Business understanding:

Background: Heart Disease is among the most prevalent chronic diseases, impacting millions of people each year. In the United States alone, heart disease claims roughly 647,000 lives each year — making it the leading cause of death. While there are different types of coronary heart disease, most individuals only learn they have the disease following symptoms such as chest pain, a heart attack, or sudden cardiac arrest.

The Behavioural Risk Factor Surveillance System (BRFSS) is a health-related telephone survey that is collected annually by the CDC. Each year, the survey collects responses from over 400000 Americans on health-related risk behaviours, chronic health conditions, and the use of preventative services. It has been conducted every year since 1984.

Goal: The goal is to create a full data mining project for the UTs "Introduction to the Data Science" course. With this project, the team hopes to find, what might cause and decrease, the risk of heart disease in a person, based on the subset of the CDCs BRFSS health study conducted in 2015. Finally create a predictive ML model around it.

Success criteria: Mentioned in the "Goal" section.

Resources: The project team consists of two UT students, with limited/entry level knowledge of the data science field and a lab assistant, who will help and guide the project and the process if need be. Also, the team will use LLMs for assistance. The team will use GitHub to manage the workflow and Jupiter notebooks for data processing and analysis.

Constraints: The project must be finished by 11th of December 2023. Also, a poster must be made to present the findings.

Risk: The main problem is the time constraint, where there is not a lot of time to process the data, which might hinder the results or the effectiveness of the model. This also brings the risk of the project not getting a passing grade.

Terminology:

- Cholesterol - Vital molecule in the human body to regulate normal body function.

Cost and benefits: Main cost is time that must be allocated for the project. Additional time must be made for team meetings to analyse work done and go over each other's progress. Main benefits are that the team can test their newfound

knowledge in a full data mining project and upon successful project, get a passing grade for the course.

Data-mining goals: Processing the data, creating a predictive model to predict, if a person has a higher change of getting a heart disease (using an ensemble of classifiers) and applying hyperparameter optimization. Finally creating a poster to present the findings to others.

Data-mining success criteria: The model must have an AUC>0.7. Finding factors (features) that have a positive moderate (r>0.3) and negative moderate (r<-0.3) correlation with heart disease.