**CSC440 Data Mining**

**Final Project Report**

**The Relationship Between Health and Health Factors**

by Zhuoyou Wang &Yang Wang

Abstract

In this report we will present our project result regarding the relationship between health factors and health outcomes.

**Introduction**

**Motivation and Background**

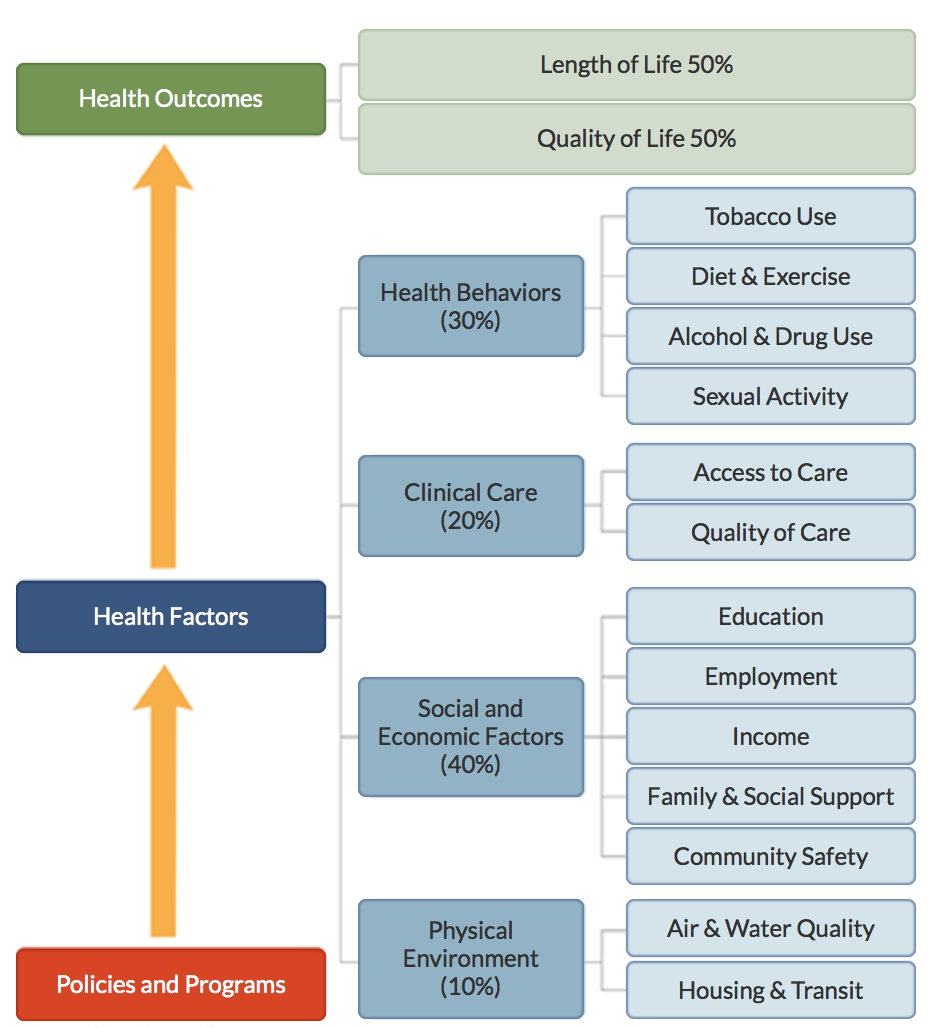
At all the time, health is the No.1 issue in individuals’ life since it is the fundamental of all activities. Without health, people are not able to perform well in casual life, studies or work. However, even with dramatic advances in medical field, people still suffer from various diseases in recent years. Some diseases (obesity) may damage people’s health and negatively influence people’s life for a small amount of time, while the other(HIV) may chronically encroach upon people’s health and even take away people’s lives. Rather than focusing on the treatment and medicine for diseases, we would like to study the root of diseases, which is the unhealthy factors that may lead to corresponding diseases. We believe that diseases are correlated with various unhealthy factors in life. Thus, in order to avoid or decrease the probability of getting unhealthy, it might be effective if we avoid or at least decrease the frequency of those unhealthy habits.

For this reason, in this project, we would study the relationship between people’s health and health factors in the United States. Most studies or data show that there is a strong correlation between US people’s health and health factors without telling the specific correlations between certain diseases and health factors. Therefore, we plan to look close into the relationship and study some specific diseases, such as obesity and sexually transmitted infections, with specific health factors, such as smoking and education.

Data mining tools such as correlation and classification analysis allow us to achieve reliable interesting findings from data related with health and health factors. Thus, in this project, we use correlations, linear regression, decision tree and ROC curve to analyze corresponding data, and finally conclude our findings and learning and give possible application based on our findings.

**Data description**

The data set we used is from County Health Rankings & Roadmaps, this website is supported by University of Wisconsin Population health institute and Robert Wood Johnson Foundation. The data measures the health of all most every county in the nation. The highest level data contains only two variables: Overall health factor and Overall health outcome. Overall health factor is made up by two parts, length of life and quality of life. Both of them weight 50% in the overall health outcome. Health factor is composed by Health behavior 30%, clinical care 20%, social and economic factor 40% and physical environment 10%.



|  |  |
| --- | --- |
| Health behaviors (30%) |  |
| Focus Area | Smoking (10%) |
| Focus Area | Diet and exercise (10%) |
| Focus Area | Alcohol and drug use (5%) |
| Focus Area | Sexual activity (5%) |
| Clinical care (20%) |  |
| Focus Area | Access to care (10%) |
| Social and economic factors (40%) |  |
| Focus Area | Education (10%) |
| Focus Area | Employment (10%) |
| Focus Area | Income (10%) |
| Focus Area | Family and social support (5%)  Income inequality (2.5%) |
| Focus Area | Family and social support (5%) |
| Focus Area | Community safety (5%) |
| Physical environment (10%) |  |
| Focus Area | Air and water quality (5%) |
| Focus Area | Housing and transit (5%) |

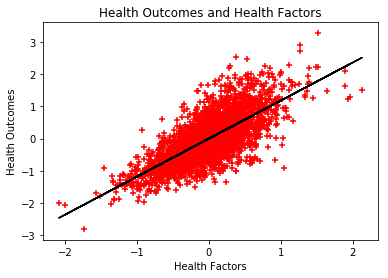
**Methodology**

**Data preprocessing**

In the original data set it contains rankings for each county which we don’t need for the purpose of this project. Moreover, for some of the data the values have been reversed due to their positive framed measure. All the values in the data are standardized using Z-score. Therefore, I don’t have to go through all the trouble to normalize values. I get off observations that has null values, and get rid of columns that show the rankings for corresponding counties.

**Correlation analysis**

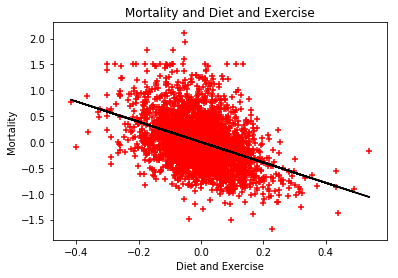
There is a strong correlation between overall health outcomes and overall health factors. Their correlation coefficient is 0.7315



I look up correlation between each factors and outcomes find some interesting correlations.

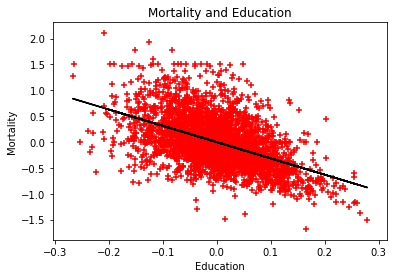
Some of the are intuitive and some are not that intuitive.

**Intuitive:**



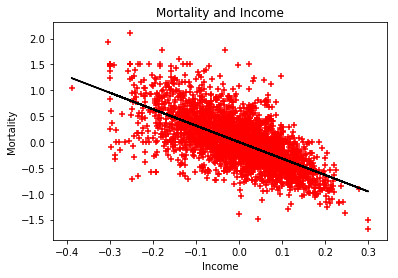
Mortality and Diet and exercise (-0.409031):

Eating health and exercise regularly “can reduce the chance of getting heart disease, diabetes, depression and several types of cancer particularly breast and colon cancers” according to The Secret to Better Health - Exercise. Therefore, good diet and regular exercise lead to less disease, which eventually imply lower mortality rate.



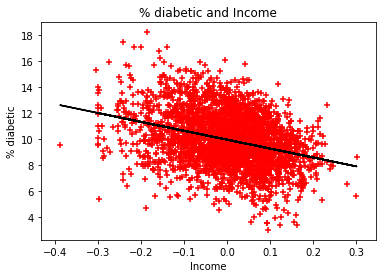
Mortality and Education (-0.509406):

Better education usually indicates the person has more knowledge, and they are more likely to have healthy behaviors. Moreover, higher level of education usually means a better job and having a better resource for good health. Therefore, they tend to live longer in the long run.



Mortality and Income (-0.655259):

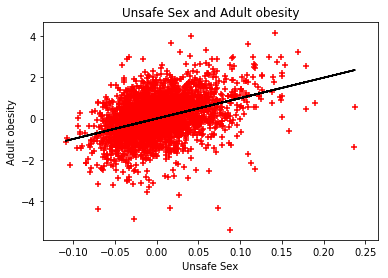
Higher income leads to better living standard and better medical attention in general, hence a better chance to live longer.



Diabetic and Income (-0.323939):

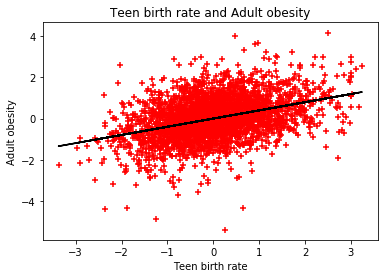
Low income is usually related with stress caused by uncertainty of food, housing and medical care etc. According to Andrew Curry “stress amps up craving for energy dense foods; in people, and also increase the level of cortisol level, which has a direct link between cortisol and diabetes.” Therefore, low income is more likely to cause diabetes.

**Not intuitive:**



Unsafe sex and Adult obesity (0.394659):

The reason behind this could be that being obese affects people’s sex drive, desire and performance; therefore, obese people is look for extra stimulations for their sex life. According to Colette Bounchez, “Over weight will affect the hormones lead to active libido”. Moreover, because of obese people’s body shape it is harder for them to have access to sex, so that obesity people are willing to do some extra to have that access to sex. Thus, there is a higher percentage of obesity people have unsafe sex than normal weight group.



Teen birth rate and Adult obesity (0.389539):

When a teenage girl is pregnant she needs a lot of nutrition to keep her body and baby growing. According to Debolina Raja, this can cause eating disorder such as bulimia. Debolina also mentions that teen moms are more likely to be depressed during pregnancy. Depression could trigger stress eating. Taking these two factors into account, we might be able to explain why Teen birth rate has a positive correlation with adult obesity.

**Classification Analysis**

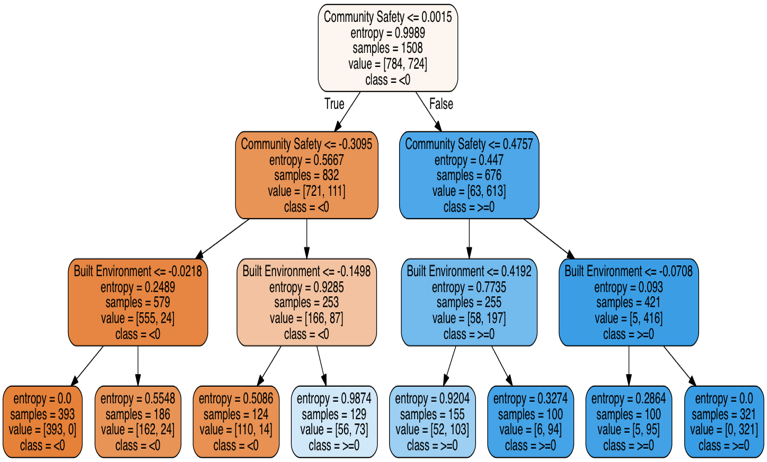
For classification analysis, we decide to use decision tree. We choose 13 health factors mentioned in ‘Data Description’ (excessive drinking, education, employment and etc.) as attributes and divide health outcomes into 2 classes (‘<0’ and ‘>=0’) according to z-scores based on county levels. When the z-scores of people’s health outcomes are smaller than 0, we claim that people’s health outcomes at that county are below the average health outcomes, and thus defined as unhealthy outcomes. Similarly, we define health outcomes that are above or equal to 0 as healthy outcomes.

In the part of attribute selection measure, we choose information gain and Gini index to build 2 decision trees separately in order to get the most appropriate decision tree with the highest accuracy.

In the part of decision tree construction, we divide the datasets into a training set and a testing set. We choose the first 2000 data as a training set and the rest 1017 data as a testing set. After setting min\_samples\_leaf=100, min\_samples\_split=200, we get the following decision trees:

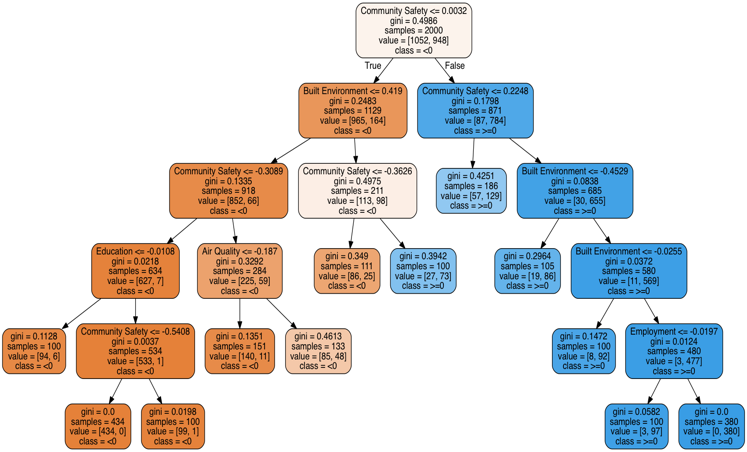
Decision Tree with Information Gain:

The mean accuracy of the decision tree with information gain (entropy) is approximately 0.89282202556538837



Decision Tree with Gini Index:

The mean accuracy of the decision tree with Gini index is approximately 0.89282202556538837



With the attributes and classes, we find that the accuracy of decision trees using information gain and Gini index are identical. Since we have many unique values for each attribute and a binary class, it should not be surprising that we get two similar decision trees. Thus, either the decision tree with entropy or that with Gini index is a good choice for us based on the high accuracy of 89.28%. In another way, we can evaluate the accuracy of the decision tree by measuring the area under ROC curve, which is the following diagram:



The accuracy measured by ROC curve is 88.13%, which is similar to our mean accuracy 89.28%. Also, with the decision tree, we are able to predict health outcomes with given attributes. For example, if we randomly generate a list of z-scores for 13 attributes, which is [-0.05,0.78,0.22,0.34,-0.55,-0.82,-0.03,0.64,0.78,0.62,-0.79,-0.66,0.03], then we are able to predict a health outcome (['<0']) with the decision tree. Most importantly, from the two decision trees with different attribute measures, we find that the first 2 attributes they choose to build trees are the same, which are community safety and built environment. Among the first two attributes, community safety is the most significant one. This indicates that we can classify whether people are healthy or not by first analyzing community safety through Gini index/entropy. It also means that community safety significantly contributes to health outcomes. In real life, it also makes sense since people live in a safe community are more likely to receive better living standards and less likely to suffer from contagious diseases such as sexually transmitted infections.

**Application**

Since the data we analyze in this project is based on county level, we can use the same methodology (correlation and classification) to mine the data on the fundamental of states/counties. In this way, we are able to find specific health factors that correlate with specific diseases in specific states. Moreover, combining the correlations of specific factors and diseases with morbidity of such kind of diseases in states, state governments may carry out policies or programs more effectively by focusing on local diseases and factors. For example, if there exists a high correlation between unsafe sex and adult obesity in California, then the local government may implement policies about decreasing prices/taxes of condoms, or programs about distribution of free condoms and booklets of unsafe sex’s harms to local residents.

**Conclusion**

With the help of data mining tools, we have found that overall health factors are highly correlated with health outcomes with correlation = 0.7315. If we take a close look into individual health factors and diseases, we are able to see separate correlations explicitly. Some interesting findings are adult obesity are correlated with unsafe sex in correlation = 0.3947 while diabetes is correlated with income in correlation = 0.3239. Furthermore, a decision tree helps us classify and predict a health outcome given 13 health factors/attributes with the accuracy of 88.13% by ROC curve and 89.28% by mean. Based on the analysis above, we can conclude that health outcomes are highly correlated with health factors (i.e. smoking, income, education, etc.), and there exist individual correlations between separate diseases and health factors (i.e. obesity vs. unsafe sex). In addition, we are able to predict health outcomes given correct input/attributes using a decision tree.

**References**

Andrew Curry 2015 *The Sheer Stress of an Environment Contributes to Obesity and Diabetes* <http://nautil.us/issue/31/stress/why-living-in-a-poor-neighborhood-can-make-you-fat>

Colette Bouchez 2005-2017 *Better Sex: What’s Weight Got to Do With it?* <https://www.webmd.com/sex-relationships/features/sex-and-weight#1>

Data set source: <https://github.com/dataiap/dataiap/tree/master/datasets/county_health_rankings>

Debolina Raja 2017 9 *Health Rishs And Realities of Teenage Pregnancy* <http://www.momjunction.com/articles/health-risks-of-teenage-pregnancy_00377831/#gref>

<http://www.countyhealthrankings.org/explore-health-rankings/use-data/exploring-data>

Virginia Commonwealth University 2015 *Why Education Matters to Health: Exploring the Causes* <https://societyhealth.vcu.edu/work/the-projects/why-education-matters-to-health-exploring-the-causes.html>