Executive Summary

Introduction

Access to healthcare is a vital need of society - the lack of having deadly consequences. In the United States of America, a persons' access is limited by their insurance status, the number of hospitals, and the number of healthcare professionals available. Further, insurance is limited by poverty, employment status, and preexisting health conditions.

The exploratory goal of this report is to find any correlation between a states' poverty levels, number of hospitals, mortality rate, and distribution of insurance types. The intent is to discover which states have the most robust healthcare coverage and which may be lacking in some of these key areas.

Two predictive machine learning models were developed: one which can determine the most likely underlying disease when given a list of symptoms - something which could be used to determine insurance availability; another that can predict the mortality rate when given the population, number of hospitals, poverty percentage, and uninsured population.

The overall goal of this analysis is to show that poverty, insurance, and hospital availability are intricately connected and can have devastating effects on the mortality rate within each state.

Process

A multivariate linear regression model was built using poverty count, uninsured population, and the number of hospitals available by state. The model was able to predict the mortality rate in a state using these three data values. The machine learning model had an accuracy of 95.63%. As for the disease prediction, a decision tree was used. This model predicted the disease of the 41 given from a set of symptoms with 100% accuracy.

Conclusion and Future Actions

The group found that the predictions made for the disease from symptoms and mortality rates from different variables yielded high accuracy results. However, when finding the correlative values for each of the variables worked with, with the exception of employer insurance and poverty, the R values were too low to conclusively answer the exploratory correlation questions. It is recommended that more investigation should be done to conclude if there is a relationship between the variables. Datasets that span longer ranges of time and data on the county level should be considered. This will allow future groups to examine if there are outliers that can be removed to improve results.