reliable, and representative of the target population.
3. The target variable "ten year chd" is a reliable indicator of CHD risk, and the data collection process for this variable is standardized and consistent.
4. The attributes in the dataset are relevant and have been preprocessed appropriately to handle any missing or erroneous data.
5. The data used to train the model is from a time period that reasonably represents the current population's risk factors.
6. There are no hidden or unmeasured confounding variables that could significantly impact the model's predictions.