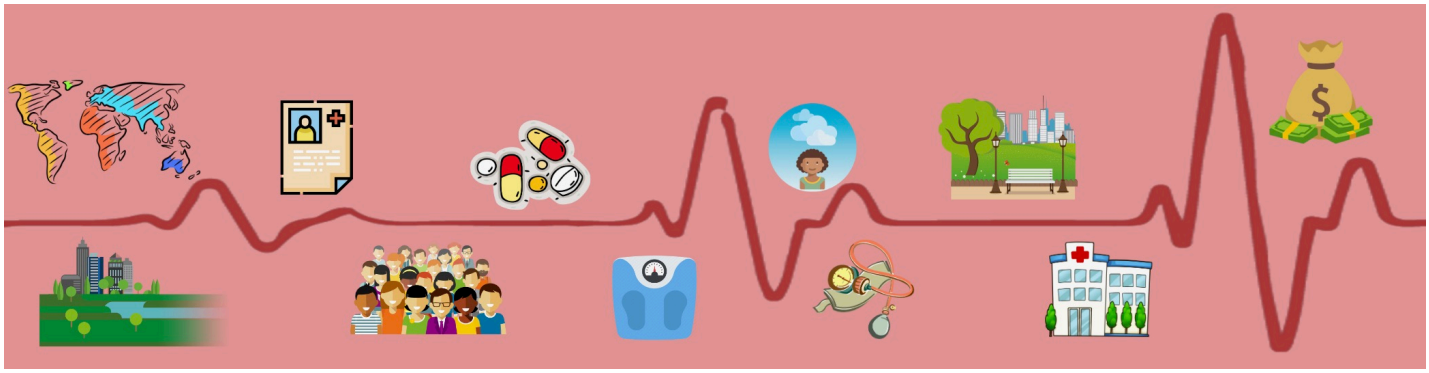
CardiovascularComplicationPrediction

Coronary Heart Disease Death Rate Risk-Level Prediction Based On Environmental, Non-Personal Parameters

Project for “CS4641 - Machine Learning” by Team 46: Aditya Kumar, Farouk Marhaba, Kinnera Banda, and Maya Rajan.



Introduction & Background

Coronary heart disease (CHD), the most common type of heart disease, kills over 300,000 people in the United States annually. It is caused by a buildup of plaque in the arteries that supply blood to the heart, limiting blood flow and increasing the risk of heart attacks. With early preventive measures, actions can be taken to significantly reduce the risk of CHD early on.

Our project aims to predict the CHD death rate for a particular region based on local environmental parameters. From this prediction, we can advise people on CHD death risk based solely on regional data, not personal information, to promote early preventative actions.

Dataset

Our dataset is from the CDC Division for Heart Disease and Stroke Prevention Interactive Atlas, and each data point contains the following heading features; each feature includes several sub-features, such as hospital #, poverty %, etc.

- County and State name

- Coronary Heart Disease death rate per 100,000 for all ages, all races/ethnicities, both genders, for 2016-2018
- Risk factors:
  - Diabetes %
  - Obesity %
  - Physical inactivity %
- Social environment:
  - Education less than high school %
  - Education less than college %
  - Female headed household %
  - Food stamp / SNAP recipients %
  - Median home value $
  - Median household income $
  - GINI coefficient (income inequality)
  - Poverty %
  - Unemployment rate %
- Demographics:
  - American Indian / Alaska Native %
  - Asian / Native Hawaiian / other Pacific islander %
  - Black %
  - White %
  - Hispanic / Latino %
  - Age 65 and older %
  - Total population
- Physical environment:
  - Air quality (annual PM2.5 level)
  - Park access %
  - Severe housing problems %
- Urban-rural status:
  - NCHS urban-rural status

Methods

Our team will predict the number of Coronary Heart Disease deaths by region based on the input parameters specified above, and will place these regions into four “risk” bins corresponding to consistent CHD death rate intervals.

Due to download constraints from the website, we exported our dataset in intervals of different features. To combine these 30+ features into one consistent table, we will concatenate the different features by referencing the county ID (or county name and state name) provided.

We plan to use unsupervised clustering algorithms such as K-Means and and GMM with different numbers of clusters to observe relationships between data points. We will compare the results of K-Means to GMM because we are not sure what the shape of our clusters will be. The goal of these clustering algorithms would be to maximize the purity of each cluster such that the points within the same cluster correspond to the same risk-level bin. We can evaluate the results by measuring the average purity across all the different clusters.

We plan to use supervised algorithms such as Random Forests to identify important features. By using RF, we can find the most predictive features in our dataset, which will inform our linear regression design. We can evaluate the results of RF by looking at an accuracy score first, and then using a confusion matrix to express whether features claimed as predictive of high risk did in fact correspond to high risk. We will use Linear Regression to help predict the specific number CHD death rates by region based on the specified input. We might modify the Linear Regression algorithm slightly by changing the evaluation metric to containerize results into our four risk-levels. We can then measure the accuracy of these bins based on the modified regression. Lastly, we believe supervised Hidden Markov Models will be helpful in looking at the evolution of events from 2005 to 2018 and predicting metrics a few months to a year out in 2019. To evaluate our Hidden Markov Model, we will compute the likelihood of different sequences of events using our test data, and we will use the forward algorithm to compute the mean error by tracking the state of the HMM and the expected observation at some future time. For all of these predictions, we will be evaluating them based on partitioning our dataset into training (80%) and testing (20%) subsets to ensure we have ground-truth data to base our evaluations off of. We will run these algorithms multiple times with randomized partitioned data points to prevent overfitting.

Results

With our supervised and unsupervised algorithms, by separating the CHD death rate into 4 discrete bins, we hope to achieve 90% accuracy in CHD death classification based on the environmental traits listed above. Using the confusion matrix, we also hope to discern the environmental traits most closely associated with CHD death rates from features that don't have much correlation at all. In regards to our evaluation metrics, we expect to get a purity score of >0.9 with our K-Means algorithm and an accuracy score of >0.9 with the Linear Regression and RF algorithm, as we are trying to maximize correct classifications of our supervised algorithms.

Discussion

Our model could also help understand which specific conditions of a region would need to be improved and to what extent in order to decrease the risk of CHD; with these results, policy changes and recommendations can be made. For example, determining if the number of hospitals or the insurance policy or cost has a direct effect on CHD risk. A potential next step would be to extend our algorithm to predicting not just heart disease risk but also the likelihood of other illnesses and analyze the effect of different combinations of features on the risk (for example, breast cancer, cervical cancer, etc.).

References

Dalen, James et. al. "The Epidemic of the 20th Century: Coronary Heart Disease." The American Journal of Medicine, 2014. Retrieved 25 September 2020 from https://www.amjmed.com/article/S0002-9343(14)00354-4/pdf

"Coronary Artery Disease: Prevention, Treatment and Research." Johns Hopkins Medicine. Retrieved 25 September 2020 from https://www.hopkinsmedicine.org/health/conditions-and-diseases/coronary-artery-disease-prevention-treatment-and-research

"Heart Disease Facts." Centers for Disease Control and Prevention, 2020. Retrieved 25 September 2020 from https://www.cdc.gov/heartdisease/facts.htm