Kyle Dillon - Final Project Proposal

Predicting Healthy BMI

Roles

Team Lead: Kyle Dillon

Recorder: Kyle Dillon

Spokesperson: Kyle Dillon

Background and Questions

Defined Research Question

Can we predict which adults have a healthy BMI primarily based on their nutrient intake, along with other demographic, laboratory, and questionnaire variables from the National Health and Nutrition Examination Survey? Which nutrients play the biggest role in making this decision?

Need

Numerous studies have been targeted towards predicting obesity and the causes behind it. However, fewer studies have been conducted on which factors keep people at a healthy BMI. The proposed question can actually provide greater insight and be a more effective way to combat this epidemic. Putting the focus on nutrient intake is important because of the role it plays in health, and it can be changed almost immediately, unlike demographics.

Why is it worth the time/ effort?

Looking at which nutrients are prominent in people who are considered healthy allows organizations, like the Department of Public Health, to take action by implementing

nutrition programs and initiatives to prevent people from becoming overweight or even obese. These programs can lead people to live healthier lifestyles and even save lives.

Novel/ Original

This question takes an alternate perspective on the question of obesity. Many studies have already been conducted trying to predict obesity using factors such as demographics, lifestyle, medical history, and calorie intake. By flipping the question on its head and focusing on what nutrients play a role in keeping people healthy, I believe this question takes more of a novel/ original path.

Stakeholder

The stakeholder I wish to target with this report is the Department of Public Health(DPH). On the DPH website under the “who we serve” section, they say the “DPH promotes and protects health and wellness and prevents injury and illness for all people” (https://www.mass.gov/orgs/department-of-public-health, 2025). With my question targeting more of the general population, it makes sense to present my findings to an organization that targets entire communities. This way, they can use the findings to implement policies or programs to help keep people healthy.

Hypothesis

Nutrition intake, along with demographic, questionnaire, and laboratory variables, will be able to predict if an adult has a healthy BMI because these variables are indicators of dietary behaviors, lifestyle patterns, and medical history, which affect a person's weight.

Prediction

Adults who have a balanced nutritional intake, a strong lifestyle, and a stable medical history will have a healthy BMI. Variables like protein, fiber, fats, sugar, and micronutrients will be the most influential when predicting healthy BMI.

Data and Methods

Data Sets

I have decided to use data sets from the National Health and Nutrition Examination Survey (NHANES), conducted by the National Center for Health Statistics (NCHS) on the Centers for Disease Control and Prevention (CDC) website (https://wwwn.cdc.gov/nchs/nhanes/continuousnhanes/default.aspx?Cycle=2021-2023). I will use features from different areas, including demographic, dietary, questionnaire, examination, and laboratory data. They are listed in the data dictionary section of this report. The NHANES data is good because it contains a variety of different studies with many relevant features for each, which can be used to help answer the research question at hand. The datasets are stored as XPT files on the CDC website. To bring them into R, the haven package within the tidyverse package is going to be needed. The haven package allows the XPT file to be read in using the read_xpt function. Once all the datasets are read in, they will need to be merged using the SEQN variable. This is the unique identifier for each individual. A couple of basic data cleaning tasks that have been conducted so far include looking at missing values and renaming variables.

Data Dictionary

Response / Outcome Variable

The outcome variable will be BMI. This variable is located in the body measures dataset within the examination section of the NHANES data and is labeled as BMXBMI. It is calculated using the metric system, and the formula is as follows:

BMXBMI = weight (kg) / height (m)2**

The threshold for being classified as healthy will be adults with a BMI between 18.5 and 24.9. BMIs under and over this threshold will be classified as unhealthy. It will be converted to binary with 0 being healthy and 1 being unhealthy.

Predictor Variables

I will be renaming the variables when data cleaning.

Demographics

- RIDAGEYR - Age in years at screening
- RIAGENDR - Gender
- RIDRETH1 - Race/Hispanic origin
- DMDBORN4 - Country of birth
- DMDEDUC2 - Education level - Adults 20+
- DMDMARTZ - Marital status
- DMDHHSIZ - Total number of people in the Household
- INDFMPIR - Ratio of family income to poverty

Dietray

I will be combining the total nutrition intake and the total dietary supplements variables to get the total amount of nutrients.

Total Nutrition Intakes

- DR1TKCAL - Energy (kcal)
- DR1TPROT - Protein (gm)
- DR1TCARB - Carbohydrate (gm)
- DR1TFIBE - Dietary fiber (gm)
- DR1TTFAT - Total fat (gm)
- DR1TCHOL - Cholesterol (mg)
- DR1TATOC - Vitamin E as alpha-tocopherol (mg)
- DR1TVARA - Vitamin A, RAE (mcg)
- DR1TVB1 - Thiamin (Vitamin B1) (mg)
- DR1TVB2 - Riboflavin (Vitamin B2) (mg)
- DR1TNIAC - Niacin (mg)
- DR1TVB6 - Vitamin B6 (mg)
- DR1TFDFE - Folate, DFE (mcg)
- DR1TCHL - Total choline (mg)
- DR1TVB12 - Vitamin B12 (mcg)
- DR1TVC - Vitamin C (mg)
- DR1TVD - Vitamin D (D2 + D3) (mcg)
- DR1TVK - Vitamin K (mcg)

- DR1TCALC - Calcium (mg)
- DR1TPHOS - Phosphorus (mg)
- DR1TMAGN - Magnesium (mg)
- DR1TIRON - Iron (mg)
- DR1TZINC - Zinc (mg)
- DR1TCOPP - Copper (mg)
- DR1TSODI - Sodium (mg)
- DR1TPOTA - Potassium (mg)
- DR1TSELE - Selenium (mcg)
- DR1TCAFF - Caffeine (mg)
- DR1TALCO - Alcohol (gm)
- DR1TMOIS - Moisture (gm)
- DR1_320Z - Total plain water drank yesterday (gm)

Total Dietary Supplements

- DSQTKCAL - Energy (kcal)
- DSQTPROT - Protein (gm)
- DSQTCARB - Carbohydrate (gm)
- DSQTFIBE - Dietary fiber (gm)
- DSQTTFAT - Total fat (gm)
- DSQTCHOL - Cholesterol (mg)
- DSQTVB1 - Thiamin (Vitamin B1) (mg)
- DSQTVB2 - Riboflavin (Vitamin B2) (mg)

- DSQTNIAC - Niacin (mg)
- DSQTVB6 - Vitamin B6 (mg)
- DSQTFDFE - Folate, DFE (mcg)
- DSQTCHL - Total choline (mg)
- DSQTVB12 - Vitamin B12 (mcg)
- DSQTVC - Vitamin C (mg)
- DSQTVK - Vitamin K (mcg)
- DSQTVD - Vitamin D (D2 + D3) (mcg)
- DSQTCALC - Calcium (mg)
- DSQTPHOS - Phosphorus (mg)
- DSQTMAGN - Magnesium (mg)
- DSQTIRON - Iron (mg)
- DSQTZINC - Zinc (mg)
- DSQTCOPP - Copper (mg)
- DSQTSODI - Sodium (mg)
- DSQTPOTA - Potassium (mg)
- DSQTSELE - Selenium (mcg)
- DSQTCAFF - Caffeine (mg)

Laboratory

- URXPREG - Urine Pregnancy Result
- LBXGLU - Fasting Glucose (mg/dL)
- LBXIN - Insulin (uU/mL)
- LBXTC - Total Cholesterol (mg/dL)

- LBXGH - Glycohemoglobin (%)

Questionnaire

- HIQ011 - Covered by health insurance
- MCQ160m - Ever told you had a thyroid problem
- MCQ550 - Has DR ever said you have gallstones
- MCQ160l - Ever told you had any liver condition
- MCQ053 - Taking treatment for anemia/past 3 mos
- BPQ101D - Taking meds to lower blood cholesterol?
- BPQ020 - Ever told you had high blood pressure
- BPQ030 - Told had high blood pressure - 2+ times
- SLQ300 - Usual sleep time on weekdays or workdays
- SLQ310 - Usual wake time on weekdays or workdays
- SLD012 - Sleep hours - weekdays or workdays
- SLQ320 - Usual sleep time on weekends
- SLQ330 - Usual wake time on weekends
- SLD013 - Sleep hours - weekends
- PAD800 - Minutes moderate LTPA
- PAD820 - Minutes of vigorous LTPA
- PAD680 - Minutes of sedentary activity
- PAD810Q - Frequency of vigorous LTPA
- PAD790Q - Frequency of moderate LTPA
- WHQ070 - Tried to lose weight in the past year

Analysis Plan

How will I know if my question is answered?

I will know if my question is answered if the models predict the class of an adult's BMI with accuracy based on their nutrient intake, along with other demographic, laboratory, and questionnaire variables. Essentially, if the model's predictions match the actual class. This will be done with accuracy tools, which are listed below in the evaluation section of the analysis plan. Also, when the models can identify which nutrition variables play the biggest role in making this decision.

How will I know if my hypothesis is supported?

I will know my hypothesis is supported if the models using nutritional intake, along with demographic, questionnaire, and laboratory variables, can accurately predict if an adult has a healthy BMI. These variables should be significant predictors of healthy BMI.

Premodeling Steps

- Data cleaning
    - Handling missing data
    - Renaming variables and inconsistencies in the data
    - Setting the threshold for “healthy weight” and converting the outcome variable(BMI) to a binary outcome. 0 being healthy and 1 being unhealthy.
- Data preprocessing
    - Creating new features
    - Normalizing or scaling numerical features
    - Encoding categorical features
- Exploratory Data Analysis EDA

    - Visualize patterns, trends, relationships, and issues with the dataset
- Deciding which features to use for the models
- Creating training and testing sets using an 80/20 split.

Models

- **Model 1: Linear Regression** - Good to use for a baseline model
- **Model 2: Random Forest** - Handling non-linear relationships, overfitting, and feature importance
- **Model 3: XGBoost** - More complex model with higher accuracy

Evulation

- ROC curves, AUC curves, F1 scores, Precision, Recall, Confusion Matrix

Pitfalls

Class Imbalance - If a majority of the people are normal weight or slightly overweight or underweight, there could be a bias, and the models might just predict the majority class.

Features - Picking the right amount of features to use for the models is going to be very important because using too many can lead to overfitting.

Multicollinearity - If two or more features have a high correlation, it could be challenging to determine how they individually affect the outcome variable.

Technical Details

Coding Language: R

Github Repo: https://github.com/Dillonkw/Data-Science-Capstone