Draft Outline – Final Project

DSE6111

Adeline Casali

Research Question #1:

What subset of predictors can be used for preventative health screening for heart disease?

Heart disease is the leading cause of death in the United States (CDC, 2023). This dataset contains 253680 survey responses from the 2015 Behavioral Risk Factor Surveillance System (BRFSS) Survey (Teboul, 2022). The variables include whether the respondents have or do not have heart disease as well as information on 21 different health-related risk behaviors and the use of preventative services. All variables are binary or ordinal. I will use classification methods such as Logistic Regression, LDA, QDA, Naïve Bayes, and KNN to determine which predictors are statistically significant and create a model that predicts heart disease as accurately as possible.

Research Question #2:

What indicators can be used to monitor for HIV response in children and adolescents?

As of 2022, 39 million people around the world were living with HIV (WHO, 2023). This dataset contains 173254 observations and includes information on the number of annual new HIV infections, mother-to-child transmission rate, number of annual AIDS related deaths, and number of people living with HIV (Bhandary, 2022). These values are grouped by country, region, year, sex, and age. I will use methods such as best subset selection, linear regression, ridge regression, lasso, PCR, and PLS to determine which predictors need to be used to create the most accurate model. I will determine the accuracy using bootstrapping or by creating a test and training dataset.

Sources:

Bhandary, K. (2022, July 19). HIV epidemiology indicators for children. Kaggle. https://www.kaggle.com/datasets/karthikbhandary2/hiv-epidemiology-indicators-for-children/data

Centers for Disease Control and Prevention. (2023, January 18). FASTSTATS - leading causes of death. National Center for Health Statistics. https://www.cdc.gov/nchs/fastats/leading-causes-of-death.htm

Teboul, A. (2022, March 10). Heart disease health indicators dataset. Kaggle. https://www.kaggle.com/datasets/alexteboul/heart-disease-health-indicators-dataset

World Health Organization. (2023). HIV. The Global Health Observatory. https://www.who.int/data/gho/data/themes/hiv-aids#:~:text=Globally%2C%2039.0%20million%20%5B33.1%E2%80%93,considerably%20between%20countries%20and%20regions.