Predicting Coronary Heart Disease
Executive Summary

Cardiovascular disease refers to all diseases that affect the heart or blood vessels. The focal point of our project is on a subcategory of cardiovascular disease known as coronary heart disease, which is one of the leading causes of death in the United States (US). We utilized predictive modeling techniques to provide insights into the risk factors of coronary heart disease. Our project focuses on human factors to help people better understand how their lifestyles and cultures impact their coronary heart disease risk. Our goal is to empower people living in the US by providing them with knowledge of their coronary heart disease risks via easily accessible language. We seek to inform policymakers like county officials and healthcare providers on human risk factors impacting their communities.

Stakeholders: *People living in the United States, policymakers, county officials, healthcare providers*
KPI: *Mean Squared Error*

Data Set: We used the Center for Disease Control and Prevention's Interactive Atlas of Heart Disease and Stroke online mapping tool to gather county-level data on coronary heart disease within the continental US (Delaware, Hawaii, Alaska, Washington DC, and US territories were excluded from analysis).

Approach
Three predictive models created using three different analysis techniques--XGBoost, Gaussian Naïve Bayes, and Linear Regression--were developed to assess coronary heart disease risk factors. Each model was trained at the state level for counties before being aggregated and tested.

Results & Recommendations
Initial exploratory analysis found that the people most at risk for coronary heart disease were people with high cholesterol, people living in a household without a computer, and people over the age of 25 with less than four years of college education. The initial exploratory analysis also found that Asian Pacific Islanders and people living in households meeting median income criteria were the least at risk for coronary heart disease.

Using the mean squared error, we measured the accuracy of each predictive model. The XGBoost provided the most accurate model, and thus, its findings are reported here. According to the XGBoost model, the top three risk factors that predict coronary heart disease are: having high cholesterol, being in a household without a computer, and being 25+ years old without a 4-year college degree. Our model provides insights into coronary heart disease risk factors and demonstrates that access to education and technology impacts health. By increasing access to education and technology, communities can help reduce the likelihood of getting coronary heart disease.

Future Project Expansion
Our results focused on risk factors influencing coronary heart disease at the continental level. However, it is likely that these risk factors vary in importance based on individual differences within the US states. To better understand these differences, we propose that case studies be conducted on individual states. Some states lack access to healthcare resources more than other states. Future project expansion hopes to capture these healthcare resource discrepancies. Further exploration can also be done on states with the highest (e.g., Ohio) and lowest (e.g., Nevada) predictive accuracy (as measured by MSE) in our XGBoost model. Understanding why the states varied in predictive accuracy can help strengthen the model. Additionally, we would like to expand the model to include Delaware, Alaska, Hawaii, Washington DC, and the US territories.