# CHD Python Project

Hans LehnDorff, Isaac Johnson, Jesse Debolt

2023-07-24

**Project Proposal: Predicting the Prevalence of Coronary Heart Disease (CHD) in U.S. Counties**

**Introduction**

Coronary Heart Disease (CHD) is a significant health issue affecting millions of people worldwide. This project aims to predict the prevalence of CHD in different counties across the United States, focusing on potential health inequities in access to care between urban and rural areas. The hypothesis is that rural areas have worse health outcomes due to various factors, including limited access to healthcare facilities and resources.

**Research Questions**

1. What is the prevalence of CHD in different counties, and how does it vary between urban and rural areas?
2. What are the key health and social environment variables that influence the prevalence of CHD?
3. How do these variables interact, and what is their relative importance in predicting CHD prevalence?

**Data Sources**

The primary dataset for this project is sourced from the Centers for Disease Control and Prevention (CDC) Interactive Atlas of Heart Disease and Stroke. This dataset is a culmination of several datasets, providing a comprehensive view of various health and social environment variables at the county level. These variables include demographic factors, health indicators, healthcare access, and social and economic factors. More details about the data sources can be found here.

**Methodology**

The project will employ various machine learning techniques using Python, including Random Forest, Principal Component Analysis (PCA), Radial Basis Function (RBF), and K-Nearest Neighbors (KNN). The Python libraries to be used include scikit-learn for machine learning, pandas for data manipulation, numpy for numerical computations, and matplotlib and seaborn for data visualization.

The project will start with data cleaning and preprocessing, followed by exploratory data analysis to understand the data's characteristics. The machine learning models will be trained and tested using the processed data, with the aim of predicting the prevalence of CHD in different counties. The models' performance will be evaluated using appropriate metrics, and the results will be interpreted to answer the research questions.

**Expected Outcomes**

The project is expected to provide insights into the prevalence of CHD in different counties and the factors influencing it. It will also shed light on potential health inequities between urban and rural areas, contributing to the broader discourse on health disparities and access to care. The findings could inform policy-making and interventions aimed at reducing CHD prevalence and improving health outcomes, particularly in underserved areas.

**Conclusion**

This project will leverage machine learning techniques to analyze a comprehensive dataset from the CDC, aiming to predict CHD prevalence and understand its influencing factors. The findings could have significant implications for public health policy and practice, particularly in addressing health inequities.