Estimating Obesity Levels Based on Eating Habits & Physical Condition

Group 7 - Vibha Gokhale, Khushwant Khatri, Maria Cristina Moreno Siguenza,
Helly Harshad Shah, Isha Thakkar, Khan Yunus

Introduction and Background

Obesity is a growing global public health concern, with rates tripling since 1975 according to the World Health Organization (WHO). Over 1.9 billion adults are classified as overweight, with more than 650 million falling into the obese category. Alarmingly, 39 million children under five were overweight or obese in 2020. This growing epidemic calls for urgent attention to prevent and manage obesity-related diseases. Predictive modeling can help in the early identification of obesity risk factors, enabling timely interventions that may reduce the prevalence of obesity and improve overall well-being.

Multiple factors contribute to obesity, including genetics, environmental factors, eating habits, and physical activity. The modern lifestyle, characterized by sedentary behavior and easy access to high-calorie foods, promotes weight gain. Additionally, unhealthy eating habits such as frequent consumption of processed foods and irregular meal patterns contribute to excessive calorie intake. Furthermore, prolonged screen time and limited physical activity elevate the risk of obesity.

Research Motivation and Objective:

This project aims to analyze how demographic factors (age, gender), eating habits, and physical condition influence obesity levels and to develop predictive models for obesity classification. Specifically, we intend to identify the most significant factors influencing obesity levels and determine whether eating habits or physical conditions are more influential. Additionally, we aim to predict an individual's obesity level based on these factors.

Dataset

The dataset used in this study is from the University of California Irvine Machine Learning Repository, titled "Estimation of Obesity Levels Based on Eating Habits and Physical Condition." The dataset, donated in 2019, contains 2111 records from individuals in Mexico, Peru, and Colombia. It includes 17 variables: 16 independent variables (demographics, eating habits, and physical condition) and a target variable representing obesity levels categorized into seven classes. The dataset also contains both synthetic and survey-generated data, with 77% of the data generated using the SMOTE technique to address class imbalance.

Feature and Target Variables

The predictor or feature variables in this study are divided into three categories. The first category, general variables, includes age, gender, height, weight, and family history of obesity, which may influence an individual's likelihood of weight gain. The second category focuses on eating habits, encompassing factors such as the intake of high calorie foods, vegetables consumption, number of daily meals, snacking frequency, and water and alcohol intake. The third category pertains to physical condition, covering aspects like calorie monitoring, exercise frequency, screen time, smoking, and mode of transportation, all of which play a role in weight management.

The target variable, obesity level, is classified into seven groups: Insufficient weight, Normal weight, Overweight I and II, and Obesity I, II, and III. These classifications help analyze weight-related trends and identify key factors contributing to obesity.

Data Summary Statistics

The table below shows the summary statistics of numerical variables. Age, Height, and Weight are continuous numeric variables. The average age is 24.3 years, with a range of 14–61, suggesting the data represents a predominantly young population. The weight varies widely from 39 kg to 173 kg, while height has a more concentrated range of 1.45 meters to 1.98 meters. Several variables like FCVC (vegetable consumption), NCP (number of main meals), CH2O (daily water intake), FAF (physical activity), and TUE (technology usage) appear as numeric but represent ordinal categorical data with limited distinct values.

Figure 1: Summary Table of Numerical Variables

	Age	Height	Weight	FCVC	NCP	CH2O	FAF	TUE
count	2111.00	2111.00	2111.00	2111.00	2111.00	2111.00	2111.00	2111.00
mean	24.31	1.70	86.59	2.42	2.69	2.01	1.01	0.66
std	6.35	0.09	26.19	0.53	0.78	0.61	0.85	0.61
min	14.00	1.45	39.00	1.00	1.00	1.00	0.00	0.00
25%	19.95	1.63	65.47	2.00	2.66	1.58	0.12	0.00
50%	22.78	1.70	83.00	2.39	3.00	2.00	1.00	0.63
75%	26.00	1.77	107.43	3.00	3.00	2.48	1.67	1.00
max	61.00	1.98	173.00	3.00	4.00	3.00	3.00	2.00

The following table shows the summary statistics for categorical variables, such as gender, family history of overweight, smoking, and transportation mode. Gender is evenly distributed in the data (51% Male, 49% Female). Most individuals report a family history of overweight (82%). Public transportation is the dominant mode of transport (75%), and non-smoking is reported by nearly all participants (98%).

Figure 2: Summary Table of Categorical Variables

	Gender	family_history_with_overweight	FAVC	CAEC	SMOKE	SCC	CALC	MTRANS	NObeyesdad
count	2111	2111	2111	2111	2111	2111	2111	2111	2111
unique	2	2	2	4	2	2	4	5	7
top	Male	yes	yes	Sometimes	no	no	Sometimes	Public_Transportation	Obesity_Type_I
freq	1068	1726	1866	1765	2067	2015	1401	1580	351

This figure shows histograms for the continuous variables age, height, and weight. Age is skewed towards the right, with most participants between 18 and 30 years old. Height and weight are more normally distributed. The slightly longer right tail of the weight distribution suggests a minority of participants with higher weights.

Figure 3: Histograms of Continuous Numeric Variables

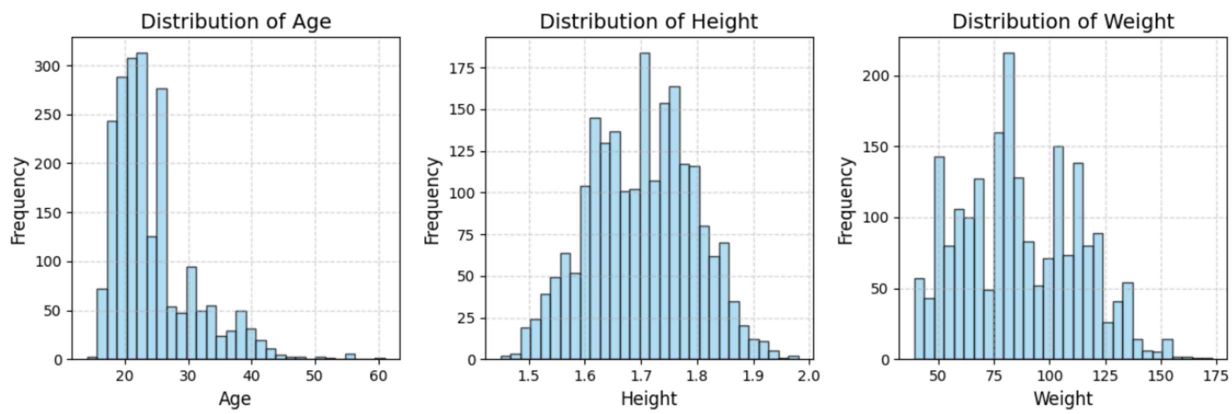
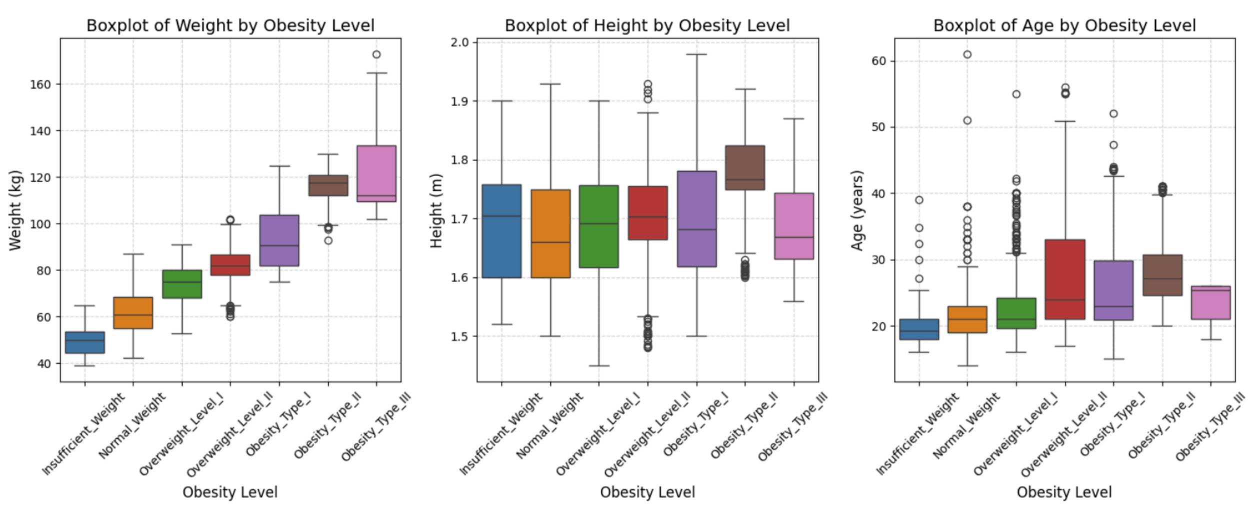


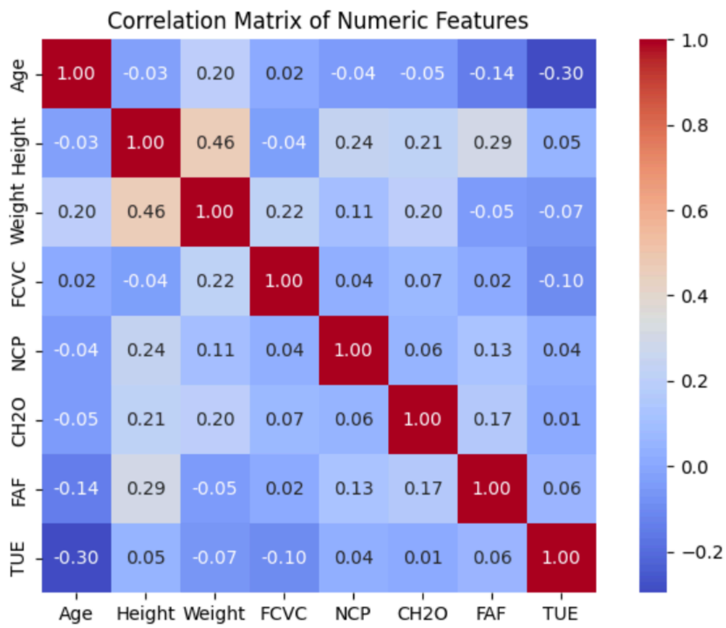
Figure 4 shows the distribution of weight, height, and age across obesity levels. As expected, weight increases steadily with obesity level, with higher variability and more outliers in the upper categories. Individuals in obesity level 3 consistently have the highest weight values, with a higher median and a wider distribution compared to individuals in levels 1 (normal weight) and 2 (overweight). Height remains relatively constant across all groups, suggesting it does not play a major role in distinguishing obesity classes. Age shows a modest increase through the overweight and moderate obesity levels, then slightly declines for the most severe categories, with broader variability and more outliers among older participants.

Figure 4: Boxplots of Continuous Variables by Obesity Levels



The correlation matrix shown in Figure 5 reveals that the strongest relationship among the predictors is between height and weight, with a correlation coefficient of 0.46. This represents a moderately positive correlation, which is expected as taller individuals may tend to weigh more, though this relationship may not be perfectly linear due to factors such as body composition. The other numeric variables in the dataset show relatively weak correlations with each other, indicating that they capture different aspects of the obesity problem.

Figure 5: Correlation Matrix for Numeric Variables



Data Mining Methodology

To analyze the factors influencing obesity levels, we applied various machine learning techniques. First, we used regression analysis by computing Body Mass Index (BMI) based on height and weight, then predicted BMI using data on eating habits and physical activity. For classification, models such as multinomial logistic regression, decision trees, random forest, and support vector machines (SVM) were used to categorize individuals by obesity levels. We assessed the significance of predictors using coefficients, feature importance, and statistical tests, depending on the model. Additionally, clustering methods, including K-Means and hierarchical clustering, helped identify patterns in dietary and physical activity behaviors.

For all models, we first preprocessed the dataset by encoding binary categorical features and applying one-hot encoding to nominal variables. We then standardized the independent variables using the StandardScaler to ensure equal contribution from all features during model training. The data was also split into test and training sets to effectively train and evaluate the models.

The performance of these models was evaluated using relevant metrics, such as R^2 for linear regression, accuracy scores for classification models, and Adjusted Rand Index (ARI) for clustering models.

Results

The following sections discuss the results of each data mining method used to analyze variables influencing obesity levels.

Linear Regression:

To identify key factors influencing obesity, we ran an OLS regression with BMI as the dependent variable.

Figure 6: Linear Regression Output

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                    BMI   R-squared:                       0.500
Model:                            OLS   Adj. R-squared:                  0.495
Method:                 Least Squares   F-statistic:                     99.43
Date:                Mon, 12 May 2025   Prob (F-statistic):          3.00e-295
Time:                        01:07:24   Log-Likelihood:                -6656.2
No. Observations:                2111   AIC:                         1.336e+04
Df Residuals:                    2089   BIC:                         1.348e+04
Df Model:                          21
Covariance Type:            nonrobust
==================================================================================================
                                       coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------------------------
const                                2.8172      5.969      0.472      0.637      -8.889      14.523
Age                                  0.3106      0.027     11.575      0.000       0.258       0.363
FCVC                                 3.3279      0.249     13.384      0.000       2.840       3.816
NCP                                  0.4330      0.166      2.601      0.009       0.107       0.759
CH2O                                 0.6829      0.216      3.159      0.002       0.259       1.107
FAF                                 -0.7898      0.158     -5.009      0.000      -1.099      -0.481
TUE                                 -0.5105      0.221     -2.305      0.021      -0.945      -0.076
Gender_Male                         -0.5069      0.273     -1.857      0.063      -1.042       0.028
family_history_with_overweight_yes   6.7661      0.364     18.563      0.000       6.051       7.481
FAVC_yes                             2.0706      0.418      4.952      0.000       1.251       2.891
SMOKE_yes                           -0.4340      0.886     -0.490      0.624      -2.171       1.303
SCC_yes                             -2.1808      0.625     -3.492      0.000      -3.406      -0.956
CALC_Frequently                     -3.5461      5.814     -0.610      0.542     -14.947       7.855
CALC_Sometimes                      -2.3249      5.782     -0.402      0.688     -13.664       9.015
CALC_no                             -4.6790      5.779     -0.810      0.418     -16.011       6.653
CAEC_Frequently                     -3.4934      0.874     -3.997      0.000      -5.207      -1.780
CAEC_Sometimes                       3.3642      0.809      4.157      0.000       1.777       4.951
CAEC_no                              2.4149      1.157      2.087      0.037       0.146       4.684
MTRANS_Bike                          2.0316      2.191      0.927      0.354      -2.266       6.329
MTRANS_Motorbike                     4.2299      1.761      2.403      0.016       0.777       7.683
MTRANS_Public_Transportation         4.5649      0.392     11.635      0.000       3.795       5.334
MTRANS_Walking                       1.6716      0.866      1.931      0.054      -0.026       3.369
```

- Model R-squared = 0.50, indicating that 50% of the variation in BMI is explained by the variables in the model and RMSE is 5.66

Significant Variables (p < 0.05):

Positive influence on BMI (increase obesity risk):

- Age
- Vegetable Intake (FCVC)
- Number of Meals (NCP)
- Water Intake (CH2O)
- Family History of Overweight
- Fast Food Consumption (FAVC)
- Snacking (CAEC_Sometimes, CAEC_no)
- Use of Public Transport

Negative influence on BMI (reduce obesity risk):

- Physical Activity (FAF)
- Technology Use (TUE)
- Calorie Monitoring (SCC_yes)
- Frequent Snacking (CAEC_Frequently)

Interpretation:

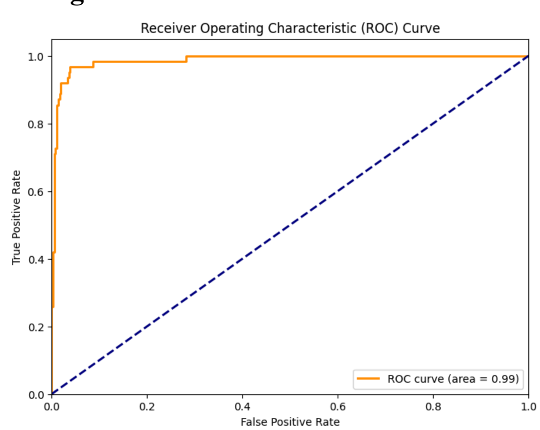
The regression analysis reveals that demographic and behavioral factors significantly affect BMI. Older individuals, those with a family history of being overweight, and those who consume fast food or have more meals per day tend to have a higher BMI, increasing the risk of obesity. On the other hand, people who are more physically active, monitor their calorie intak**e**, or snack frequently tend to have a lower BMI. These insights underline the critical role of lifestyle choices in managing obesity.

<u>**Multinomial Logistic Regression:**</u>

Logistic regression was used to classify obesity levels into 7 categories using all available features. The BMI variable we calculated in the earlier step was excluded along with Height and Weight to avoid redundancy. The logistic regression model's performance was evaluated using cross-validation accuracy, overall accuracy, the classification report, and the confusion matrix.

- Cross-validation Accuracy:
  The model achieved an average accuracy of 61.15% across 5-fold cross-validation.
- Overall Accuracy:
  The model reached an accuracy of 62.88% on the test set.
- Classification Report:
  Precision and recall varied widely across classes, with some (e.g., class 4) predicted very accurately, while others (e.g., classes 1 and 6) had poor performance.
- Confusion Matrix:
  The confusion matrix showed frequent misclassifications among adjacent weight categories.

Figure 7: ROC Curve of Obese Class



The logistic regression model underperformed despite the dataset being artificially balanced across obesity categories. This may be due to the exclusion of Height and Weight, which are not redundant in the data and may contain independent predictive power. As part of exploratory analysis, we reintroduced them into the logistic regression model, which significantly improved its performance, with an updated accuracy score of 86.52%. When Height and Weight were included, most classes had precision, recall, and F1-scores above 0.85 and the confusion matrix showed minimal misclassification.

Decision Trees:

Decision Trees are a widely used machine learning model that make predictions by recursively splitting the data based on the most informative features.

Figure 8: Decision Tree Output

```
Accuracy Score: 0.933806146572104
Classification Report:
                     precision    recall  f1-score   support

Insufficient_Weight       0.92      0.96      0.94        56
      Normal_Weight       0.84      0.87      0.86        62
     Obesity_Type_I       0.96      0.92      0.94        78
    Obesity_Type_II       0.95      0.95      0.95        58
   Obesity_Type_III       1.00      1.00      1.00        63
 Overweight_Level_I       0.91      0.88      0.89        56
Overweight_Level_II       0.96      0.96      0.96        50

           accuracy                           0.93       423
          macro avg       0.93      0.93      0.93       423
       weighted avg       0.93      0.93      0.93       423
```

1. **Accuracy**: The model has an overall accuracy of 93.38%, which indicates that it performs well in predicting the correct class for most of the samples.

2. **Precision, Recall, and F1-Score**:
   - **Precision** refers to the proportion of true positive predictions (correct classifications of a given class) out of all the predicted positives (including false positives). The model is generally good at precision for most classes, especially for "Obesity_Type_III" and "Overweight_Level_II," both with high precision values (1.00).
   - **Recall** measures how well the model identifies all actual positives for each class. Recall values are also strong, with the model achieving near-perfect recall for most classes, particularly "Obesity_Type_III" and "Insufficient_Weight."
   - **F1-Score** is the harmonic mean of precision and recall. It provides a balance between the two and is helpful when considering both false positives and false negatives. The F1-score is high across all classes, indicating good model performance.

3. **Class-Level Performance**:
   - **Obesity_Type_III** has perfect scores across all metrics (precision, recall, and F1-score), indicating that the model correctly classifies every sample in this category.

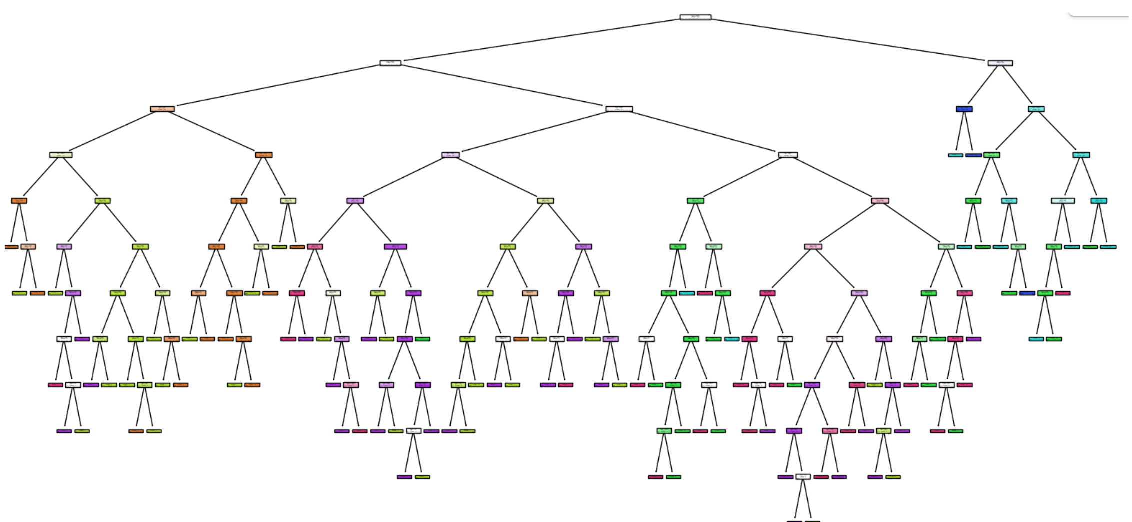
- **Normal_Weight** and **Overweight_Level_I** show slightly lower scores in precision and recall, but they still fall within a reasonable range (above 0.80).
- **Obesity_Type_I** and **Obesity_Type_II** have very good precision and recall, with the F1-scores of 0.94 and 0.95 respectively, indicating that they are well classified.

4. **Balanced Metrics**:
   - The **macro average** and **weighted average** for precision, recall, and F1-score are all 0.93, which indicates that the model is balanced in its performance across all classes, taking both smaller and larger class sizes into account.

Interpretation:

This model demonstrates good overall performance with an accuracy of 93.38%. It effectively distinguishes between obesity categories, with some classes, like **Obesity_Type_III**, performing exceptionally well. The performance is slightly lower for **Normal_Weight** and **Overweight_Level_I**, but still within acceptable ranges. The balanced metrics suggest that the model is robust and handles all classes with a reasonable degree of success.

Figure 9: Decision Tree



<u>Random Forest:</u>

The Random Forest classifier was used to predict obesity levels based on all 16 input features, including demographic, dietary, and physical activity variables. The model was trained on an 80/20 train-test split and evaluated using accuracy and classification metrics.

Key Results:

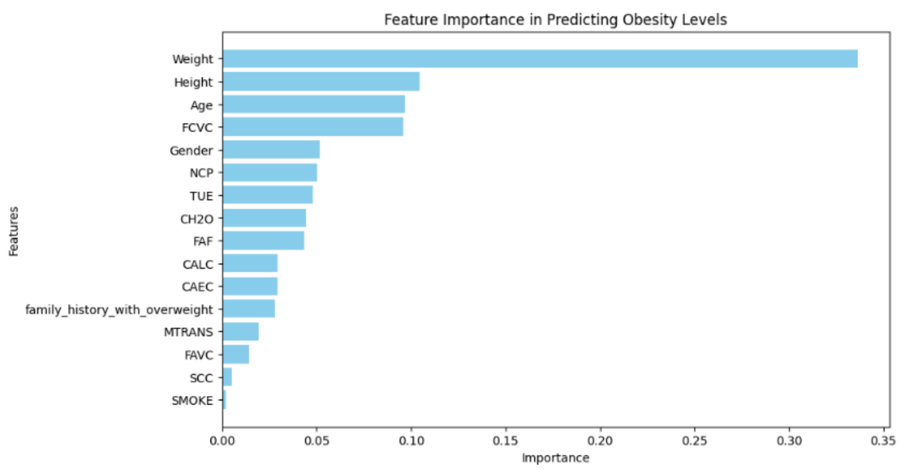
- Accuracy: The model achieved a high accuracy of 95.5% on the test set, indicating excellent performance in predicting obesity categories.

- Precision & Recall: The model demonstrated high precision and recall across most classes, particularly:
    - Obesity_Type_I, II, III: F1-scores close to or at 1.00
    - Insufficient_Weight: F1-score of 0.97
    - Normal_Weight and Overweight_Level_I: Slightly lower scores, with F1-scores around 0.88–0.90
    - Macro F1-score: 0.95 — indicating strong overall balance across all classes.
    - Weighted F1-score: 0.96 — confirming robustness even with imbalanced class sizes.

<u>Feature Importance:</u>

Using the model's built-in feature importance metric, we identified Weight, Height, and Age as the top predictors, followed by lifestyle-related features such as vegetable intake (FCVC) and physical activity (FAF).

Figure 10: Random Forest Feature Importance



<u>Support Vector Machines (SVM):</u>

Our final classification model for predicting obesity based on various features was Support Vector Machine. The SVM model was trained on standardized data and evaluated using two different kernels: a Radial Basis Function (RBF) kernel with default hyperparameters, and a Linear kernel to understand feature importance and improve interpretability. The Support Vector Machine with the RBF kernel achieved an accuracy score of 83.74% on the test set, while the model using the Linear kernel performed significantly better, achieving an accuracy score of 93.85%.

The ten most important features in predicting obesity, based on the magnitude of the absolute coefficients from the linear SVM model, are shown in the table. These results show that Weight and Height are the most influential predictors, with lifestyle factors such as eating habits and mode of transport also contributing significantly. Notably, MTRANS_Automobile (using cars as primary mode of transportation) and TUE (time spent using technology) indicate a potential association with obesity risk, aligning with known patterns of sedentary behavior and unhealthy eating.

Figure 11: Top 10 Features by Linear SVM Coefficient

Top 10 Features by SVM Coefficient

	Feature	Importance
0	Weight	6.663700
1	Height	1.558450
2	CAEC_Always	0.199455
3	Age	0.190714
4	FCVC	0.155934
5	CAEC_Frequently	0.155846
6	CH2O	0.148100
7	TUE	0.141319
8	MTRANS_Automobile	0.111468
9	CALC_Frequently	0.110998

K-Means Clustering

To uncover hidden groupings within the dataset, we applied the K-Means clustering algorithm using numeric features. This unsupervised learning method allowed us to explore natural patterns in individuals' physical and lifestyle characteristics without relying on labeled obesity levels.

Prior to modeling, all numeric variables were standardized using StandardScaler to ensure that features like age, height, and weight contributed equally during clustering. We selected k = 5 clusters based on exploratory testing and consistency with class examples. The model was then trained on the standardized data, and each individual was assigned to one of five clusters.

Figure 12 below visualizes the cluster assignments using two key features - Age (x-axis) and Height (y-axis). Each color represents a distinct cluster identified by the algorithm.

The figure illustrates clear segmentation across the age and height dimensions. For example, one cluster consists primarily of younger individuals with average height, while another spans a wider age range with greater variability in height. Despite not using obesity levels as an input, the clusters reveal potentially meaningful differences in physical profiles, which may correlate with distinct obesity risk levels.

Since K-Means is an unsupervised method, traditional performance metrics like accuracy or F1-score do not apply. Instead, we evaluated model performance based on:

- Cohesion within clusters, indicating that similar individuals were grouped together,
- Separation between clusters in 2D space, suggesting distinct subgroups,

- And interpretability of the cluster distributions, which align with realistic health and lifestyle patterns.

The clustering results provide valuable insight into the underlying structure of the data and demonstrate how unsupervised learning can aid in identifying behavioral and physical trends, even in the absence of labeled outcomes.

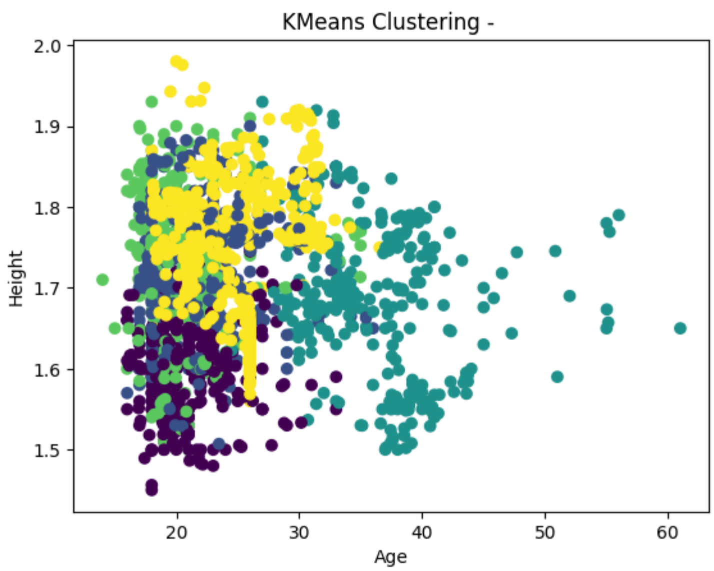


Figure 12: Scatterplot showing K-Means cluster assignments based on Age and Height. Each color represents a different cluster (k = 5).

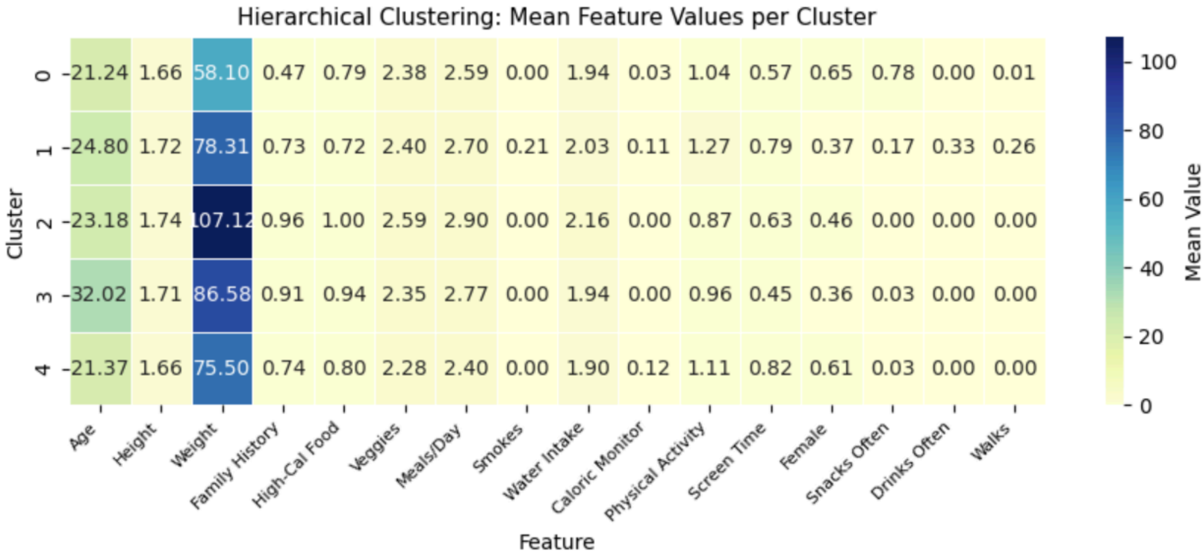
Hierarchical Clustering:

The second unsupervised model we used is hierarchical cluster analysis. We applied agglomerative clustering to the standardized data and identified 5 meaningful clusters. We chose 5 clusters to ensure compatibility with our K-Means model and to allow for a meaningful comparison of ARI scores between the two methods.

These clusters revealed distinct patterns in weight, eating habits, physical activity, and other health-related behaviors, as shown in the heatmap of mean feature values per cluster. For example, Cluster 2 includes individuals with the highest average weight, frequent eating habits, and low physical activity. In contrast, Cluster 0 contains individuals with the lowest average weight, higher physical activity levels, and less screen time.

The ARI score is 0.14, suggesting low similarity between the unsupervised clusters and actual obesity levels. However, the distinct cluster profiles suggest that clustering may help tailor health interventions to specific behavioral profiles.

Figure 13: Mean Feature Values per Cluster



__**Conclusions**__

We compared the performance of all the supervised and unsupervised models, as summarized in the table below. The linear regression model produced an R^2 of 0.50 and an RMSE of 5.66, indicating moderate predictive power. This suggests that while the model captured some variance in obesity levels, it likely missed key nonlinear patterns in the data.

Among the **c**lassification models, Random Forest performed the best with an accuracy of 95.51%, followed closely by SVM and Decision Tree, both achieving accuracy scores above 93%. These models benefited from their ability to handle non-linear relationships and multiple feature interactions.

Logistic Regression showed a noticeably lower test accuracy of 62.88%, largely due to the initial exclusion of key features like height and weight. When these features were reintroduced, accuracy improved significantly to 86.52%, but still lagged behind the performance of the tree-based and SVM models, suggesting that Logistic Regression was less capable of capturing complex class boundaries.

For clustering, we evaluated K-Means and Hierarchical Clustering using the ARI to compare cluster labels to the true obesity levels. K-Means achieved an ARI of 0.191, slightly higher than Hierarchical clustering's ARI of 0.14, indicating that both models captured some structure, but not strongly aligned with the actual categories.

Figure 14: Evaluating Model Performance

Model	Task	Metric	Score
Linear	Regression	R^2, RMSE	0.50 ,5.66
Logistic	Classification	Accuracy	62.88%
Decision Tree	Classification	Accuracy	93.38 %
Random Forest	Classification	Accuracy	95.51 %
SVM (linear kernel)	Classification	Accuracy	93.85 %
K-Means	Clustering	Adjusted Rand Index	0.191
Hierarchical	Clustering	Adjusted Rand Index	0.14

Overall, our best classification performance came from tree-based methods like Random Forest, while linear SVM also demonstrated high accuracy and interpretability. Clustering techniques provided additional insight by uncovering lifestyle-based patterns without relying on labels. Each model contributed unique insights. Together, they offer both predictive accuracy and exploratory value.

Our findings highlight the effectiveness of machine learning models in predicting obesity and uncovering influential health factors. Demographic attributes (age, weight), dietary habits (meal frequency, vegetable intake), and lifestyle patterns (exercise, screen time) all significantly contribute to obesity levels.

Practical Implications

The outcomes of our research could have significant real-world implications:

1. **Health Risk Prediction**: Individuals can receive early warnings about obesity risks, encouraging proactive health measures.
2. **Diet & Lifestyle Recommendations**: Based on our predictive models, we can offer personalized recommendations for healthier eating and activity patterns.
3. **Healthcare & Policy Insights**: Public health organizations can leverage our findings to design more effective obesity management programs and policies.

By translating data insights into actionable recommendations, our project can contribute to both individual and societal health improvements.

References

Centers for Disease Control and Prevention. (2024, May 14). Adult Obesity Facts. CDC. https://www.cdc.gov/obesity/adult-obesity-facts/index.html

Estimation of Obesity Levels Based On Eating Habits and Physical Condition [Dataset]. (2019). UCI Machine Learning Repository. https://doi.org/10.24432/C5H31Z.

Gillette, H. (2024, February 22). Obesity: Is It Genetic of Environmental? Healthline. https://www.healthline.com/health/obesity/is-obesity-genetic-or-environmental

Jie Wei Zhu, Parsa Charkhchi, Shadia Adekunte, & Akbari, M. R. (2023). What Is Known about Breast Cancer in Young Women? Cancers, 15(6), 1917–1917. https://doi.org/10.3390/cancers15061917

Markelle Kelly, Rachel Longjohn, Kolby Nottingham, The UCI Machine Learning Repository, https://archive.ics.uci.edu

World Health Organization. (2024, March 1). Obesity and overweight. World Health Organization. https://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight