Project 1 Week 1

Kristie Kooken

DSC 680-T301

Due: March 17, 2024

The topic that I will investigate for my first project will be exploring the relationship between different health indicators and diabetes and hypertension to determine how these factors can predict stroke. The name of this project will be Diabetes, Hypertension and Stroke Prediction.

Business Problem:

Diabetes is a well-established risk factor for stroke (Chen, et al., 2017). Likewise, high blood pressure (hypertension) can often have no symptoms and if untreated, may lead to conditions such as heart disease and stroke (www.mayoclinic.org). The annual estimated costs of loss of productivity from these two chronic diseases is staggering, about 600 billion US dollars (www.cdc.gov). This investigation will explore how different health behaviors and physical characteristics (e.g., age and sex) as well as diabetes and hypertension diagnoses can predict who can be at risk for stroke.

Dataset:

Using a dataset from kaggle.com that is based on the 70,692 survey responses from the Behavioral Risk Factor Surveillance System (BRFSS) 2015 (www.kaggle.com), I will investigate if participants with diabetes and/or hypertension have a higher degree of stroke outcomes based on the collected data. I am looking to explore what health indicators from regular exercise to number of alcoholic beverages consumed in a week to see if these factors are higher or lower in those with diabetes and / or hypertension and how this may influence incidence of stroke. The data from this survey was collected in 2015. This survey collects health behavioral risk factor data and is an established data collection tool since 1984 (www.cdc.com).

Methods:

This data will be analyzed with Python using Jupyter Notebook. Data cleaning will be performed on this data to remove any duplicate rows and replace or remove any missing data. Likewise, creation of meaningful categorical variables will be done as applicable. Frequencies and descriptive statistics will be explored to understand data distributions. Additional exploratory data analysis will be conducted to visualize the data and determine if there are underlying relationships when the different factors are viewed by those with and without diabetes, hypertension, and stroke. Logistic regression modeling will be conducted with stroke as the dependent variable and other factors including diabetes and hypertension as independent variables.

Ethical considerations:

Ethical considerations include knowing that survey data may not represent all people who suffer from these chronic diseases but only those who are willing to participate in this research. Likewise, results of this survey may not be representative of all populations who have these chronic diseases like diabetes and hypertension, there are limited demographics collected so it will be a challenge to know how balanced the sample is. Because of this, it will be important to interpret results with these limitations in mind. Likewise, this data is limited to survey data which is self-reported from the participants and could differ from what they do in everyday life. This will be another important consideration when interpreting results.

Challenges/Issues:

I could face different challenges in combining this data (there are three datasets, and I may try to combine them) and in developing a robust model to predict stroke. This type of model development work can take months or even years, and I will have more limited time though I am looking forward to the challenge. Other challenges could include not being able to create a meaningful model or not being able to replicate results published by the BRFSS.

References:

The incidence of each chronic disease could be compared to the overall population incidence to see if this sample was representative of those populations. Likewise, the results of this analysis will be validated by looking at CDC statistics to determine if data from this survey are the same as what is presented based on larger dataset or data collected over multiple years. It would be interesting to compare model results to with other models for diabetes and stroke or hypertension and stroke as well.

References:

Behavioral Risk Factor Surveillance System (accessed 2024, March 17). The Center for Disease Control and Prevention. https://www.cdc.gov/brfss/about/brfss_faq.htm.

Chen R, Ovbiagele B, Feng W. Diabetes and Stroke: Epidemiology, Pathophysiology, Pharmaceuticals and Outcomes. Am J Med Sci. 2016 Apr;351(4):380-6. doi: 10.1016/j.amjms.2016.01.011. PMID: 27079344; PMCID: PMC5298897.

Diabetes, Hypertension and Stroke Prediction (accessed 2024, March 17). Kaggle. https://www.kaggle.com/datasets/prosperchuks/health-dataset/data.

Health Topics – Diabetes (accessed 2024, March 17). The Center for Disease Control and Prevention. https://www.cdc.gov/policy/polaris/healthtopics/diabetes/index.html#:~:text=The%20total%20estimated%20cost%20of,90%20billion%20in%20reduced%20productivity.

Health Topics – High Blood Pressure (accessed 2024, March 17). The Center for Disease Control and Prevention. https://www.cdc.gov/policy/polaris/healthtopics/highbloodpressure/index.html#:~:text=Economic%20Burden,to%20%24198%20billion%20each%20year.

High blood pressure (hypertension) (accessed 2024, March 17). The Mayo Clinic. https://www.mayoclinic.org/diseases-conditions/high-blood-pressure/symptoms-causes/syc-

20373410?utm_source=Google&utm_medium=abstract&utm_content=Hypertension&utm_campaign=Knowledge-panel