Open-ended Exploration Report

Haiming Li, Shuxian Chen, Witty Wen

2023-11-06

Introduction

In this report, we delve into two fundamental questions using the NHANES data. First, we explore the potential relationships between various features - such as demographic, dietary, and lifestyle factors - and cholesterol levels. This analysis aims to uncover patterns and correlations that could aid in better understanding the dynamics of cholesterol levels across the U.S. population. Second, we investigate how the importance of these features varies across different age groups. Given that the risk factors and health profiles can significantly differ with age, this analysis could provide insights into age-specific strategies for managing cholesterol levels effectively.

In addressing the two pivotal questions outlined, our study adopts a robust analytical approach by employing Support Vector Machine (SVM) and XGBoost (XGB) models. These advanced machine learning techniques are selected for their efficacy in handling complex, multidimensional data, allowing us to decipher intricate patterns and relationships within the NHANES dataset.

A critical aspect of our analysis involves the use of SHAP (SHapley Additive exPlanations) values. This innovative method provides a detailed and interpretable understanding of feature importance within our models. By leveraging SHAP values, we can gain insights into how different variables, such as demographics, dietary habits, and lifestyle choices, influence cholesterol levels. This approach not only aids in identifying key drivers behind cholesterol variations but also allows us to compare the relative importance of these features across the SVM and XGB models, ensuring a comprehensive and nuanced understanding.

Further, to capture the nuances associated with age-related variations in cholesterol levels, we partition the dataset into three distinct age-based subgroups. This stratification enables a more granular analysis, allowing us to assess and compare feature importance across different age categories. Such an age-specific approach is vital, considering the varying risk factors and health profiles across age groups. It enables us to tailor our findings and recommendations to specific age demographics, enhancing the relevance and applicability of our study in clinical and public health settings.

Data

We utilize the 2017-March 2020 Pre-Pandemic Demographics data. The dataset contains information on total cholesterol for all participants examined who were 6 years or older before the pandemic. Since the suspended field operations of NHANES programs was stopped in March 2020 due to the pandemic, the part of the dataset collected between 2019 and March 2020 was combined with the data from the 2017-2018 cycle to form a full representative sample of that period.

The focus of our research is the relationship between features and cholesterol level and the difference of the importance of the feature across all age groups, so we utilized the following main variables:

LBXTC:(mg/dL) total cholesterol level

BMXWT:(kg) weight

BMXBMI: body mass index, a nonlinear combination of weight & height

LBXBPB:(ug/dL) blood lead concentration

LBXBCD(ug/dL) blood cadmium concentration

LBXTHG:(ug/dL) blood mercury concentration

LBXBSE:(ug/dL) blood selenium concentration

LBXBMN:(ug/dL) blood manganese concentration

RIDAGEYR:(yrs) age

INDFMPIR: Ratio of family income to poverty; lower means poorer

RIAGENDR: gender(1: male/2: female)

For the above variables, except for LBXTC, we have not made any changes to the original data of the variables. For the LBXTC variable, since the model we ultimately form is using binary classification, as the response variable of the model, we replace all values greater than 200 in this variable with 1, and other values are replaced with 0.

(adding plots. . .)

Methods

Results

Conclusions

Contributions

Reproducibility

To reproduce the same analysis as us,