Technical Report

Eway Cai, Jeanny Zhang

March 16, 2022

For the final project, our team built a shiny app to investigate 7 socioeconomic and cultural factors we hypothesized to have significant impacts on alcohol consumption of individuals in different countries. The factors included gross domestic product (GDP) per capita, homicide rate, happiness index, working hours, corruption and more. We believe such factors are either directly or indirectly correlated to the amount of alcohol consumed, whether for leisure or work purposes.

The first tab introduces the users to the background of drinking cultures, the main goal of our website and the data sets used. Various HTML operations are used for the formatting of text to boost user experience.

We studied the topic of interest from two approaches, shown on the second and third tab of the shiny application. The second tab provides an overview of each of the data sets used for the study. The first subtab displays a line plot showing the trends of the user's factor of interest through time (for example, the number of working hours of a country over time). The user can select one or multiple countries of interest, and corresponding trends for that country will be displayed. The second subtab displays a heatmap of alcohol consumption distribution. We use this to let the users have a deeper understanding of what the distribution is like overall, which lays the foundation for exploring the potential relevant factors on the next tabs. Methods used in this tab included basic shiny tools to add selection buttons, using 'Plotly' package to visualize data and add interactivity with the user. Reactive functions are also incorporated to enhance our ability in reacting to user input and making relevant changes to our data sets.

The third tab panel displays our attempt at observing trends between various socioeconomic factors and alcohol consumption levels, as well as utilizing methods in machine learning to perform supervised and unsupervised learning. The first sub-tab shows scatterplots of a factor of interest as the explanatory variable and alcohol consumption as the response variable. The user has the option to select between 7 different factors, while alcohol consumption is always displayed on the y-axis. In the second sub-tab, linear regression modeling is used to explore possible correlations between the variables. The third sub-tab uses K-means clustering to categorize each data point. An elbow plot is displayed to inform users about the optimal number of k. Users can also change k by will, but the default number for k is three as it is shown to be the optimal k based on the elbow method.

Methods used in the fourth tab panel are all machine learning related such as decision trees and random forest. Users have the freedom to choose more than one variable, use decision trees algorithm to find out how these variables affect alcohol consumption level, and use random forest algorithm to find out the importance of each variable. The alcohol consumption data is pre-processed and clustered into three categories: "Low Alcohol Consumption", "Median Alcohol Consumption", and "High Alcohol Consumption". K-means clustering was conducted and a cluster number of three is determined to be the optimal. The default selected variables are "Corruption Perception Index", "GDP Per Capita", "Life Satisfaction", "Literacy Rate", and "Working Hours", because they are the variables that exhibit statistically significant correlation with alcohol consumption based on simple linear regression (p-value < 0.001). We are aware that our class did not cover anything regarding having three categories in decision trees or random forest. We do not know any method to check the accuracy of our models. However, we believe that it is beneficial to build the models with the optimal clustering in order to yield more accurate results. We would love to learn more about the advanced model checking methods if we have more time.