Hello World

data-to-paper

July 2, 2023

1 Methods

1.1 Data Source

The data used in this study was obtained from the "diabetes_binary_health_indicators_BRFSS2015.c
file. This dataset consists of various health indicators collected through the 2015 Behavioral Risk Factor Surveillance System (BRFSS) survey. The data includes information on diabetes status, high blood pressure, high cholesterol, cholesterol check, body mass index (BMI), smoking behavior, history of stroke and heart disease or heart attack, physical activity, fruit and vegetable consumption, heavy alcohol consumption, healthcare access, general health, mental health, physical health, difficulty with walking, sex, age, education, and income. Each row represents an individual record, and each column represents a specific health indicator or demographic characteristic.

1.2 Data Preprocessing

Prior to conducting the analysis, several preprocessing steps were performed on the dataset using Python. First, any missing values in the dataset were filled in using the mean value of the respective column. This approach ensured that all records contained complete information for further analysis.

Next, the BMI values were normalized using the MinMaxScaler from the scikit-learn library. This transformation rescaled the BMI values to a range between 0 and 1, allowing for a standardized comparison.

1.3 Data Analysis

To investigate the association between BMI and the prevalence of diabetes, a chi-squared test of independence was performed. First, the data was divided into groups based on BMI quartiles. This was achieved by applying the qcut function from the pandas library, which evenly partitioned the BMI values into four groups, ensuring an approximately equal number of observations in each group.

Next, a contingency table was created by cross-tabulating the diabetes status (binary) and the BMI groupings. This contingency table represented the observed frequencies of individuals with and without diabetes within each BMI group.

The chi-squared test of independence was then applied to the contingency table. This test determined whether there was a statistically significant association between BMI and diabetes. The test calculated the chi-squared statistic and the corresponding p-value. The chi-squared statistic measures the overall deviation from independence, while the p-value indicates the probability of obtaining such an extreme result, given the assumption of independence.

Finally, the results of the analysis were recorded in a results.txt file. This file included general information about the dataset, such as the total number of observations, the mean BMI value, and the percentage of subjects with diabetes. Additionally, two tables were created to summarize the results. Table 1 presented descriptive statistics of BMI for individuals with and without diabetes, while Table 2 reported the chi-squared statistic and p-value resulting from the test of independence.

The code implementation provided a comprehensive analysis of the association between BMI and the prevalence of diabetes.

s