Hello World

data-to-paper

June 27, 2023

1 Methods

1.1 Data Source

The data for this study was obtained from the CDC's Behavioral Risk Factor Surveillance System (BRFSS), specifically from the year 2015 survey. The BRFSS is an annual health-related telephone survey that collects information on health-related risk behaviors, chronic health conditions, and the use of preventative services from over 400,000 Americans. The dataset used for this study consists of 253,680 responses with 22 features, including diabetes status, fruit and vegetable consumption, physical activity level, and demographic variables. The dataset was provided as a comma-separated values (CSV) file.

1.2 Data Preprocessing

The pre-processing of the data was performed using Python programming language. First, missing values were removed from the original dataset, resulting in a clean dataset of 253,680 responses. This step ensures that the subsequent analysis is conducted on complete data. Next, a new variable called "FruitVeg" was created by combining the "Fruits" and "Veggies" variables using a logical AND operation. This new variable represents whether an individual consumes at least one fruit and one vegetable each day. These pre-processing steps were performed using the pandas library in Python.

1.3 Data Analysis

To examine the association between fruit and vegetable consumption, physical activity, and the risk of diabetes among adults, logistic regression analysis was conducted using the statsmodels library in Python. In the first analysis step, a logistic regression model was fitted with the "Diabetes_binary" variable as the dependent variable and "FruitVeg," "Age," "Sex," "BMI," "Education," and "Income" as independent variables. This analysis aimed to determine the association between fruit and vegetable consumption and the risk of diabetes, while controlling for demographic and health-related factors.

In the second analysis step, an interaction term between fruit and vegetable consumption ("FruitVeg") and physical activity level ("PhysActivity") was introduced in the logistic regression model. The model included the main effects of "FruitVeg" and "PhysActivity," as well as the interaction term "FruitVeg_PhysActivity." This analysis aimed to investigate whether the association between fruit and vegetable consumption and diabetes risk is modified by physical activity level.

The results of the logistic regression analyses, including odds ratios and corresponding p-values, were obtained from the fitted models. Additionally, descriptive statistics for the dataset were calculated using the pandas library. The results were written to a text file named "results.txt" for further examination and reporting.

These analysis steps provide insights into the association between fruit and vegetable consumption, physical activity, and the risk of diabetes among adults, while controlling for potential confounding factors.