Title: Detecting Obesity Level with Predictive Modeling

BUS AN 516 | Team 9

Hyewon Jeong, Samvit Patankar, Lucy Tsai, Weihan Weng, Yoey Zhang

March 12, 2024

1. Executive Summary
    a. Overview: This report presents a comprehensive analysis of obesity levels across Mexico, Peru, and Colombia by examining a range of factors including dietary habits, physical activity, lifestyle choices, and demographic characteristics. The research leveraged predictive modeling techniques to identify individuals at high risk of obesity. Three distinct models were developed: Linear Regression to predict BMI, Ordinal Logistic Regression to classify obesity levels, and CART (Classification and Regression Tree) model for categorical classification of obesity levels.
    b. Main Findings:
        i. Linear Regression showed limited success, performing best in predicting Obesity Type I but with an overall accuracy of 37%.
        ii. Ordinal Logistic Regression demonstrated an improved performance with an accuracy of 47%, effectively predicting Insufficient Weight and higher obesity categories.
        iii. The CART model yielded the highest accuracy at 78%, showing effectiveness across most obesity categories, particularly in the higher obesity levels.
    c. Methodology: The analysis utilized a dataset from Kaggle with 2111 records and 16 variables covering a variety of health factors, eating habits, and physical conditions. Descriptive statistics and variable importance were explored, with models evaluated based on accuracy, precision, and recall.
    d. Conclusions: Among the models evaluated, the CART model was the most successful, indicating its suitability for this type of classification problem. Key variables influencing obesity levels included age, height, vegetable consumption frequency, number of main meals, and time spent using technology devices.
    e. Significance: The study underscores the global health concern posed by obesity and its impact on individual well-being. Understanding the contributing factors to obesity is critical for healthcare professionals and policymakers to formulate effective prevention and management strategies. The ability to predict obesity levels can lead to targeted interventions and informed decisions, potentially reducing the prevalence and impact of obesity in populations. The decision to exclude 'Weight' from the key variables, in favor of 'Height', is justified by the relative stability of height in adults and its less frequent fluctuations compared to weight, providing a more consistent predictor for the models employed.

2. Introduction
    a. Project Background
        i. Obesity is a pressing public health concern globally, with its prevalence steadily increasing across various regions. In response to this growing

issue, our project aims to conduct a comprehensive analysis of factors influencing obesity levels across Mexico, Peru, and Colombia. By delving into dietary habits, physical activities, lifestyle choices, and demographic factors, we seek to gain a deeper understanding of the multifaceted nature of obesity in these countries. Our research endeavor involves the construction of predictive models designed to identify individuals at high risk of obesity within the aforementioned regions. By leveraging advanced data analytics techniques, we aim to develop insights that can inform targeted interventions and policy initiatives aimed at mitigating the prevalence of obesity and its associated health risks.

b. Research Purpose and Questions: The primary objective of our research is to investigate the multifaceted determinants of obesity levels across Mexico, Peru, and Colombia. To achieve this goal, we have formulated the following research questions:
    i. Can we accurately predict obesity levels based on dietary habits, physical activity, lifestyle, and demographic factors?
    ii. What are the key factors influencing obesity levels?

    These questions were propelled by the necessity to understand the underpinnings of obesity to aid in its prevention and management. The study is significant as it directly addresses the needs of healthcare providers and policymakers who are the primary audience. By identifying high-risk individuals and understanding the dynamics of obesity, these stakeholders can devise targeted health interventions, influence policy changes, and allocate resources more efficiently to where they are most needed.

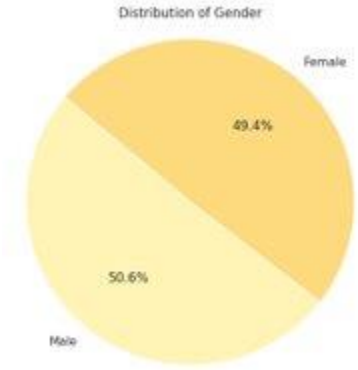
3. Dataset Overview
    a. Detailed Description: The dataset employed for this study consists of 2111 individual records, was sourced from Kaggle and collected by a survey conducted via a web-based platform. The data was meticulously curated to ensure no missing values are present, allowing for a more robust analysis. The dataset is composed of 16 variables that span across demographic information, health factors, eating habits, physical condition, and lifestyle choices.
    b. Data Scope and Timespan:
        i. Scope
            1. Geographical scope: our dataset encompasses Mexico, Peru, and Colombia, three diverse countries within the Latin American region. By focusing on these countries, we aim to capture variations in dietary habits, physical activity levels, lifestyle

choices, demographic characteristics, and obesity prevalence across different socio-cultural contexts.

2. Thematic scope: our dataset primarily revolves around health conditions, with a specific emphasis on obesity. However, it also includes variables related to dietary habits, physical activity, lifestyle factors, and demographic characteristics.

ii. Timespan: The data is assumed to represent a snapshot at the time of data collection. While the specific timespan of data collection is not explicitly provided, we can infer that it reflects a relatively recent timeframe, given the relevance of the data for addressing contemporary public health concerns.

4. Data Exploration and Methodology
   a. Variable Details: Total of 16 variables provide a comprehensive overview of demographics, health factors, eating habits, and physical condition, enabling a thorough exploration of factors influencing obesity levels.
      i. Binary
         1. Gender: Indicates the gender of the respondent.



Genders are evenly distributed.

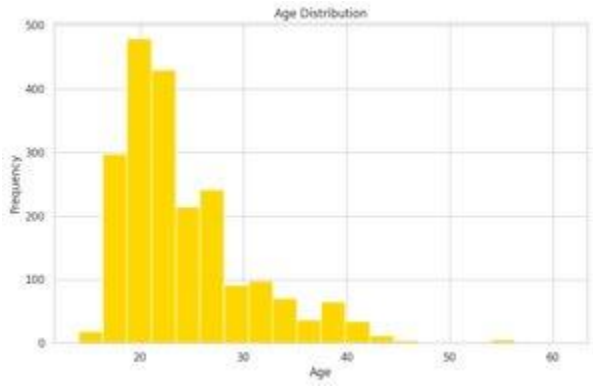
         2. Family history: Indicates whether the respondent has a family history of obesity or related health conditions.
         3. FAVC (Frequent consumption of high-caloric food): Indicates the frequency of consuming high-caloric foods.
         4. SMOKE: Indicates whether the respondent is a smoker.
         5. SCC (Calories consumption monitoring): Indicates whether the respondent monitors their calorie consumption.

      ii. Categorical
         1. CALC (Consumption of alcohol): Indicates the frequency of alcohol consumption.

2. MTRANS (Transportation mode used): Indicates the mode of transportation typically used by the respondent.
3. Obesity: Indicates the obesity status of the respondent.

iii. Numeric

1. Age: Represents the age of the respondent. The dataset predominantly comprises younger individuals, possibly due to the online survey method.



Age ranges from 14 to 61, with a mean age of 24.3.

2. Height: Represents the height of the respondent in centimeters.
3. Weight: Represents the weight of the respondent in kilograms.
4. FCVC (Frequency of vegetables consumption): Indicates the frequency of consuming vegetables.
5. NCP (Number of main meals): Indicates the number of main meals consumed per day.
6. CH2O (Consumption of water daily): Represents the daily water intake of the respondent.
7. FAF (Physical activity frequency): Indicates the frequency of engaging in physical activity.
8. TUE (Time using technology devices): Represents the time spent using technology devices in hours.

b. Methodological Framework

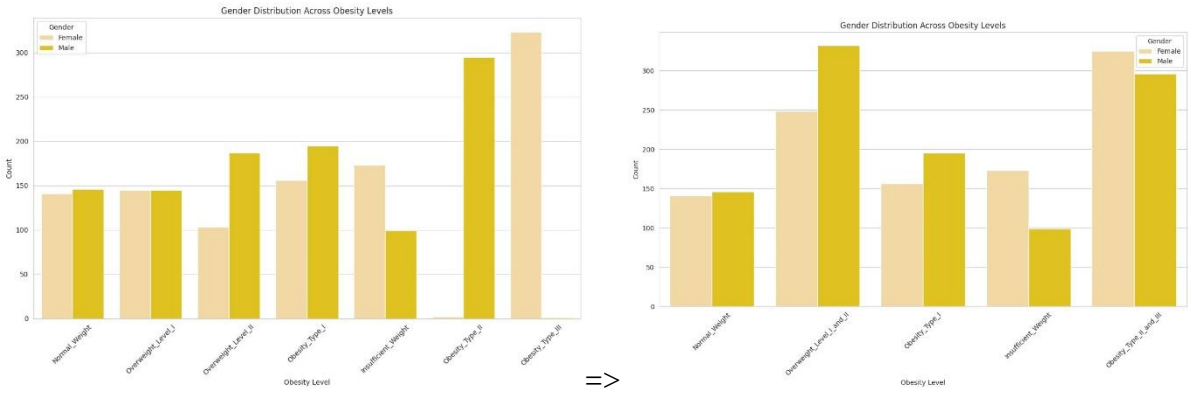
Height and Weight show a strong positive correlation, which is expected as taller individuals generally weigh more.

Either Height or Weight can be dropped from our key variables, and we decided to remove Weight. Height tends to remain relatively constant for adults, while weight can fluctuate due to various factors. (*We also tested about dropping Heigh & Weight both or just dropping Weight, only dropping Weight will get the models better accuracy.)*

	Age	Height	Weight	FCVC	NCP	CH2O	FAF	TUE
Age	1.000000	-0.025958	0.202560	0.016291	-0.043944	-0.045304	-0.144938	-0.296931
Height	-0.025958	1.000000	0.463136	-0.038121	0.243672	0.213376	0.294709	0.051912
Weight	0.202560	0.463136	1.000000	0.216125	0.107469	0.200575	-0.051436	-0.071561
FCVC	0.016291	-0.038121	0.216125	1.000000	0.042216	0.068461	0.019939	-0.101135
NCP	-0.043944	0.243672	0.107469	0.042216	1.000000	0.057088	0.129504	0.036326
CH2O	-0.045304	0.213376	0.200575	0.068461	0.057088	1.000000	0.167236	0.011965
FAF	-0.144938	0.294709	-0.051436	0.019939	0.129504	0.167236	1.000000	0.058562
TUE	-0.296931	0.051912	-0.071561	-0.101135	0.036326	0.011965	0.058562	1.000000

Splitting the gender categories for obesity distribution reveals a higher prevalence of males in other obesity categories, with noticeable gender bias observed in obesity types II and III.

The data suggests a higher count of females in lower obesity levels. Combining obesity type II and III into one category, as well as grouping overweight categories together, would reduce model complexity.



i. Linear Regression
   1. In our analysis, we use a Linear Model to predict the BMI of a person by the given dependent variables. The BMI is calculated using the Height and Weight parameters in the dataset and is set as the target variable. We then removed the Height and Weight

variables to remove direct correlation between the dependent and the independent variables.

2. We have also run a Linear model without removing Height parameter from the dataset as we were getting a very low accuracy score for the model above. This method was implemented with the thought that with only one variable the BMI might not be calculated, and we did see only a slight improvement in the model performance.

ii. Ordinal Logistic Regression

We utilize Ordinal Logistic Regression as our second approach in order to predict obesity levels categorized ordinally. We chose this methodology in conjunction with Linear Regression to effectively address the ordinal nature of the target variable and comprehensively capture the multifaceted factors influencing obesity.

The target variable in our study represents different levels of obesity, spanning from Insufficient Weight, Normal Weight, Overweight levels 1 and 2, to Obesity levels 1, 2, and 3. This ordinal structure aligns seamlessly with the capabilities of Ordinal Logistic Regression. By preserving the ordinal nature of the outcome variable, this modeling technique allows us to maintain the meaningful order and magnitude of obesity levels within our predictive framework.

iii. CART Model

1. Splitting Data

The target variable y is the obesity level we aim to predict, while X includes variables that might influence obesity. Excluding weight and transportation mode from the features ensures that the model is not biased by direct indicators of obesity (like weight) or potentially unrelated variables, focusing on more nuanced relationships. Also, splitting the data into training and test sets and feature scaling are fundamental steps in preparing for machine learning tasks, including obesity level classification and prediction. The 80/20 split ensures substantial learning from training data and unbiased evaluation on unseen test data, balancing the model's ability to learn and generalize.

2. Decision Tree Classifier

   Hyperparameter tuning using GridSearchCV combined with DecisionTreeClassifier can best solve the classification and prediction of obesity levels. GridSearchCV carefully searches for combinations of hyperparameters such as max_depth, min_samples_split, and min_samples_leaf to fine-tune the model to strike a balance between underfitting and overfitting, thus enhancing its generalization ability. DecisionTreeClassifier is known for its interpretability and ability to manage numerical and categorical data without feature scaling (although scaling can improve efficiency), so the choice of DecisionTreeClassifier makes it an ideal tool for this task.

3. Evaluation on Test Set

   Finally, evaluating the best model on the test set provides an estimate of how the model performs on unseen data, which is critical to understanding its real-world applicability. This step confirms the model's ability to generalize beyond the training data.

c. Model Specification: Include detailed model specifications, explaining variable selection, model assumptions, and any validation procedures.
   i. Linear Regression
      1. Model 1:
         a. Variable Selection:

            As per our research in data exploration, we calculated the BMI parameter using both the Height and Weight parameters in the dataset. This was then set as the target variable and the linear model was run to predict it. All the other variables except Height and weight were included in the independent variables section and used to predict the target variable. As BMI can be mapped to Obesity levels, we attempted to use a linear model to gain better accuracy predicting a numerical target variable (BMI) and then check the classification of the actual target variable (Obesity).

         b. Model Assumptions:

            Linearity: The relationship between the independent variables (predictors) and the dependent variable (response) should be

linear. This means that the change in the response variable is proportional to the change in the predictor variables.

No Multicollinearity: The predictor variables should not be highly correlated with each other. Multicollinearity can lead to unstable estimates of the regression coefficients and makes it difficult to interpret the effects of individual predictors.

c. Model Validation:

For model validation, we employed a train-test split approach and divided the data with a test set of 20% of the whole dataset. We then trained the linear model on the train set and used this model to predict the BMI on the test set. These values for BMI were mapped to their respective Obesity levels and then these predicted labels were compared to the true labels for accuracy.

2. Model 2:

We had similar variable selection, methodology, assumptions, and evaluation criteria for this model as well. The only difference was that in the variable selection section, we included the Height parameter in the model to improve the accuracy of the model.

ii. Ordinal Logistic Regression

1. Variable Selection:

As per our research in data exploration, we decided to include all 15 variables available in the dataset, excluding Weight. These variables encompass various factors that are relevant to obesity outcomes.

2. Model Assumptions:

- Ordered Dependent Variable: The dependent variable should be ordered or ranked. In other words, the categories of the dependent variable should have a meaningful order, representing different levels of outcome.
- No Multicollinearity: There should be no multicollinearity among the independent variables.
- Proportional odds: Proportional odds assumption states that the relationship between the independent variables and the dependent variable is consistent across all levels or categories of the outcome variable. In other words, the

effect of the predictors on the odds of being in a higher category versus a lower category of the outcome variable is constant across all levels of the outcome variable.

3. Model Validation:

   To validate our model, we employed a stratified train-test split approach to ensure the balance of each class in the test set. This procedure involves partitioning the dataset into training and testing subsets while preserving the distribution of the ordinal target variable in both subsets.

   Furthermore, we utilized the BFGS optimization algorithm within the Ordinal Logistic Regression framework. This algorithm allows us to minimize the loss function associated with the model, seeking to optimize the fit of the model to the training data.

iii. CART Model

The DecisionTreeClassifier was chosen because it has minimal reliance on data distribution assumptions, unlike its parametric counterpart which presupposes specific relationships between features and outcomes.

1. variable selection: Variables were selected from a broad range of demographics, lifestyle habits, and health indicators, intentionally excluding direct obesity indicators such as weight. (The accuracy will drop by 2% if we exclude both Weight and Height, so we only dropped Weight.)
2. model assumptions: The model assumes that individual independent variables are nonlinear (which is also satisfied by previous data exploration section), which provides the model with the ability to explore the data in a hierarchical manner, allows for the integration of numerical and categorical data, and facilitates the exploration of multifaceted variable interactions.
3. validation procedures: Central to the model refinement are its hyperparameters—max_depth, min_samples_split, and min_samples_leaf—that were optimized through diligent application of GridSearchCV. This process, based on cross-validation, helps identify balanced and optimal model configurations, deftly navigating the boundaries between complexity and generalizability, which is crucial to avoid overfitting and underfitting. The validation strategy involves splitting the dataset into training and test sets, coupled with cross-

validation in hyperparameter tuning, ensuring a robust evaluation framework. This allows the model's performance to be evaluated against unseen data, thereby validating its predictive power and generalizability

.

5. Analysis and Findings
    a. Question-by-Question Analysis:

Research Question 1: "*Can we accurately predict obesity levels based on dietary habits, physical activity, lifestyle, and demographic factors?*"

Methodology Application: The Linear Regression model was utilized to predict BMI, a continuous surrogate marker for obesity, based on the specified factors. The variable selection was informed by exploratory data analysis and prior knowledge of obesity's determinants.

Model Details: BMI was calculated using height and weight from the dataset. The model was then trained using demographics and health-related factors as independent variables.

Assumption Checks: Standard assumptions for linear regression were verified, including the linearity of the relationship between independent and dependent variables, independence of errors, homoscedasticity, and normal distribution of error terms.

Research Question 2: "*What are the key factors influencing obesity levels?*"

Methodology Application: Ordinal Logistic Regression and CART models were employed to classify discrete obesity levels, respecting the ordinal nature of the outcome.

Model Details: Both models utilized the same set of independent variables minus 'weight' due to its high correlation with 'height'. The CART model was fine-tuned using grid search to optimize tree depth and minimum sample splits and leaves.

Assumption Checks: For the Ordinal Logistic Regression, the proportional odds assumption was checked. CART models do not rely on many of the parametric assumptions that regression analyses do, but overfitting and underfitting were checked through cross-validation.

b. Results Interpretation:

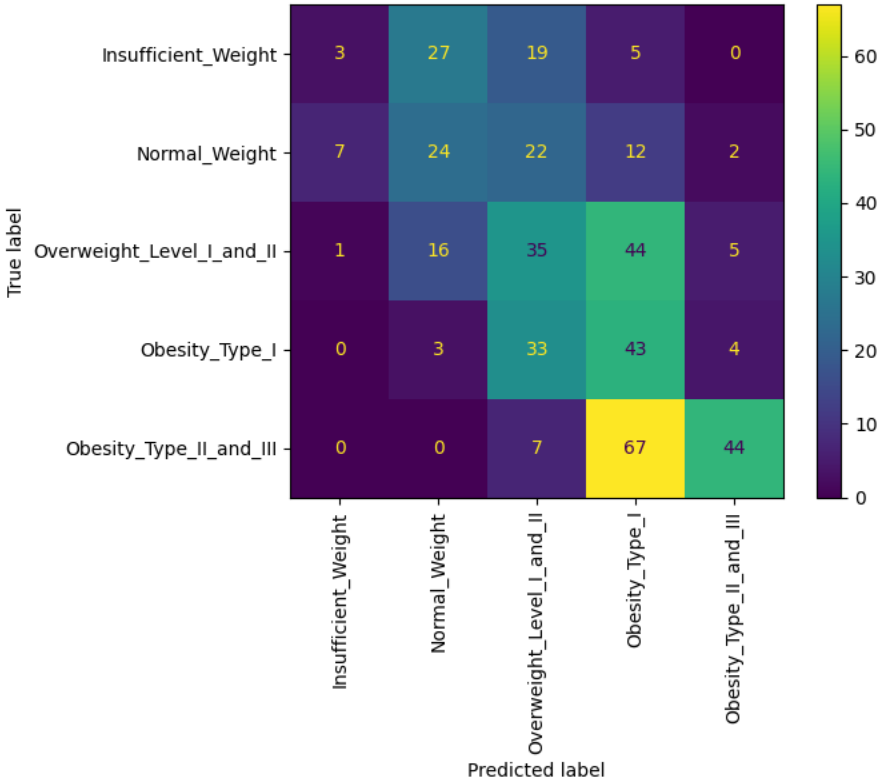
Linear Regression

Model 1 (Height and Weight excluded):

This Linear model results in a model accuracy of 35.2% with weighted precision and recall scores of 43% and 35% respectively. The overall accuracy suggests moderate predictive performance.

The confusion matrix compares the actual BMI classifications (true labels) on the right to the model's predicted BMI classifications. For instance, in the top left corner, the cell value 3 shows that out of 60 people who were underweight (according to their true label), the model predicted 3 of them correctly. The model also incorrectly classified 27 underweight people as normal weight, 19 as overweight level I or II, and 11 as obese (types I, II, or III).

Looking at the diagonal cells, which represent the number of correctly predicted labels, we can see that the model performed best at classifying people with obesity type II and III (7 out of 7) and people with normal weight (24 out of 50). The model struggled the most with classifying people who are overweight level I or II, and underweight people.



The coefficients below provide insights into the relationship between various predictor variables and obesity levels. Among the most influential factors are Family History with Overweight, Frequency Consumption of High-Calorie Food,

and Consumption of Food Between Meals, each possessing an absolute coefficient greater than 0.5.

Family History with Overweight: This predictor variable exhibits a positive coefficient of approximately 7.73. This suggests that individuals with a family history of obesity are more likely to have higher obesity levels. The positive coefficient implies a direct relationship between family history and obesity risk.

Frequency Consumption of High-Calorie Food: With a coefficient of around 2.21, higher frequency consumption of high-calorie food is associated with an increased likelihood of obesity. The positive coefficient indicates that as the frequency of consuming high-calorie food increases, so does the risk of obesity.

Consumption of Food Between Meals: This predictor variable shows a negative coefficient of approximately -3.76. This negative contribution to obesity risk suggests that avoiding food between meals may reduce the risk of obesity. Individuals who refrain from consuming food between meals are likely to have lower obesity levels.

Other noteworthy factors include Consumption of Alcohol with a coefficient of 2.11, indicating its positive association with obesity risk, and Physical Activity Frequency with a negative coefficient of -1.10, suggesting that higher physical activity frequency is linked to lower obesity levels.

Overall, these coefficients provide valuable insights into the complex interplay between various factors and their impact on obesity levels. Understanding these relationships can inform targeted interventions and public health strategies aimed at obesity prevention and management.

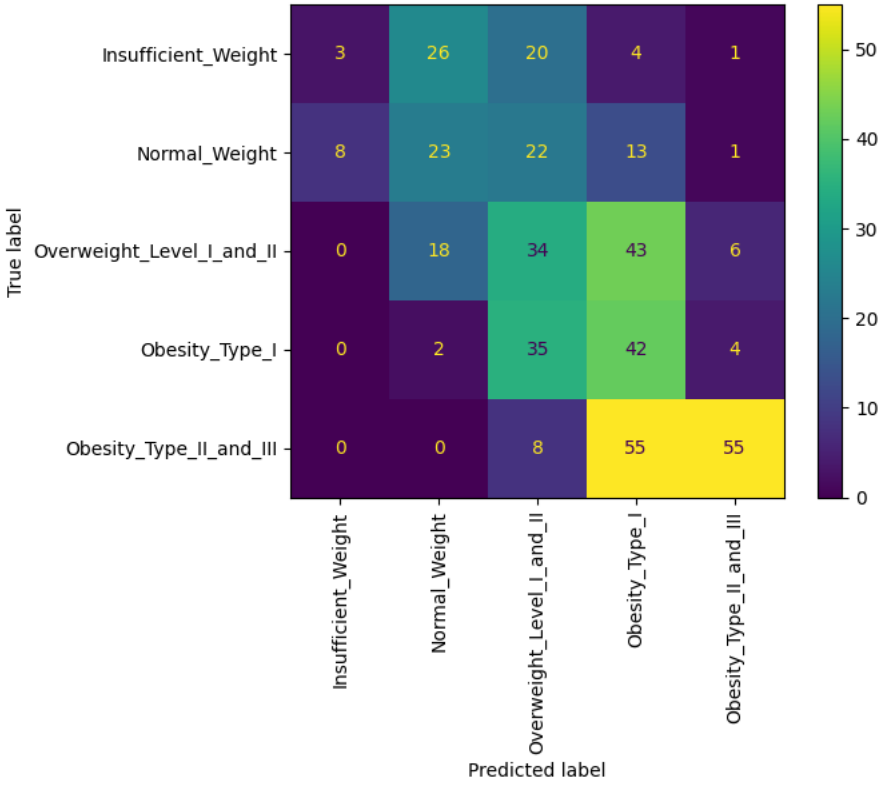
Feature	Coefficient
family_history_with_overweight	7.731409
frequency_consumption_of_vegetables	3.646186
frequency_consumption_of_high_calorie_food	2.214409
consumption_of_alcohol	2.107529
consumption_of_water_daily	0.603189
number_of_main_meals	0.320713
age	0.125633
time_using_technology_devices	-0.363283
smoker	-0.450590
gender	-0.742696
physical_activity_frequency	-1.097800
calorie_consumption_monitoring	-2.717516
consumption_of_food_between_meals	-3.763010

In summary, although the model exhibits moderate overall accuracy, there are evident variations in its predictive performance across different obesity categories.

Model 2 (Weight excluded):

This Linear model results in a model accuracy of 37.11% with weighted precision and recall scores of 43% and 37% respectively. The overall accuracy suggests moderate predictive performance.

As per the confusion matric below, the model performs fairly well for obesity type 1 and Obesity levels 1,2 and 3. But fails to perform well for other types.



The coefficients below provide insights into the relationship between various predictor variables and the target variable (height in this case). Among the most influential factors are Height, Family History with Overweight, and Frequency Consumption of Vegetables, each possessing an absolute coefficient greater than 3.5.

Height: This predictor variable exhibits a significant positive coefficient of approximately 8.54. This suggests that there is a strong positive correlation between height and the target variable. Individuals with greater height tend to have higher values for the target variable.

Family History with Overweight: With a coefficient of around 7.39, family history with overweight also shows a substantial positive association with the target variable. This implies that individuals with a family history of overweight are more likely to have higher values for the target variable.

Frequency Consumption of Vegetables: This predictor variable displays a positive coefficient of approximately 3.52. It indicates that higher frequency consumption of vegetables is associated with higher values for the target variable. This suggests a positive impact of vegetable consumption on the target variable.

Other noteworthy factors include Consumption of Alcohol with a coefficient of 1.92, indicating its positive association with the target variable, and Physical Activity Frequency with a negative coefficient of -1.28, suggesting that higher physical activity frequency is linked to lower values for the target variable.

It's interesting to note that Consumption of Food Between Meals has the most negative coefficient (-3.83), indicating a strong negative impact on the target variable. This suggests that avoiding food between meals may lead to higher values for the target variable.

Overall, these coefficients provide valuable insights into the complex relationship between various factors and the target variable.

Feature	Coefficient
height	8.544006
family_history_with_overweight	7.390981
frequency_consumption_of_vegetables	3.523110
consumption_of_alcohol	1.918285
frequency_consumption_of_high_calorie_food	1.883182
consumption_of_water_daily	0.529462
number_of_main_meals	0.171334
age	0.132913
time_using_technology_devices	-0.383855
smoker	-0.676684
physical_activity_frequency	-1.275340
gender	-1.635551
calorie_consumption_monitoring	-2.463016
consumption_of_food_between_meals	-3.832921

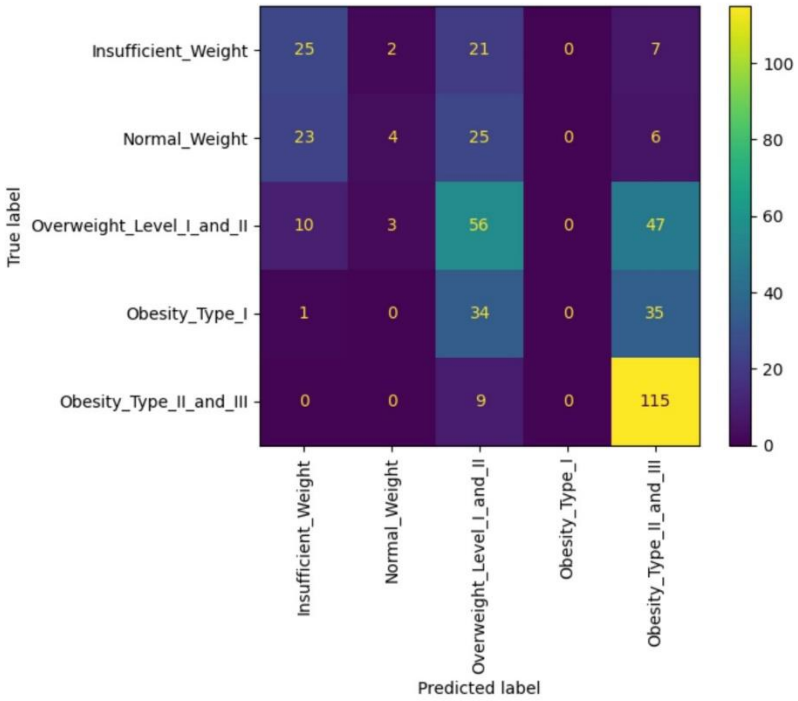
Ordinal Logistic Regression

The Ordinal Logistic Regression results in a model accuracy of 47%, with weighted precision and recall scores of 38% and 47%, respectively. The overall accuracy suggests moderate predictive performance.

The confusion matrix below shows the predictive capabilities of Ordinal Logistic Regression model across different obesity levels. The results for Insufficient Weight, Obesity Type II and III, and Overweight Level I and II demonstrate

relatively robust performance, with F1-scores ranging from 0.43 to 0.69. However, the model's performance for Normal Weight and Obesity Type I is poorer, with F1-scores of 0.12 and 0.00, respectively. This indicates challenges in accurately predicting individuals within these categories.

In summary, although the model exhibits moderate overall accuracy, there are evident variations in its predictive performance across different obesity categories. To explore potential improvements in predictive performance, we have decided to implement the CART model to assess whether it can provide a more enhanced predictive capability.



Below table shows the coefficients that provide insights into the relationship between various predictor variables and obesity levels. Among the most influential factors are *Family History with Overweight*, *Frequency Consumption of High-Calorie Food,* and *Consumption of Food Between Meals*, each possessing an absolute coefficient greater than 0.5.

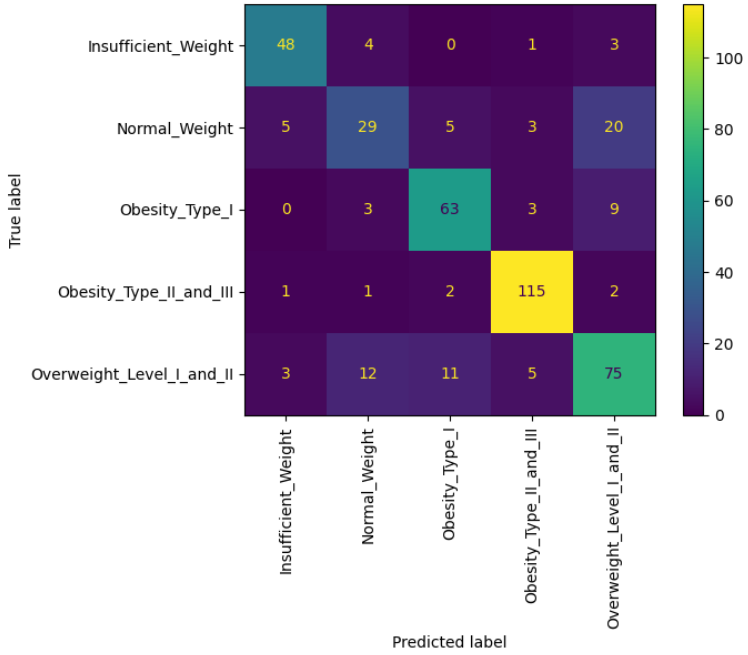
A positive coefficient for Family History with Overweight suggests that individuals with a family history of obesity are more likely to have higher obesity levels. Similarly, higher frequency consumption of high-calorie food is associated with an increased likelihood of obesity, as indicated by their positive coefficients. In contrast, Consumption of Food Between Meals demonstrates a clear negative contribution to obesity risk, suggesting that avoiding food between meals may

reduce the risk of obesity. Overall, these coefficients provide valuable insights into the complex interplay between various factors and their impact on obesity levels, informing targeted interventions and public health strategies aimed at obesity prevention and management.

Variable	Coefficient	Interpretation
Family History with Overweight	2.2888152	Individuals with a family history of overweight are more likely to have higher obesity levels.
Frequency Consumption of High-Calorie Food	0.9303970	Higher frequency consumption of high-calorie food is associated with an increased likelihood of obesity.
Frequency Consumption of Vegetables	0.4503549	Higher frequency consumption of vegetables is associated with an increase likelihood of obesity.
Consumption of Alcohol	0.3324229	Consumption of alcohol is positively associated with obesity risk.
Age	0.3122281	Older individuals tend to have higher obesity levels.
Height	0.1833093	Taller individuals are more likely to have higher obesity levels.
Smoker	0.1624738	Smokers may have a higher risk of obesity.
Gender	0.0955089	Gender does not have a significant impact on obesity levels.
Consumption of Water Daily	0.0647191	Daily water consumption has a minor positive association with obesity risk.
Number of Main Meals	0.0483555	The number of main meals per day has a low impact on obesity risk.
Time Using Technology Devices	-0.140226	Spending more time on technology devices is associated with a slight decrease in obesity risk.
Physical Activity Frequency	-0.328183	Higher frequency of physical activity is associated with a lower risk of obesity.
Calorie Consumption Monitoring	-0.475757	Monitoring calorie consumption is associated with a decreased likelihood of obesity.
Consumption of Food Between Meals	-1.261014	Consuming food between meals is strongly associated with a decreased risk of obesity.

CART Model

In the confusion matrix, the diagonal cells represent correct classifications by the model, with the highest number of correct predictions occurring for 'Obesity_Type_II_and_III'. This suggests that the model is particularly effective at identifying individuals with higher obesity levels. However, the model appears to have challenges in correctly classifying 'Normal_Weight' and 'Overweight_Level_I_and_II', as evidenced by the lower values in those cells and the spread of values across their rows and columns.



For insufficient weight, the precision is high at 0.84, indicating a strong ability to correctly label cases as insufficient weight when they are indeed such. However, the model exhibits challenges with normal weight, reflected in a precision of 0.59 and recall of 0.47; it struggles to identify normal weight cases correctly and often misclassifies them as another category. Obesity type I shows better predictive performance with a precision of 0.78 and recall of 0.81, suggesting the model can reliably identify this class. The standout performance is for obesity type II and III, with an impressive precision of 0.91 and recall of 0.95, indicating a very high likelihood that predictions for this class are correct and most actual cases are captured. Overweight level I and II fall in the middle ground, with a precision of 0.69 and recall of 0.71, reflecting moderate reliability.

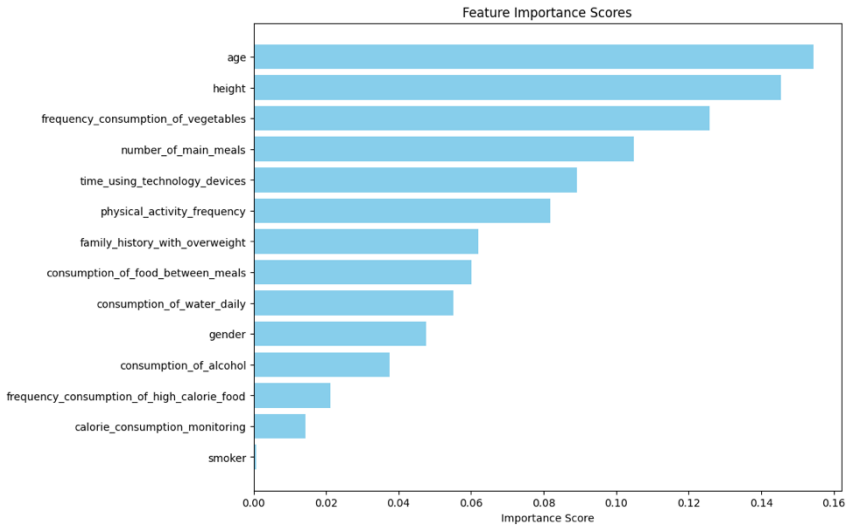
	Precision	Recall	F1-score	Support
Insufficient_Weight	0.84	0.86	0.85	56
Normal_Weight	0.59	0.47	0.52	62
Obesity_Type_I	0.78	0.81	0.79	78

Obesity_Type_II_and_III	0.91	0.95	0.93	121
Overweight_Level_I_and_II	0.69	0.71	0.70	106
accuracy			0.78	423
macro avg	0.76	0.76	0.76	423
weighted avg	0.77	0.78	0.78	423

The overall accuracy is 0.78, demonstrating that the model correctly predicts the obesity level in approximately 78% of cases. Despite this, the 'macro avg' and 'weighted avg' values for precision, recall, and F1-score—which consider the imbalanced class distribution—suggest that the model's predictive accuracy is not uniform across all classes. The high performance in obesity type II and III could be overshadowing weaker performance in other categories, especially normal weight, which is pivotal to address in a practical health context.

6. Conclusion
   a. Summary of Findings: Concisely summarize the project's key outcomes and their implications.

The CART model's analysis reveals age as the primary predictor of obesity, affirming the influence of demographic elements. Following age, height and vegetable consumption frequency emerge as critical factors, signifying the impact of physical traits and dietary habits on obesity risk. The model also underscores the significance of meal frequency and technology usage duration, highlighting the role of eating patterns and sedentary behavior in obesity prevalence.



Further findings reveal physical activity frequency and family history as notable factors, illustrating the complex and multifaceted nature of obesity. In contrast, smoking status, calorie consumption monitoring, and frequency of high-calorie food intake appeared less influential, suggesting they are comparatively minor factors in obesity prediction within this dataset's context and model configuration.

To comprehend how variations in these important variables correlate with obesity trends—whether increasing or decreasing—we should delve deeper into the CART model's splits and structure (chart in the appendix). This examination would unveil the specific variable values linked to higher or lower obesity rates, thereby refining our understanding of the causative factors in obesity dynamics.

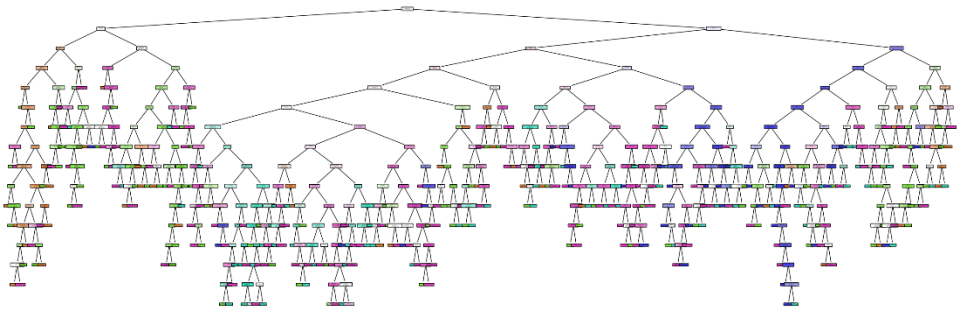
7. Appendices
    a. Technical Details: Include any in-depth technical details, additional data, or complex methodological explanations here.

    When dropping both Weight and Height, the precision will drop to 76%, around 2% decrease from 78%. So we determine to keep Height for CART model.

    CART Model Validation of Dropping both Weight and Height:

	precision	recall	f1-score	support
Insufficient_Weight	0.73	0.84	0.78	56
Normal_Weight	0.55	0.58	0.57	62
Obesity_Type_I	0.76	0.68	0.72	78
Obesity_Type_II_and_III	0.90	0.94	0.92	121
Overweight_Level_I_and_II	0.74	0.68	0.71	106
accuracy			0.76	423
macro avg	0.74	0.74	0.74	423
weighted avg	0.76	0.76	0.76	423

    b. Additional Visualizations: Provide any supplementary charts or tables that support the report but are too detailed for the main sections.



8. References
    a. Citation of Sources:

Appendix

CART

The codes refers to:
https://colab.research.google.com/drive/1y2iLK_12TEPiNK1_H_DIPcDKJvf1l0dc?usp=sharing