

## Kyle Dillon - Preliminary Project Proposal

### Predicting Obesity

#### **Roles**

**Team Lead:** Kyle Dillon

**Recorder:** Kyle Dillon

**Spokesperson:** Kyle Dillon

#### **Background and Questions**

##### **Defined Research Question**

Can we categorize adults (people 18+) as either obese or not obese based on demographic data, macronutrients, micronutrients, and questionnaire data? Which features play the biggest role in making this decision?

##### **Need**

In today's society, obesity is a significant problem in the World. Obesity rates continue to increase with no signs of slowing down. Identifying which people are classified as obese and what factors contribute to this is crucial for various programs and initiatives targeted towards combating this epidemic.

##### **Why is it worth the time/ effort?**

It is worth the time and effort because of the positive implications it can have. Knowing which people are at risk of becoming overweight or obese, and the factors behind it, can help address the obesity problem. This can lead to people living healthier lifestyles and can even save lives.

##### **Novel/ Original**

The proposed question is novel/ original. I am not taking the improvement route, and as far as I know, I do not know of anyone else who is doing a question like this.

### **Stakeholder**

- Department of Public Health
- Population Health Services
- Primary Care Doctors

### **Hypothesis**

Demographic, dietary, questionnaire, and examination data will accurately predict the class of someone's BMI.

### **Prediction**

Dietary features, including total sugar, carbs, protein, energy, dietary fiber, total fat, and more, will be some of the most important features when predicting obesity risk.

### **Data and Analysis**

#### **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.

### **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:

$$\text{BMI} = \text{weight (kg)} / \text{height (m)}^2$$

### **Predictor Variables**

This list of features is not exhaustive

#### Demographics

- Age
- Gender
- Country of birth
- Education level
- Ratio of family income to poverty
- Race
- Marital status
- Household size

#### Dietray

- Grams
- Energy
- Protein
- Carbs
- Total sugar
- Dietary Fiber
- Total fat

- Total saturated fatty acid
- Cholesterol

#### Questionnaire

- Income questionnaire
- Physical activity questionnaire
- Medical conditions questionnaire
- Diet behavior and nutrition questionnaire
- Insulin questionnaire

#### Examination

- Weigh
- Standing height
- BMI index

#### Tentative Analysis Plan

- Final Dataset Selection
- EDA
- Data Preprocessing
- Modeling
- Evaluation Metrics

#### 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.

**How will I know if my question is answered?**

I will know if my question is answered if the models predict the class of people's BMI with accuracy. Essentially, if the model's predictions match the actual class. This will be done with accuracy tools, including accuracy, ROC curves, AUC curves, F1, and more. Also, determining the probabilities of each of the categories to see any early indicators that someone may be shifting to another category. Finally, identifying what features 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 accurately predict the class of someone's BMI based on the data listed in the hypothesis. Essentially, if the predicted class is the same as the actual class.

**Technical Details**

**Language:** R

**Resources:** No other resources

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

