United States Death Trends

First Capstone Project Proposal

Introduction

Study and analysis of Mortality and its causes is critical for the interpretation of national trends and international comparisons. Mortality statistics are a mirror to the situation of health prevalent in a population. The mortality trends depict the picture of population's health status. In depth analysis of deaths and its indicators can substantiate the progress of health programs

Problem:

The life expectancy of the United States is in decline. What are the leading causes for the deaths of Americans? Finding the major causes and change in epidemiological outbreaks can help mitigate some of the major death leading causes.

Goal:

Build a predictive model to predict the deaths causes and its prevalence in future population.

Specific Clients:

This project is aimed at biological and public health researchers. However, this can be very useful for government agencies, pharmaceutical companies and the general public. The approach used to classify different causes and its prediction based on age gender could potentially help government agencies to make better policies which could be instrumental in dealing and mitigating different health risk factors.

Data:

The data for this project comes from the Centers for Disease Control and Prevention (CDC). The dataset contains 10868 rows and 6 attributes depicting the information about the death rate in different states of United States with different causes between the year 1999 and 2017.

Approach

This project will explore the data for correlation between the state and cause of death. The total dataset is divided into training set and test set. 70% of data is taken for training sets and remaining 30% for test sets. Initially the model will be fitted in training sets and on the later phase test set will be used to provide an unbiased evaluation of a final model fit on the training sets. To determine which predictor variables are significantly associated with the multicategory outcome, Event type, categorical regression model such as multinomial logistic regression model will be built. Once the significant predictor variables are identified, a machine learning method like a classification decision tree will be trained to create the best prediction model possible.

Deliverables

The final deliverables for this project will be on GitHub. The deliverable will include the final Python Code. A final report will explain the approach in thorough details.