An Analysis of Obesity Levels & BMI from Nutrition, Lifestyle, & Physical Characteristics

DS 6021: Final Project

Bela Barton, Ben Doniger, Emily Garman,
Natalie Seah, & Erin Siedlecki

Group 10

**BMI and obesity categories are metrics that are limited, oversimplified, and outdated. While we use them for modeling for the purposes of a large, population-based analysis, they should not be used to judge an individual's beauty, health, or value.

Introduction & Data Cleaning

The Dataset:

- "Dataset for estimation of obesity levels based on eating habits and physical condition in individuals from Colombia, Peru and Mexico" by Fabio Mendoza Palechor and Alexis de la Hoz Manotas
- Collected via anonymous online survey
- Generated additional synthetic data to balance the dataset

Variables:

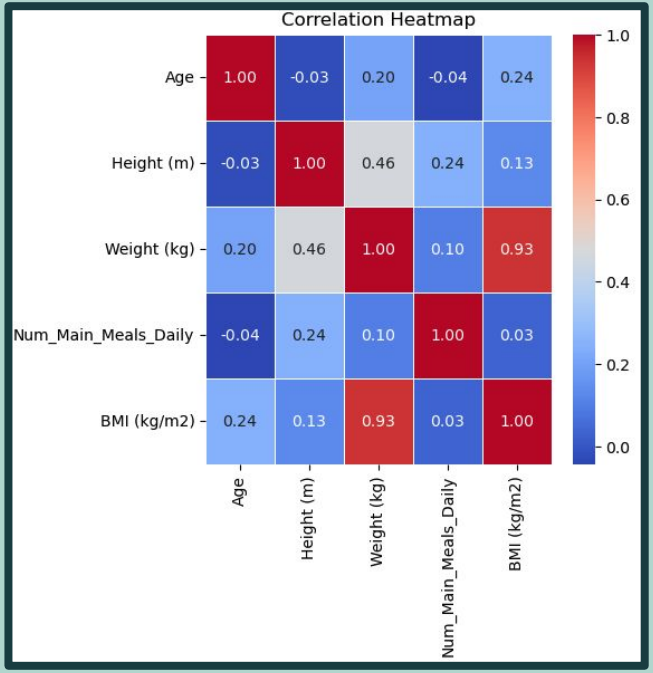
- Predictors: Explored features such as family history, age, gender, daily meals, water and alcohol intake, exercise, and transportation
- Response: Obesity Level (Insufficient Weight, Normal Weight, Overweight Levels I and II, Obesity Types I, II, and III) and BMI (kg/m^2)

Data Cleaning Process:

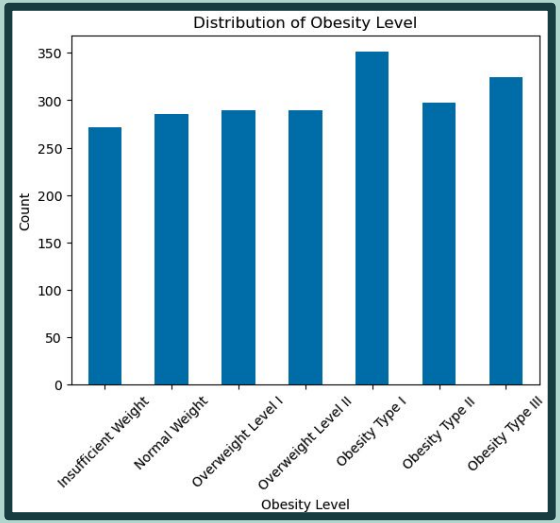
1. Initial Audit and Validation
2. Column Remaining for Interpretability
3. Categorical Standardization and Consistency Fixes
4. Repairing Ordinal Variables Based on Codebook
5. Categorical Type Conversion
6. Outlier Diagnostics
7. Class Distribution and Stratification
8. Output and Reproducibility

#	Column	Non-Null Count	Dtype
0	Gender	2111 non-null	object
1	Age	2111 non-null	float64
2	Height (m)	2111 non-null	float64
3	Weight (kg)	2111 non-null	float64
4	Overweight_Family_History	2111 non-null	object
5	High_Calorie_Consumption_Often	2111 non-null	object
6	Vegetable_Consumption_Often	2111 non-null	object
7	Num_Main_Meals_Daily	2111 non-null	int64
8	Eat_Between_Meals	2111 non-null	object
9	Smoke_Regularly	2111 non-null	object
10	Water_Drank_Daily	2111 non-null	object
11	Calories_Monitored_Daily	2111 non-null	object
12	Workout_Frequency	2111 non-null	object
13	Time_Using_Technology_Daily	2111 non-null	object
14	Alcohol_Consumption_Frequency	2111 non-null	object
15	Means_of_Transportation	2111 non-null	object
16	Obesity_Level	2111 non-null	object
17	BMI (kg/m2)	2111 non-null	float64

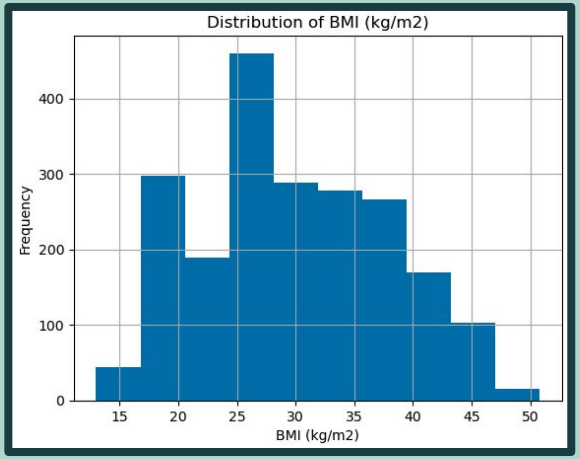
Exploratory Data Analysis



- Weight and BMI are highly correlated, which is expected since weight is used to calculate BMI. Because height and weight are used to calculate BMI we excluded them from our modeling.
- The other numeric variables do not have a strong correlation with each other



- The value counts of Obesity Levels are pretty consistent, indicating there are no major class imbalances that would hinder our models

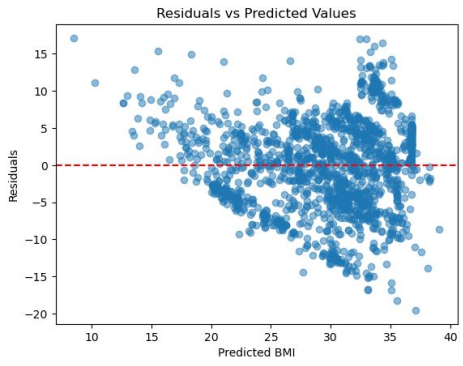


- The distribution of BMI (kg/m^2) is relatively unimodal
- There is a slight right skew and the majority of individuals fall in between 25-35 kg/m^2

Linear Regression

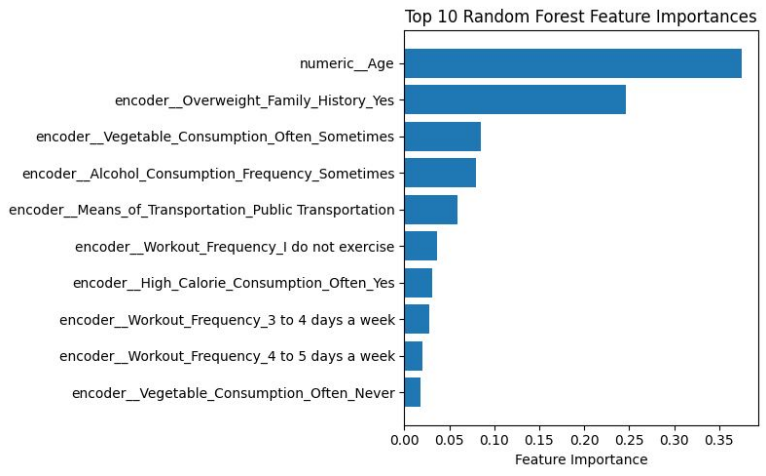
Which eating habits, lifestyle factors, and physical activity factors are most predictive of BMI?

- **Overall F-test: significant** → regression model explains a significant portion of the variability in BMI
- **Significant predictors:** Overweight family history, high calorie consumption, vegetable consumption, calories monitored daily, workout frequency, alcohol frequency, transportation, age
- **Performance:** R^2 = 0.35, RMSE = 6.57
- **Issues:** Linearity & constant variance of errors assumptions violated



Random Forest

- **Performance:** R^2 = 0.66, RMSE = 4.77
- Captures non-linear patterns & interactions
- Substantially better fit than linear models



Multiclass Logistic Regression

1. **Which eating habits, lifestyle factors, and physical activity factors are most predictive of obesity level?**
2. **Can we accurately classify an individual's obesity level using our predictors?**
    - Classifying Obesity Level (1 of 7 types): Insufficient Weight, Normal Weight, Overweight Level I, Overweight Level II, Obesity Type I, Obesity Type II, & Obesity Type III

Results ***(rounded)***

Evaluation Metrics	Baseline Logistic Regression Model	Lasso Regression Model
Training Accuracy	0.633	0.630
Testing Accuracy	0.637	0.618
Log Loss	1.063	1.089
ROC AUC	0.888	–
Mean CV Accuracy	0.607	–
Mean CV Log Loss	1.053	–

Most Predictive Features

Features related to:

- Overweight family history
- Calorie consumption & monitoring
- Gender
- Eating between meals
- Means of Transportation

Best Model?

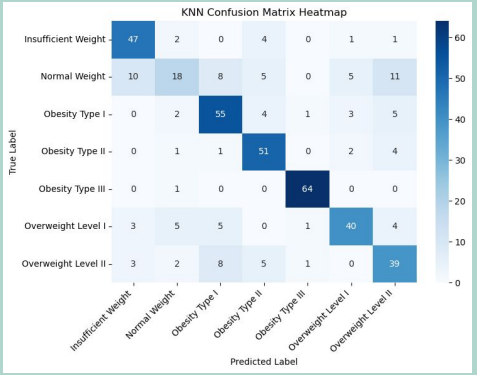
- In terms of lowest log loss and highest test accuracy, the baseline logistic regression model is best
- In terms of interpretability and variable selection, the lasso model is best

K-Nearest Neighbors

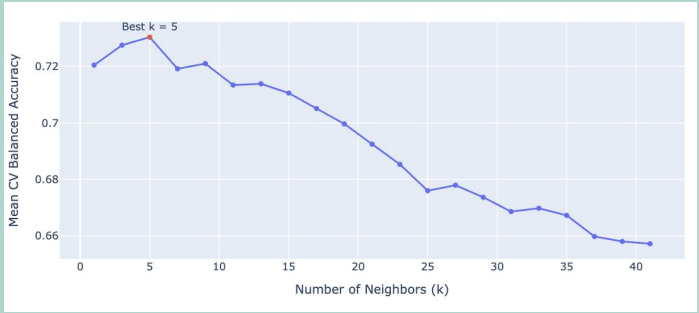
Best k = 5

CV Balanced Accuracy	Test Accuracy	Test Balanced Accuracy
0.730	0.744	0.738

- Strong recall for extreme classes
  - Obesity III: 0.98
  - Obesity II: 0.86
  - Insufficient: 0.85
- Normal weight hardest to classify (recall 0.32)
- Misclassifications mostly among adjacent weight categories



Individual	Key Traits	Predicted Level
Male, 54	High-calorie diet, little exercise, snacking, etc	Overweight Level II
Female, 28	Regular exercise, good hydration, frequent vegetable consumption	Normal Weight



K-Prototypes

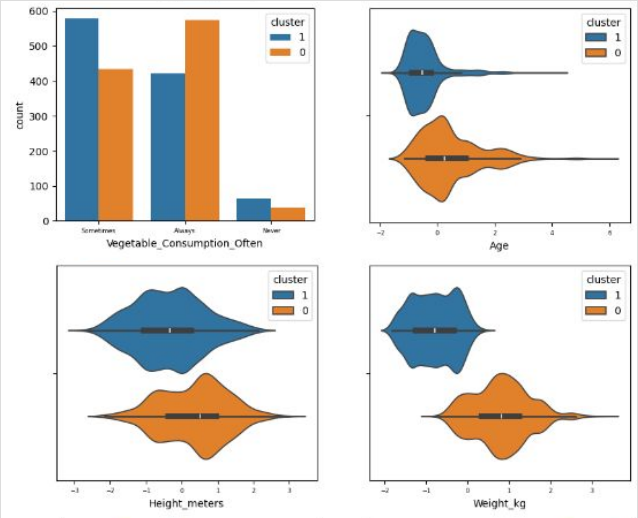
Do patterns emerge from the data relating to BMI?

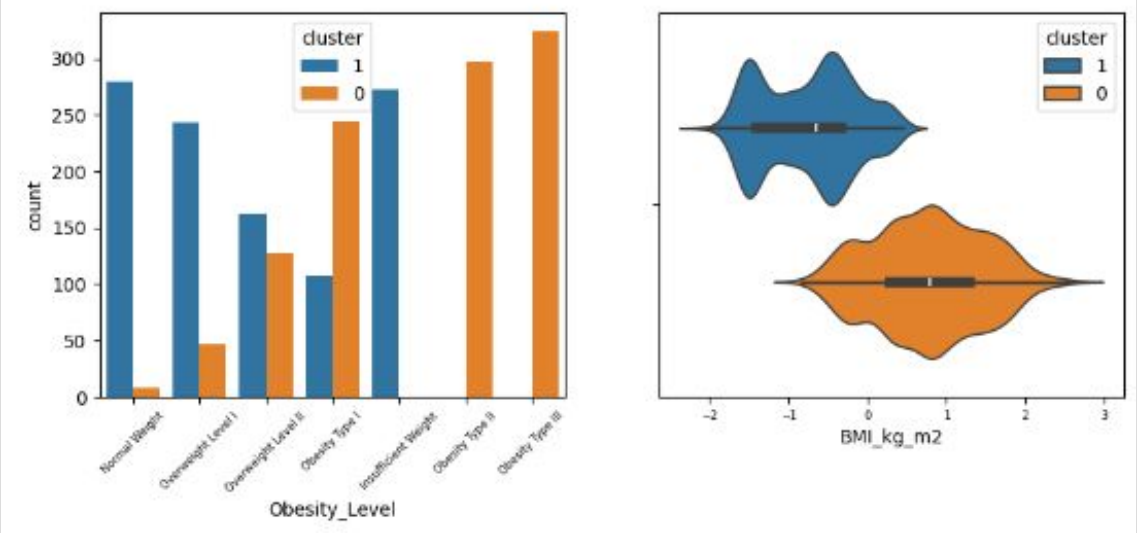
Variables that differentiated the most:

- Age, Height, Weight
- Vegetable Consumption, Obesity Level, BMI

Conclusions:

- Cluster 0 is higher BMI values and higher risk Obesity Levels
- Cluster 1 is lower BMI values and lower risk Obesity Levels



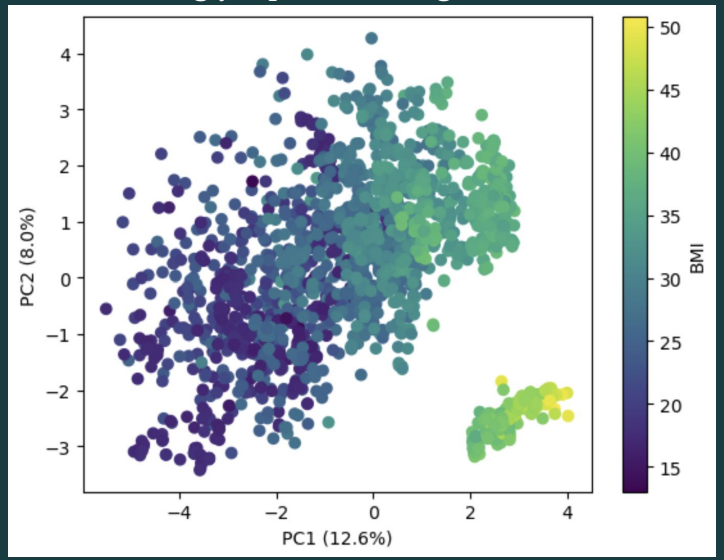


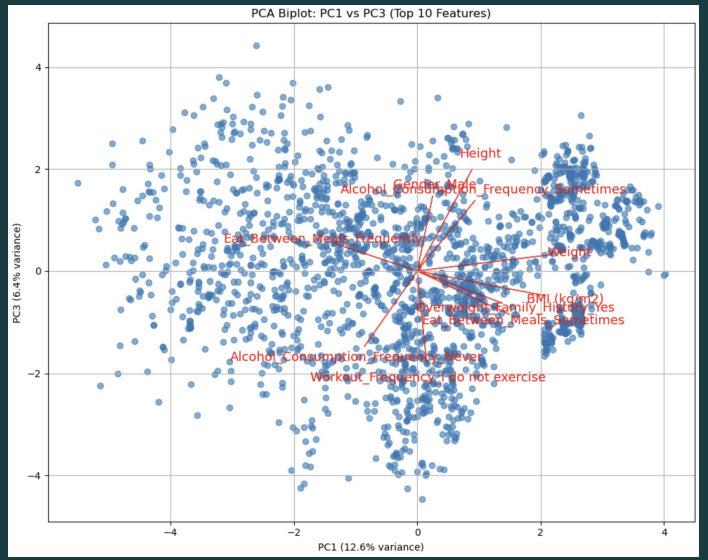
Principal Component Analysis

- Using PCA clustering to uncover more relationships between predictors and BMI
- PC1 captures BMI-relevant variation (2 clear clusters)
- PC2 does not seem to be strongly related to BMI
- PC3 has very small loadings and appears to be influenced by different variables, but people are not strongly separated along PC3

PC1 Contributive Features:

- Eat between meals sometimes
- Overweight family history
- Obesity level type III
- Often high calorie consumption





Shiny App Demo

https://natalieseah.shinyapps.io/bmi_prediction_dashboard/