What are the main drivers of obesity, and which communities are impacted most significantly?

Lauren Chen, Doma Ghale, Tanisha Gittens, Pakize Sanal

https://github.com/laurenc8/ds4a-obesity

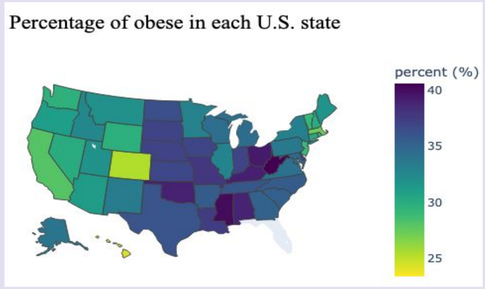
Highlights

- Washington DC shows the healthiest weight distribution with the lowest obesity (23%) and overweight (33%) rates, and highest normal weight (42%) among the given states.
- Average BMI starts to decrease above income of 75,000
- Among all models, Random Forest Classifier model provided the most accurate predictions.

Background

Adult obesity in the U.S. has become a significant concern, with the prevalence of obesity increasing dramatically over the past few decades. According to the CDC approximately 43% of U.S. adults were classified as obese in 2020. This growing obesity rate has led to a corresponding rise in serious health conditions and has equally concerning economic implications. By 2018, the healthcare costs associated with obesity treatments were estimated to reach approximately $344 billion [Journal of Health Economics, 2018]. Addressing this issue requires a comprehensive and coordinated effort across various sectors, including healthcare, education, business, and government [WHO, 2023].

Individuals with Body Mass Index (BMI) value at or above 3000 is considered to be obese, where BMI value is calculated as (weight in pound ÷ square of height in inches) x 703.



Data

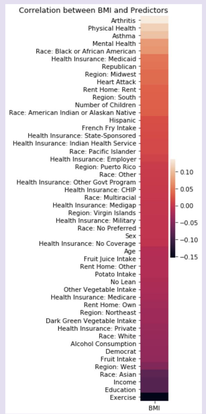
We used 2021 data from the Behavioral Risk Factor Surveillance System (BRFSS). Data are collected from a random sample of U.S. adults (one per household) through a telephone survey. There were 438,693 respondents and 303 features.

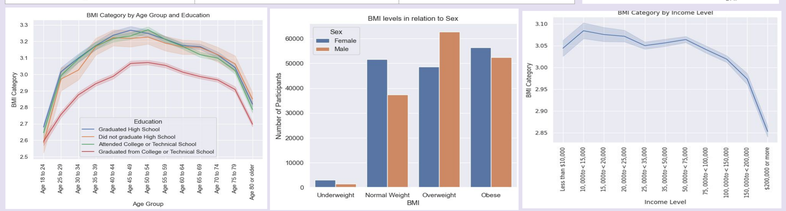
1. **Data Cleaning:** we first extracted 25 relevant columns, spanning demographics (e.g. age, income, race), fruit and vegetable intake, lifestyle (e.g. exercise, alcohol consumption), and medical history (e.g. asthma, mental health). We then handled missing data by either dropping it or imputing with a median, as well as checked for outliers and erroneous data. In the end, we were left with 201,286 data points.
2. **Exploratory Data Analysis:** we plotted the distributions of the variables in the form of histograms, box plots, and scatter plots. In addition, we plotted the correlation matrix between the variables.
3. **Data Preprocessing:** depending on the data type of each variable, we used binary encoding, one hot encoding, and ordinal encoding to convert categorical variables into numeric ones.

Exploratory Data Analysis

- Correlation plot on right shows no correlation
- Mean BMI is lower for younger and older population and higher for middle age population across all education level
- Mean BMI for college graduates is lower than other education levels across all age group
- Mean BMI decreases as individual's income increases
- Men have higher number of overweight and lower number of normal weight than female

U.S. states with highest and lowest percentage in each BMI category		
	Highest %	Lowest %
Obese	West Virginia 42%	Washington DC 23%
Overweight	Rhode Island 38%	Washington DC 33%
Normal weight	Washington DC 42%	West Virginia 24%
Underweight	Hawaii 2%	Alaska 1%





Hypothesis Tests

Using independent t-test we found that the mean BMI between male and female are not significantly different, mean BMI of home owners is significantly lesser than non-home owner and mean BMI of east and west coast regions is significantly lesser than mid-regions. We note that these tests are more likely to show significant differences because of large sample size .

.

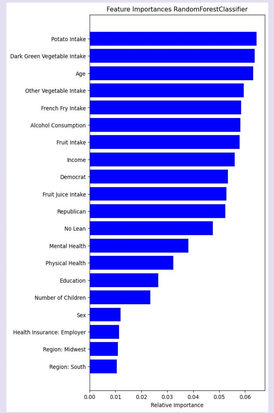
Model

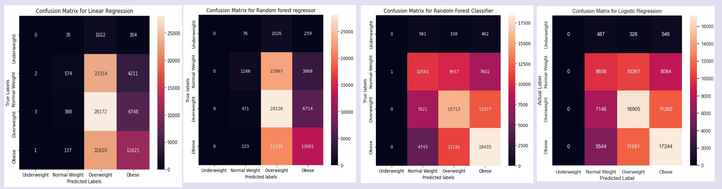
We build four types of models: linear regression and random forest regression to predict BMI, and logistic regression and random forest classifier to predict BMI categories (underweight, normal weight, overweight, or obese). The choice of these models hinged on their interpretability and appropriateness for the task. We prioritized our models' ability to accurately predict obese class. However, these models may have limitations in accuracy and individual sensitivity.

Results

- Based on recall on obese class, random forest classifier performed the best.
- recall = number of individuals correctly predicted as obese out of all obese individuals.

Model	Mean Absolute Error (MAE)	'Obese' Class Accuracy (Recall)
Linear Regression	4.71	33.81%
Random Forest Regression	4.56	37.82 %
Logistic Regression	NA	50.17%
Random Forest Classifier	NA	53.63%





Conclusions

Our study used 2021 survey data to examine obesity in the U.S. We found obesity rates vary by state, age, and income. West Virginia had the most obesity, and Washington DC the least. Middle-aged people and lower-income individuals had higher obesity rates. We used four types of models to predict obesity. All models showed some accuracy but also highlighted the complexity of predicting obesity. The results emphasize the need for public health interventions, particularly in higher-risk communities.

Future Work

Some of the ways we could possibly improve our model's performance is by finding more relevant features (currently only 8% is used), performing hyperparameter tuning on our best mode (Random Forest Classifier) and using different models.