Maha Mapara
Revised Problem Description

Recap:

My problem statement was about predicting when a public health intervention is necessary without having to wait for a certain disease (caused by modifiable risk factors) to rise in the population. A reframing of this is: Can we predict that a disease will be an issue in a community before it becomes a burden?

By public health intervention I mean behavioral interventions like health promotions, programming done by government, nonprofits and even hospitals to raise awareness before the diseases become major issues in a community/area.

Vocabulary:

1. Prevalence: reflects the number of existing cases of a disease.
2. Incidence: reflects the number of new cases of disease and can be reported as a risk or as an incidence rate.

Updates:

1. Upon further thought, I want to frame my research question as predicting whether a community is at risk of a particular disease becoming prevalent due to human behavior like unhealthy diet, sedentary lifestyle, tobacco consumption etc. If it is predicted to be at risk, then there should be public health interventions.
2. I have narrowed down the diseases I will focus on to:
    a. Heart disease
    (not including diseases that are genetic or someone was born with like hypertrophic cardiomyopathy and congenital heart defects)
    b. Cancer
    (only ones linked to modifiable risk factors like liver cancer, pancreatic cancer)
    c. Diabetes- type 2
    (Type 1 is not caused by lifestyle choices)

The reason behind choosing these diseases is that they are some of the leading causes of death in the US and are linked to modifiable risk factors (mentioned in point 1).

3. Looking at data from 50 states is too big of a task, so I have chosen to narrow my work to the following states:
    a. States with highest cancer incidence (as compared to national average):

        i. Kentucky, Delaware, Pennsylvania, New Hampshire, New York, New Jersey, Iowa, Maine, Louisiana, Virginia, Arkansas, Minnesota, Nebraska

    b. States with highest incidence of heart disease (as compared to national average):
        i. Oklahoma, Arkansas, Louisiana, Mississippi, Alabama, West Virginia, Kentucky, Tennessee, Missouri, Ohio, Nevada, Michigan
    c. States with highest incidence of type 2 diabetes (as compared to national average):
        i. West Virginia, Mississippi, Tennessee, Alabama, Arkansas, Kentucky, Oklahoma, South Carolina, Ohio, Georgia, New Mexico, Louisiana, Florida, Pennsylvania, Delaware, Missouri
        (states underlined are the ones that are present in the other high incidences categories)

Data:

1. Google trends data on searches related to smoking, alcohol consumption and diseases.
2. Data on tobacco purchase.
3. Nutrition, physical activity and obesity data.
4. Alcohol consumption- worldwide
5. Alcohol consumption per capita in the US from 1850 to 2018
6. Tobacco products in the US
7. Stats of the States - Heart Disease Mortality
8. Stats of the States - Cancer Mortality

Modeling approaches:

1. If I am predicting whether a community is at risk or not, or different classes of risk (very high risk, high risk, low risk etc), I can use logistic or multinomial regression.
2. It would be interesting to use k-means clustering to see how age groups, ethnicities, risk factors might get clustered. I could use the cluster(s) as a variable in the classification task.
3. Support vector machine to find optimal boundary between possible outputs.
4. Decision trees are a good way of seeing how something is being predicted and is used for classification.

Questions I'm thinking about and may be you are too :

1. I can't use google trends data to predict communities at risk of disease incidence as google started collecting and making search data available from 2011. However, I still think it would be interesting to incorporate it somehow. Perhaps I can use it as a way to confirm the areas I predict to be at risk?

2. Are there better or more interesting modelling approaches I could be using?
3. How does this all come together?
    a. I'll use state/county level data on disease incidence, consumption data on tobacco and alcohol, and race, age, gender, socio-economic status data for a state/county.
    b. For each state listed right now, it took quite some time for them to be at the incidence level they are. Delving into past data and consumption habits in these states should highlight the kind of behavior that leads to the current incidence level.
    c. How will all this different kind of data come together in a model? I don't know, I'll have to do more research on how this will work.
4. What do I mean by predicting if a community is at risk?
    a. A community can't be a state or even a county as that's too large of a geographical area. For a big city like New York, it would be divided into its districts. For smaller cities and towns, I can count each one of them as a community.
5. Something interesting to look at will be fast food restaurants opening up in these states, the particular locations and seeing if there is a correlation between number of fast food chains in an area and how long they have been there to disease incidence.
6. Will I be able to generalize my model if I'm using past data from only a few states? All states differ in the age, race, gender, socio-economic status of their residents, so this is something to think about.
7. Is this project too large? Should I reduce its scope?

Bibliography

FastStats - Leading Causes of Death

Do You Live in a Cancer "Hot Zone"? These Are The 20 States With the Highest Cancer Rates

Stats of the States - Cancer Mortality

USCS Data Visualizations - CDC

State-wise cancer profile

Cancer is a Preventable Disease that Requires Major Lifestyle Changes

Lifestyle Behaviors Contributing to the Burden of Cancer - Fulfilling the Potential of Cancer Prevention and Early Detection

Heart Disease Statistics and Maps | cdc.gov

American Health Ranking

Stats of the States - Heart Disease Mortality

Annual Report 2019- US Heart Disease Rankings

Alcohol Facts and Statistics- USA

Diabetes Rankings in the United States