Can total cholesterol level be predicted without a laboratory test for adults age 20 and over?

For my final capstone for the Thinkful Data Science Flex program I am planning on working on a supervised learning regression project that uses health, lifestyle, and diet information to predict total cholesterol levels.

This could be beneficial to a health and fitness tracker app which tracks your daily fitness levels from exercise you have done, nutrition levels from food you have eaten that you input, pulse and blood pressure (some fitness trackers do take blood pressure). It takes the tracked information along with persoal information you have input such as age, sex, race, weight, height, smoking and drinking habits, and any chronic conditions you may have and the app can then use the information to estimate total cholesterol. In the event total cholesterol is estimated to be high, a notification can pop up, reading “Have you had your cholesterol checked recently?”

The data I plan to use come from the National Health and Nutrition Examination Survey. https://www.cdc.gov/nchs/nhanes/index.htm. The survey and examination is conducted on a yearly basis. I plan to use the 2017 - 2018 data. I have not yet done an exploratory data analysis. If I feel more data is needed after performing my EDA, I will take data from the prior years.

The NHANES data comes in multiple files. These files are in SAS transport file format and can be read in using pandas read_sas function. The files I intend to use are: Demographic Variables and Sample Weights; Dietary Interview - Total Nutrient Intakes, First Day; Dietary Interview - Total Nutrient Intakes, Second Day; Blood Pressure; Cholesterol - Total, as well as a number of the questionnaire data files to gather information about chronic conditions, physical activity, and smoking, alcohol, and other substance use. I was also thinking of including medications but realized a person might not be willing to give their medication list to a health app.

The target variable will be LBXTC from the cholesterol total data file which measures Total Cholesterol (mg/dL). For modeling I plan to try linear regression, random forest regression, as well as neural networks. I anticipate using dimensionality reduction to visualize data and assist machine learning and also DBSCAN clustering to check for outliers.

The biggest challenge I anticipate facing is not finding a model that I feel performs well. I plan to use RMSE to compare models. As total cholesterol is usually below 300, a minimum RMSE of 50, for example, is very high. To improve RMSE I plan to play with different hyperparameters as well as play with features used. For example, I am planning on using the total nutrient data files, however maybe the individual food data files may be better.

On a final note, the reason I chose adults age 20 and over is because the suggested cholesterol levels are the same for men and women 20 and over according to https://medlineplus.gov/cholesterollevelswhatyouneedtoknow.html