Predicting Health Outcomes Using Nutritional Intake, Socioeconomic Factors, and Sleep Data

Huy Duc Pham, Hanush Kumar, Neha Anil Chede

Abstract

This study aims to develop a predictive model to health outcomes using NHANES data, applying machine learning and association rule mining to identify patterns and classify individuals into risk groups. Performance will be evaluated using ROC curves and F1 scores, with the goal of providing actionable insights for public health interventions and identifying at-risk populations for disease prevention.

Keywords: Health outcome prediction, nutrition, chronic diseases, socioeconomic factors, sleep patterns, association rules, diabetes, depression, insomnia, classification, clustering

A. Introduction

Understanding the relationship between factors such as demographic factors, dietary habits, and biomarkers is critical for the development of preventative medicine and personalized health care strategies. As large-scale health datasets become more accessible, there is an unique opportunity to use this data to better understand how these factors contribute to the prevalence of chronic diseases and health risks. In this study, we aim to investigate the association between demographic variables, dietary habits, and biomarkers using data provided by the Centers for Disease Control and Prevention[1]. Through statistical analysis, association rule mining and predictive modeling, we aim to identify patterns that can help classify the likelihood of individuals contracting specific diseases like essential hypertension, type 2 diabetes mellitus, asthma, major depressive disorder, and insomnia. By leveraging association rule mining, we will uncover frequent itemsets and strong associations between lifestyle factors and the development of chronic health conditions. By examining high-dimensional data structures and identifying these meaningful patterns, our goal is to provide insights into disease risk factors that could inform public health interventions and strategies for disease prevention.

B. Methods

The data comprises multiple files, including demographics (age, gender, race, age, education, household structure), dietary (meal intake, dietary patterns, sodium, B12), examination (BMI, blood pressure, median liver stiffness), laboratory (cholesterol, ferritin, glucose), medications (prescription drug use, primary & secondary diagnosis), and questionnaire (alcohol use, diabetes) datasets, each containing unique predictors essential to our analysis.

To examine the effects of nutritional intake, socio-economic factors and sleep data on health outcomes, we will develop classification models to categorize individuals into risk groups: low, high, and very-high risk. First, we will perform data preprocessing, including encoding categorical variables (e.g., gender) into numerical formats suitable for machine learning models and normalizing numerical variables (e.g., BMI, cholesterol levels) to ensure consistency.

Missing values will be handled through imputation methods, such as mean or predictive imputation, and outliers will be detected and addressed to prevent skewed results.

We will then apply Principal Component Analysis (PCA) to reduce the dimensionality of the dataset, removing noise and improving data quality while retaining the most significant features. To uncover patterns and groupings, we will perform K-means clustering to identify natural groupings within the data and explore how various factors, such as demographics and nutrient intake, relate to each other and to health outcomes.

We will employ multi-level classification algorithms, including Decision Trees, Random Forests and XGBoost, to identify patterns and associations between the predictor variables (nutritional intake, sleep and socioeconomic factors) and target outcomes, i.e. the risk level of chronic diseases. The model's performance will be evaluated using metrics such as the ROC curve, precision, recall, and F1 score.

We will also use the Apriori algorithm for multi-level association analysis to explore the relationships between socioeconomic status, nutrient intake, and health outcomes, aiming to uncover key associations between lifestyle factors and chronic conditions.

References

1. Centers for Disease Control and Prevention (CDC). National Health and Nutrition Examination Survey (NHANES), 2021-2023; https://wwwn.cdc.gov/nchs/nhanes/continuousnhanes/default.aspx?Cycle=2021-2023
2. Ambika Satija, Edward Yu, Walter C Willett, Frank B Hu Understanding Nutritional Epidemiology and Its Role in Policy 2015; 10.3945/an.114.007492
3. Sareen S Gropper The Role of Nutrition in Chronic Disease 2023; 10.3390/nu15030664
4. Kimokoti R.W., Millen B.E. Nutrition for the Prevention of Chronic Diseases. Med. Clin. N. Am. 2016;100:1185–1198. doi: 10.1016/j.mcna.2016.06.003
5. Colten HR, Altevogt BM, editors. Sleep Disorders and Sleep Deprivation: An Unmet Public Health Problem. Washington (DC): National Academies Press (US); 2006
6. Cátia Reis , Sara Dias , Ana Maria Rodrigues , Rute Dinis Sousa , Maria João Gregório , Jaime Branco , Helena Canhão , Teresa Paiva Sleep duration, lifestyles and chronic diseases: a cross-sectional population-based study 2018; 10.5935/1984-0063.20180036