Predicting Obesity:

A Machine Learning Approach to Analyzing Lifestyle Factors in Latin America

Assiba Lea Apovo
Cassidy Cruz
Daniel Pineda
Edward Tabije
Widchy Joachim

Background

Overweight and obesity have immediate and potentially long-term health impacts that include:

- Respiratory difficulties
- Increased risk of fractures
- Hypertension
- Early markers of cardiovascular disease
- Insulin resistance
- Psychological effects
- Increased risk of non-communicable diseases

The worldwide prevalence of obesity increased

+200%

from 1975 to 2022

Source: World Obesity Federation

Obesity is now recognised as one of the most important public health problems facing the world today.

This is especially true in Latin America

57%

of adults in Latin America are considered overweight

24%

are considered obese

vs **13%**

global estimate

Source: FAO, IFAD, PAHO,UNICEF and WFP. 2023. Latin America and the Caribbean - Regional Overview of Food Security and Nutrition 2023: Statistics and trends. Santiago.

Goal

Develop a machine learning model that can support health initiatives to manage overweight and obesity by predicting obesity levels based on lifestyle parameters.

Obesity Levels

$$BMI = \frac{weight\,(kg)}{height^2\,(m^2)}$$

Underweight

BMI: < 18.5

Normal Weight

BMI: 18.5 to 24.9

Overweight I

BMI: 25.0 to 26.9

Overweight II

BMI: 27.0 to 29.9

Obese I

BMI: 30.0 to 34.9

Obese II

BMI: 35.0 to 39.9

Obese III

BMI: > 40.0

Model Requirements

Minimum 75% accuracy

Model must meet a minimum accuracy threshold to be considered useful

Interpretability

Model must be able to convey the weight or importance of the lifestyle factors it uses to make its predictions so as to inform what actions may be taken to address obesity in a population

Flexible

Model must be robust against overfitting and not be biased towards the data used to train it

The Data

Source

[Palechor FM, Manotas AH. Dataset for estimation of obesity levels based on eating habits and physical condition in individuals from Colombia, Peru and Mexico. 2019 Aug]()

SMOTE

A technique that addresses imbalanced datasets by generating synthetic data points.

ETL

- Checked for Nulls & inconsistencies
- Renamed Columns
- Changed Data types
- Loaded refined data into a PostgreSQL Database

Data Source

Palechor FM, Manotas AH.
Dataset for estimation of obesity levels based on eating habits and physical condition in individuals from Colombia, Peru and Mexico.
2019 Aug

Data Collection Method

Online survey

Countries

Mexico, Peru, and Colombia

Age range

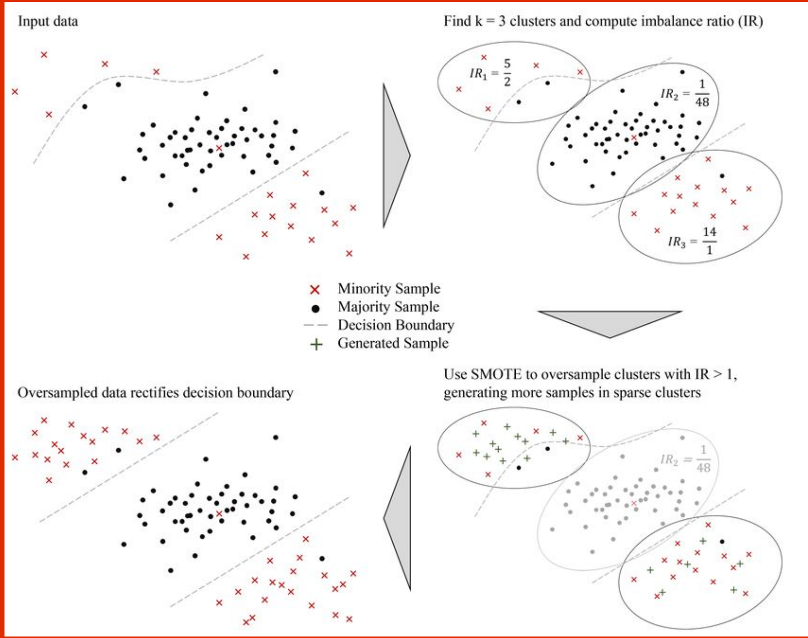
14 - 61 years old

Attributes Collected

- Frequent consumption of high caloric food
- Frequency of consumption of vegetables
- Number of main meals consumed daily
- Consumption of food between meals
- Consumption of water daily
- Consumption of alcohol
- Calories consumption monitoring
- Physical activity frequency
- Time using technology devices
- Transportation used
- Gender
- Age
- Height
- Weight
- Obesity Level (based upon BMI derived from height and weight)

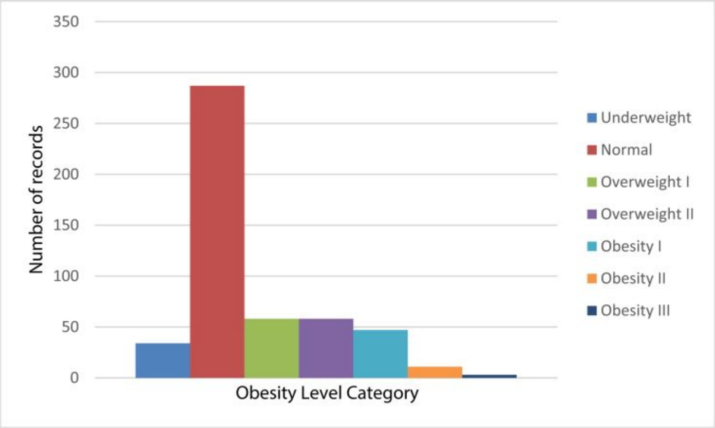
SMOTE

Synthetic Minority Oversampling Technique

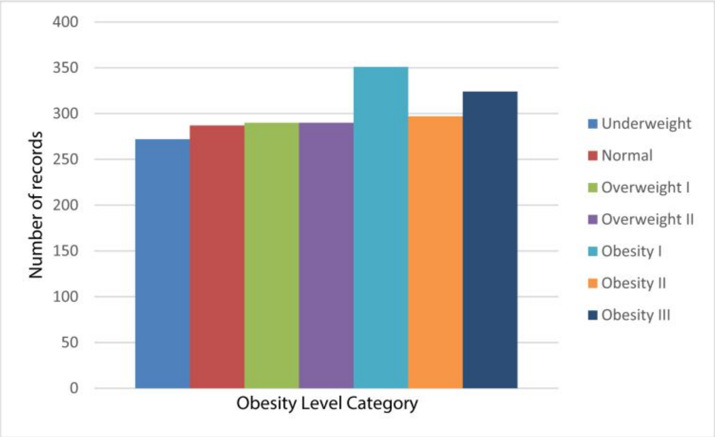


Last, Felix & Douzas, Georgios & Bação, Fernando. (2017). Oversampling for Imbalanced Learning Based on K-Means and SMOTE. 10.48550/arXiv.1711.00837.

Initial Data Collected



Data Balanced via SMOTE



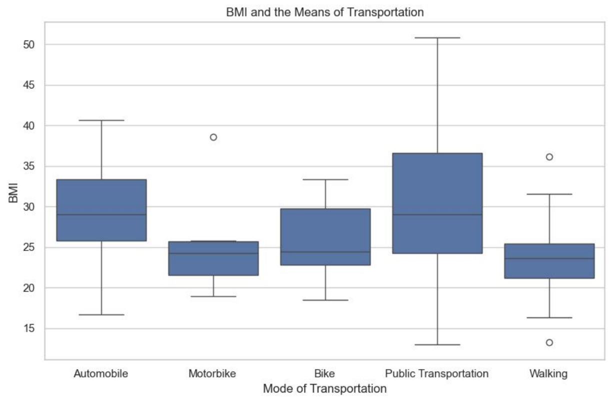
Database Entity Relationship Diagram (ERD)

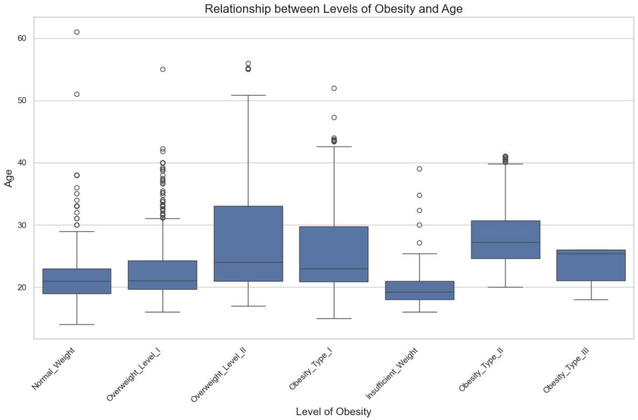
After the raw data was reviewed and refined, the cleaned data was uploaded to a single table in a PostgreSQL database

participants	
id	serial
gender	varchar
age	float
height_m	float
weight_kg	float
family_history_with_overweight	boolean
high_calorie_intake	boolean
vegetable_consumption	float
daily_meal_count	float
food_between_meals	varchar
smoking_habit	boolean
water_consumption	float
tracks_daily_calories	boolean
exercise_frequency	float
tech_usage_time	float
alcohol_intake	varchar
transportation_used	varchar
obesity_level	varchar

Feature Selection

- Exploratory data analysis established that all other attributes in the dataset were viable candidates based on relationships/patterns connected to BMI
- Height and weight excluded given their direct relationship to BMI/Obesity level classifications





Features

List of initial features to include in the initial model aimed at predicting obesity levels (target)

Gender
Male | Female

Age
Years

Frequent consumption of high caloric food
Yes | No

Frequency of consumption of vegetables
Never | Sometimes | Always

Number of main meals consumed daily
One | Two | Three | More than Three

Consumption of food between meals
No | Sometimes | Frequently | Always

Consumption of water daily
Less than a liter | Between 1 and 2 L | More than 2 L

Consumption of alcohol
No | Sometimes | Frequently | Always

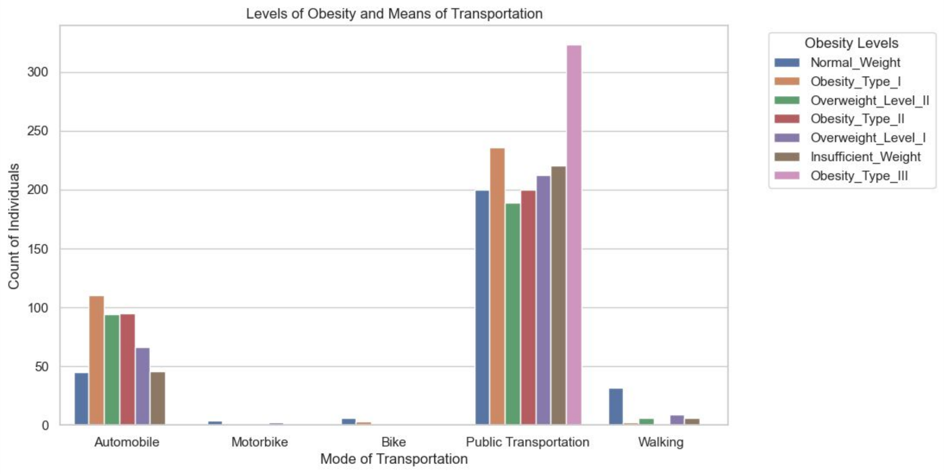
Calories consumption monitoring
Yes | No

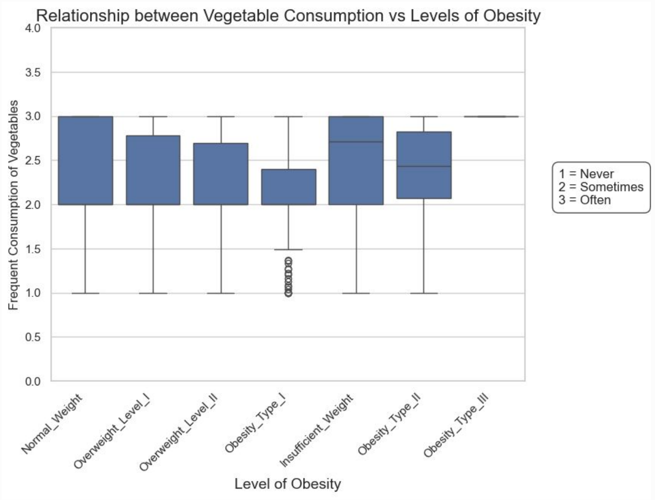
Physical activity frequency
I do not have any | 1 or 2 days | 2 or 4 days | 4 or 5 days

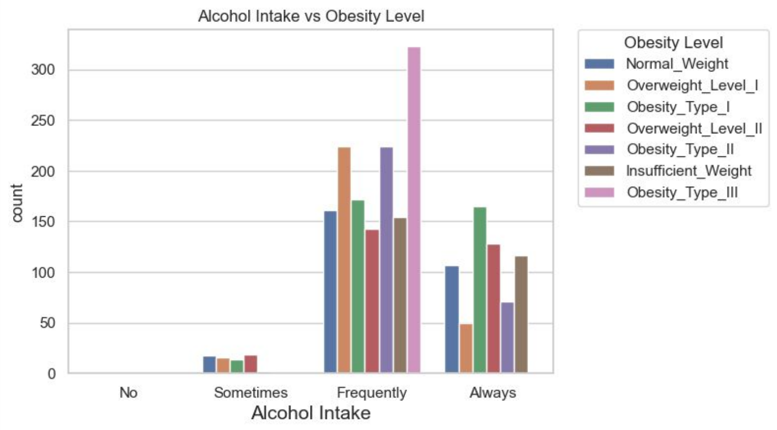
Time using technology devices
0–2 hours | 3–5 hours | More than 5 hours

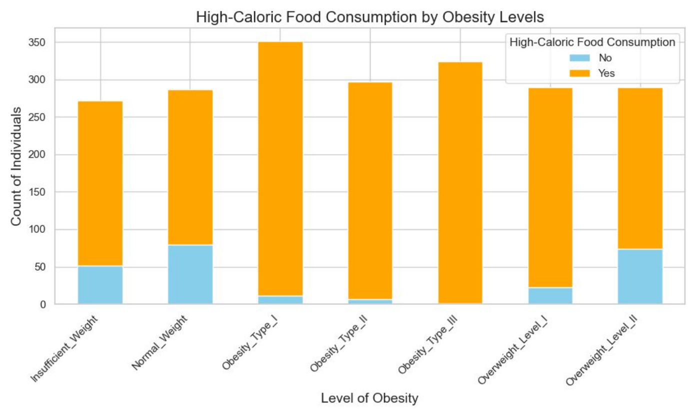
Transportation used
Automobile | Motorbike | Bike | Public Transportation | Walking

Relationship between Obesity and Impacting Factors prior to Model Optimization

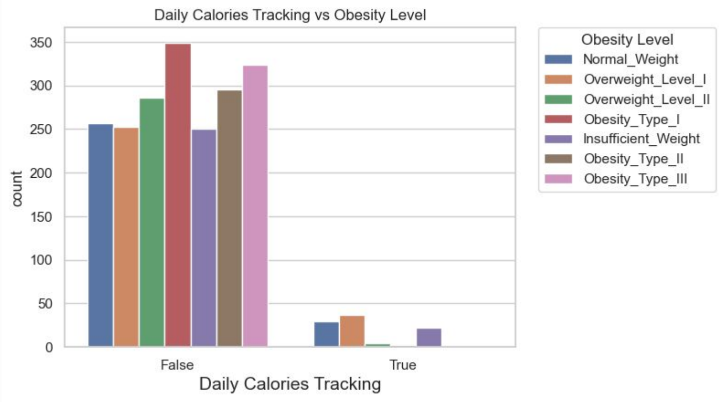


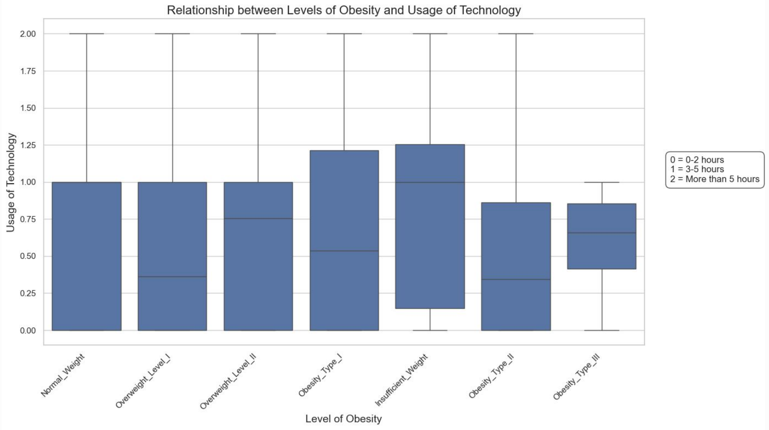


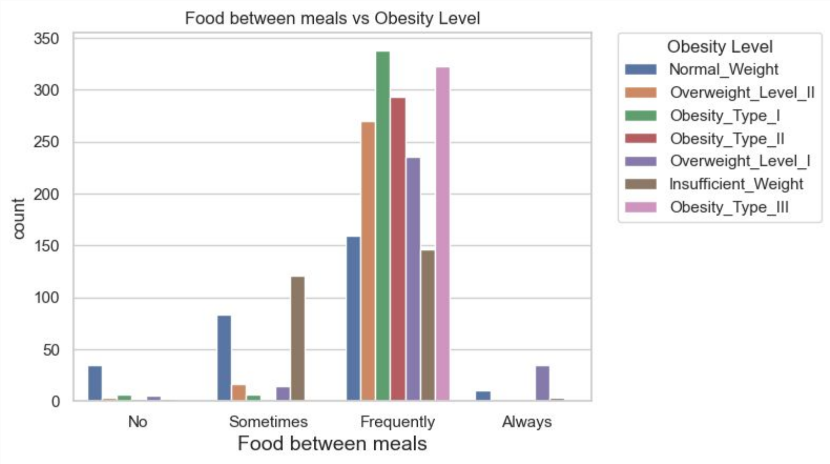


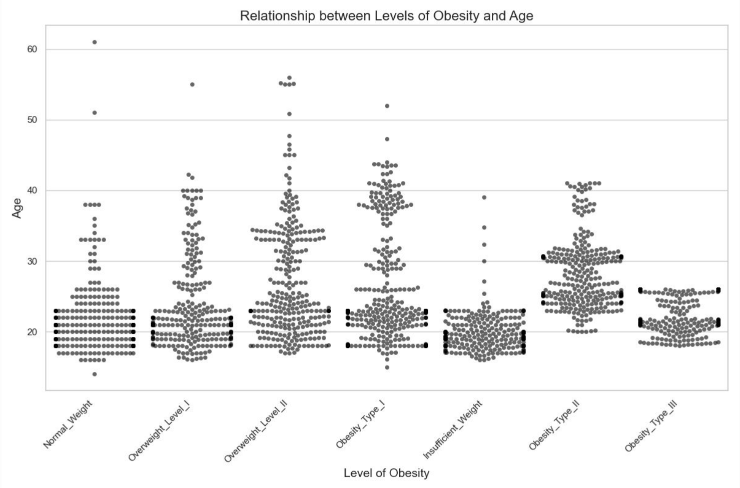


Relationship between Obesity and Impacting Factors prior to Model Optimization









Selecting the appropriate model type

Decision Tree	Gradient Boosting	Random Forest 🏆
Minimum 75% accuracy ✅ Prediction Accuracy: 77.30%	**Minimum 75% accuracy ✅** Prediction Accuracy: 78.49%	**Minimum 75% accuracy ✅** Prediction Accuracy: 84.87%
Interpretability ✅ Influence of features upon decision making can be measured via Gini Importance	**Interpretability ✅** Influence of features upon decision making can be measured via Gini Importance	**Interpretability ✅** Influence of features upon decision making can be measured via Gini Importance
Flexible ❌ Average Cross-Validation Accuracy: 74.95%	**Flexible ✅** Average Cross-Validation Accuracy: 79.80%	**Flexible ✅** Average Cross-Validation Accuracy: 85.19%

Optimizing our model

I. Feature Engineering

Manipulating, removing, or adding to the data or features included in the overall model

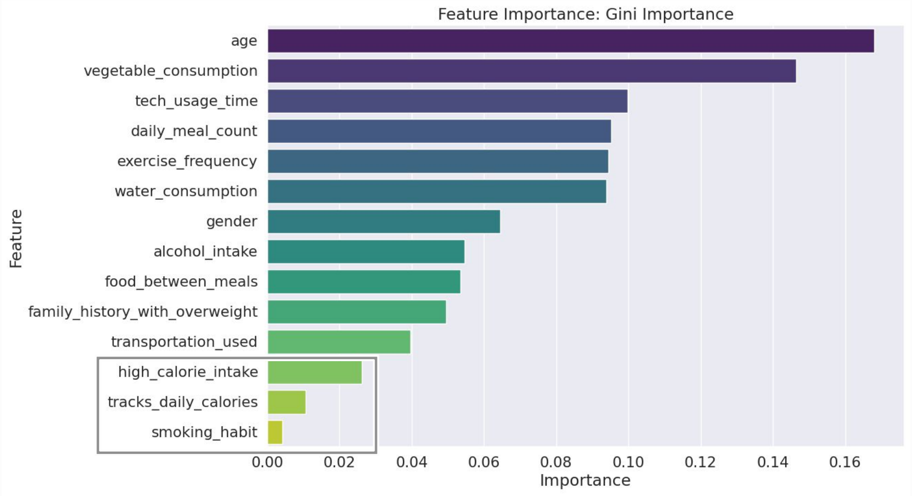
II. Hyperparameter Tuning

Changing the actual structure of the model itself

I. Feature Engineering

- Adding new data presented challenges given the complexity of the data and the presence of synthetic samples

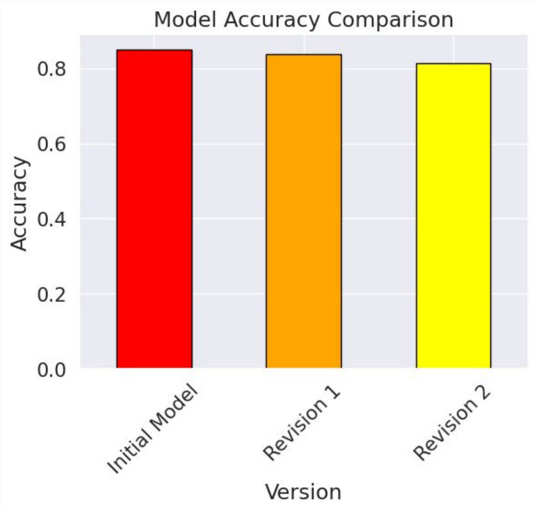
- Model performance could still be improved by removing features of lower Gini importance that may be causing ‘noise’

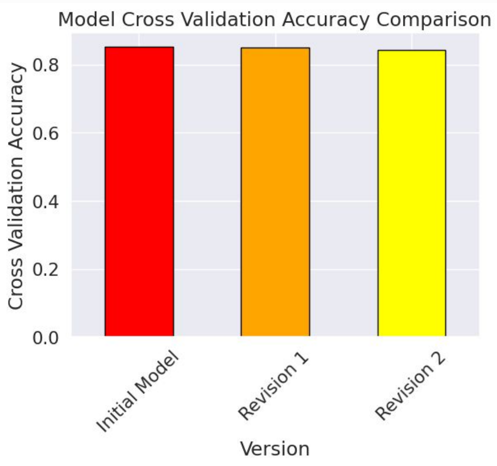


I. Feature Engineering

- **Optimization Attempt# 1:**
  - Removes the two features with the lowest feature importance (smoking_habit, tracks_daily calories)
  - Accuracy **decreased** to 83.69%
  - Average CV Accuracy **decreased** to 85.13%

- **Optimization Attempt# 2:**
  - Removes the three features with the lowest feature importance (smoking_habit, tracks_daily calories, high_calorie_intake)
  - Accuracy **decreased** to 81.32%
  - Average CV Accuracy **decreased** to 84.36%





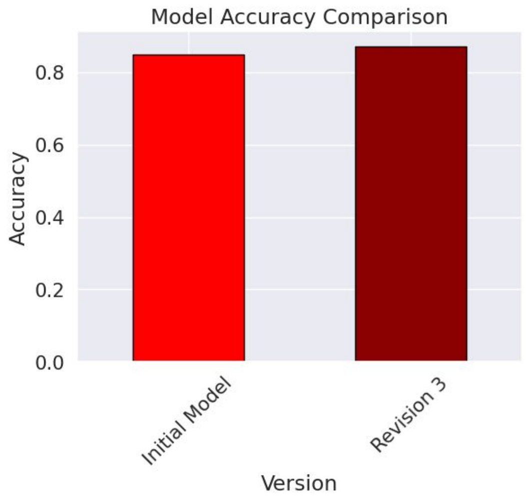
II. Hyperparameter Tuning

Optimization Attempt #3

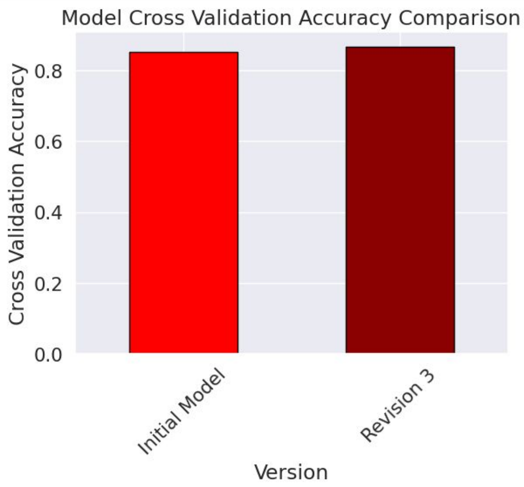
- Leverage Random Search Cross Validation to determine the hyperparameters that optimize accuracy

```
Initial Model Parameters
{'bootstrap': True,
 'ccp_alpha': 0.0,
 'class_weight': None,
 'criterion': 'gini',
 'max_depth': None,
 'max_features': 'sqrt',
 'max_leaf_nodes': None,
 'max_samples': None,
 'min_impurity_decrease': 0.0,
 'min_samples_leaf': 1,
 'min_samples_split': 2,
 'min_weight_fraction_leaf': 0.0,
 'monotonic_cst': None,
 'n_estimators': 100,
 'n_jobs': None,
 'oob_score': False,
 'random_state': 42,
 'verbose': 0,
 'warm_start': False}
```

```
Revision 3 Model Parameters
{'bootstrap': False,
 'ccp_alpha': 0.0,
 'class_weight': None,
 'criterion': 'gini',
 'max_depth': 70,
 'max_features': 'sqrt',
 'max_leaf_nodes': None,
 'max_samples': None,
 'min_impurity_decrease': 0.0,
 'min_samples_leaf': 1,
 'min_samples_split': 5,
 'min_weight_fraction_leaf': 0.0,
 'monotonic_cst': None,
 'n_estimators': 1577,
 'n_jobs': None,
 'oob_score': False,
 'random_state': 42,
 'verbose': 0,
 'warm_start': False}
```

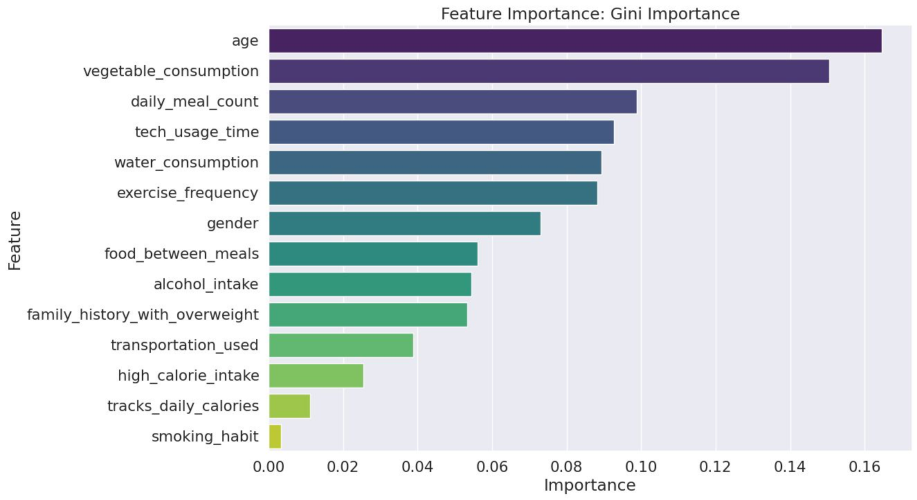


Accuracy **increased** to 87.00%



Average CV Accuracy **increased** to 86.55%

Results of Final Model



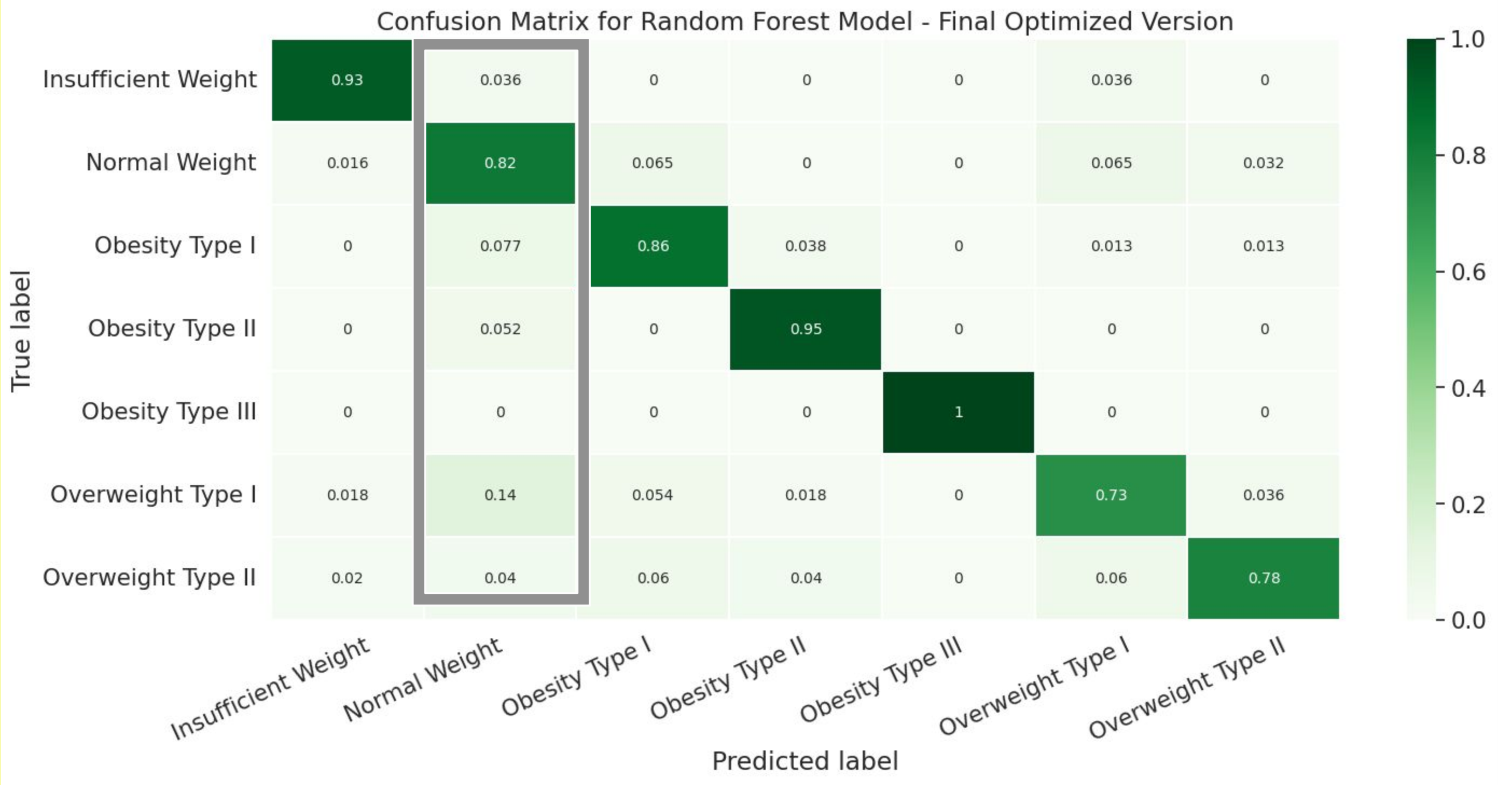
Random Forest Classifier

Overall Accuracy: 87.00%
Average Cross Validation Accuracy: 86.55%

Classification Report:

	precision	recall	f1-score	support
Insufficient_Weight	0.95	0.93	0.94	56
Normal_Weight	0.71	0.82	0.76	62
Obesity_Type_I	0.87	0.86	0.86	78
Obesity_Type_II	0.90	0.95	0.92	58
Obesity_Type_III	1.00	1.00	1.00	63
Overweight_Level_I	0.80	0.73	0.77	56
Overweight_Level_II	0.89	0.78	0.83	50

Results of Final Model



All Model requirements are met

Minimum 75% accuracy✅

Interpretability✅

Flexible✅

Potential real world applications

Guide Health Initiatives

The features with the highest Gini importance can inform policy makers and healthcare providers on what major factors they can focus on shifting in order to manage the prevalence of obesity. Possibile examples include:

- Age: Strategize how they reach out to different age demographics
- Vegetable Consumption: Launch campaigns that encourage the consumption of fresh vegetables

Diagnose Health Conditions

Have the model make predictions on obesity levels based upon new data obtained from an individual or group

- If the model's prediction does not align with the person's actual obesity level (via BMI derived from their height and weight), it could be a potential link to another underlying problem

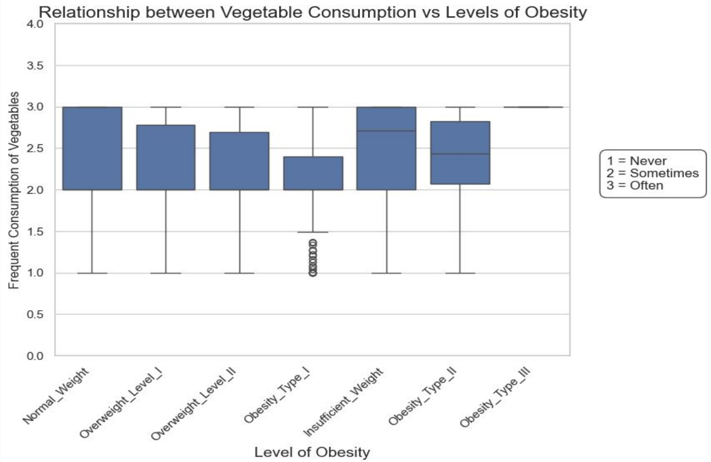
Questions?

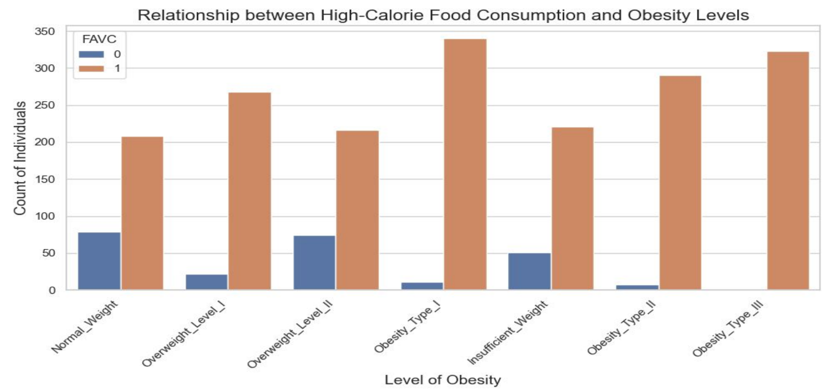
Thank you!

Appendix

Attributes: Eating Habits

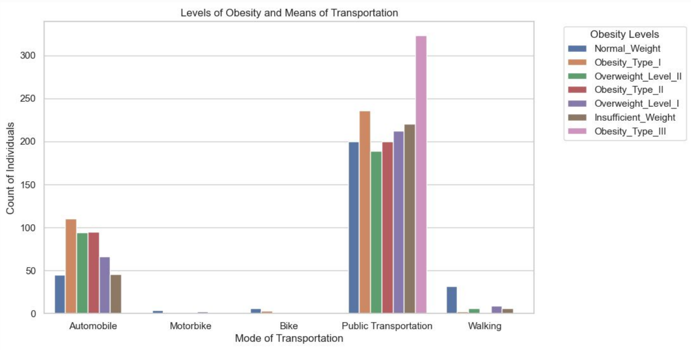
Exploratory data analysis on the relationship between BMI/Obesity level classifications and attributes relating to eating habits

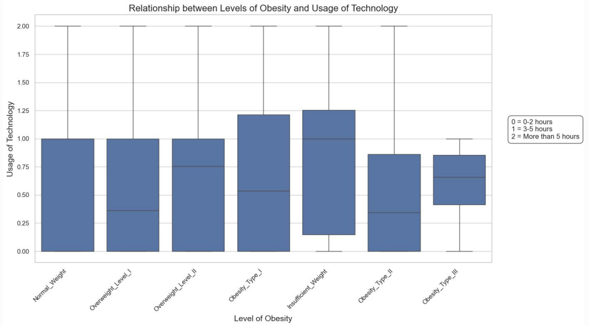


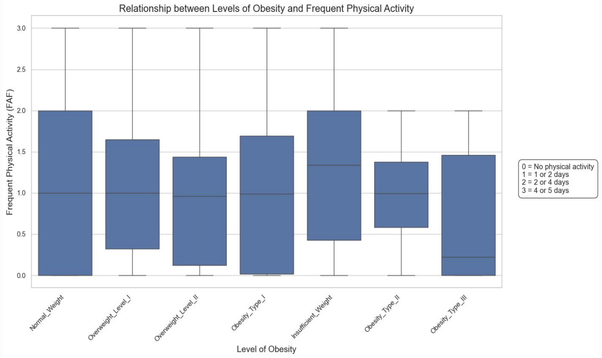


Attributes: Physical Activity

Exploratory data analysis on the relationship between BMI/Obesity level classifications and attributes relating to physical activity







Attributes: Age

Exploratory data analysis on the relationship between BMI/Obesity level classifications and age