Data Mining for Health: Investigating Weight Categories and Contributing Factors

Sagar Bangera[1], Nikhil Chhatraband[1], Sumona Mondal[2], and Naveen Ramachandra Reddy[1]

[1]Department of Data Science, Clarkson University, Potsdam, NY
[2]Co-Director, MS Program in Applied Data Science, Clarkson University, Potsdam, NY

Abstract

This data set focuses on the health and lifestyle characteristics of individual obesity-related parameters, such as age, sex, height, weight, and behavioral markers, including physical activity, water intake, and food. It is a helpful resource for researching the association between lifestyle choices and obesity levels by categorizing people into several groups ranging from insufficient weight to extreme obesity.

1 Introduction

In this study, our objective is to unravel the intricate relationships between obesity and a variety of demographic, behavioral, and genetic factors. The data set used for this analysis provides a comprehensive view of the characteristics that contribute to obesity. It includes key physical attributes such as height and weight, along with lifestyle indicators such as diet habits, water intake, physical activity, and a family history of overweight. These diverse variables offer a holistic perspective on how different factors interact to influence weight changes and obesity risk.

By categorizing individuals into different levels of obesity, ranging from underweight to severe obesity, the data set allows us to examine patterns and associations that may not be immediately apparent. Inclusion of behavioral factors such as the frequency of vegetable consumption, snacking habits, and alcohol intake allows us to delve deeper into how lifestyle choices shape weight outcomes. Similarly, variables such as water consumption and the frequency of physical activity highlight the importance of hydration and exercise in maintaining a healthy weight. The family history component adds another layer of depth, pointing to the genetic predispositions that often underlie obesity.

The overarching goal of this study is to uncover meaningful insights into the connections between obesity categories and lifestyle decisions. To achieve this, we employ a combination of statistical and machine learning techniques. These methods are chosen not only for their ability to identify patterns and associations, but also for their potential to make accurate predictions. By integrating traditional statistical approaches with advanced computational models, this project demonstrates the power of combining data-driven methods to tackle complex real-world problems.

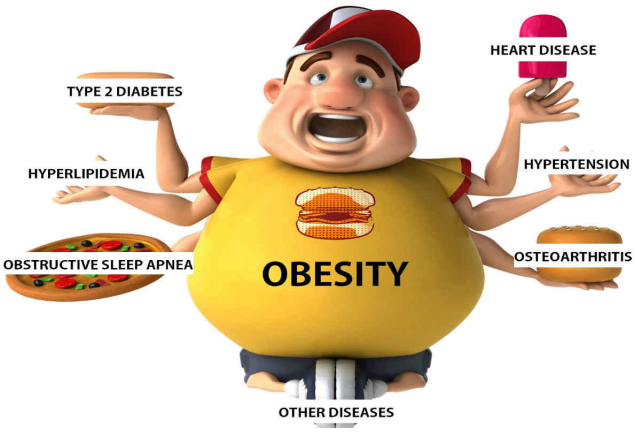

Figure 1: Problems from Obesity

This research is not merely an academic exercise; it has practical implications that extend to public health and policy-making. The findings of this study can inform interventions aimed at promoting healthier lifestyles and preventing obesity. For instance, identifying key behavioral factors associated with obesity can guide the design of targeted educational campaigns and community programs. Similarly, understanding the role of genetic predisposition can pave the way for personalized healthcare approaches that take into account the unique risk factors of an individual.

In addition to its societal relevance, this project showcases the application of data science techniques in addressing health-related challenges. The use of statistical

tests such as ANOVA and Pearson's Chi-squared, combined with machine learning models like Random Forest and Decision Trees, illustrates how these tools can be leveraged to gain a deeper understanding of complex datasets. The results not only provide actionable insights but also highlight the strengths and limitations of different analytical methods.

By shedding light on the multifactorial nature of obesity, this study contributes to a growing body of literature that seeks to understand and address this pressing global issue. It underscores the importance of adopting a multidisciplinary approach, combining insights from data science, public health, and behavioral research to develop effective strategies for combating obesity. Through this work, we aim to bridge the gap between data analysis and real-world applications, demonstrating the value of evidence-based decision-making in improving health outcomes.

2 Background

Data-driven methods are now essential for comprehending complex systems and obtaining useful insights in a variety of fields. In order to investigate relationships, spot patterns, and forecast results, this research focuses on analyzing a dataset that includes demographic, behavioral, and health-related factors. High dimensionality and interdependence are two problems that arise frequently when several variables interact, necessitating the use of stringent statistical and computational methods to guarantee reliable analysis. The dataset depicts real-world situations where a thorough understanding of behaviors and results is offered by both continuous and categorical variables. Gaining insight into these connections aids in answering more general queries, such as how particular actions affect health metrics or how demographics affect results. The research aims to improve interpretability, produce accurate predictions, and reveal hidden structures in the data by utilizing statistical and machine learning approaches. The background information emphasizes how crucial it is to combine computational efficiency and statistical rigor in order to evaluate datasets successfully. Informed decision-making and perceptive conclusions are made possible by this method, which addresses typical issues like noise, multicollinearity, and non-normal distributions while permitting meaningful interpretations.

3 Objective

The primary objective of this research is to leverage a combination of statistical and machine learning techniques to comprehensively analyze and interpret the intricate relationships between demographic, behavioral, and health-related variables in the dataset. Given the multidimensional nature of the data, this study aims to unravel the interplay between various factors that contribute to obesity, a complex and multifaceted global health challenge. The overarching goal is to bridge the gap between raw data and actionable insights, providing a holistic understanding of the factors that influence obesity while showcasing the practical application of advanced analytical methods.

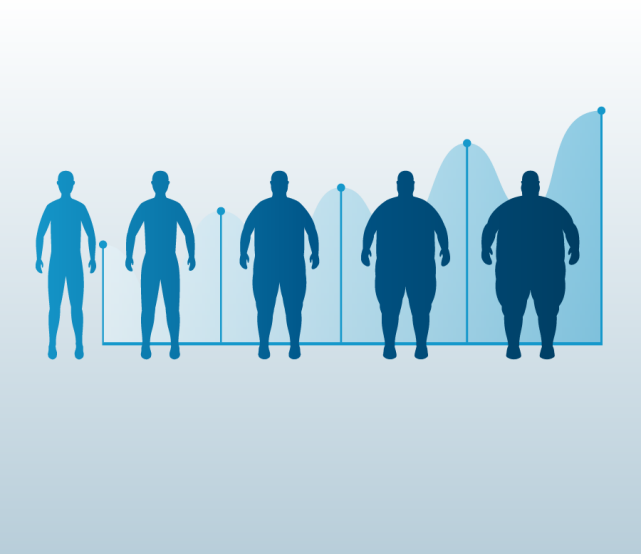

Figure 2: Which category do you lie in?

To achieve this, the study begins by assessing the quality and structure of the data. One of the initial tasks involves checking for normality, multicollinearity, and the variance-covariance structure of the variables. These assessments are crucial for ensuring the reliability of subsequent statistical analyses. For instance, evaluating normality using tests such as the Kolmogorov-Smirnov test helps determine whether parametric or non-parametric methods are more suitable. Similarly, identifying and addressing multicollinearity through variance inflation factor (VIF) analysis ensures that the predictors in regression models are independent enough to produce meaningful and interpretable results.

Exploratory and inferential statistical methods are employed to identify significant patterns and relationships among variables. For example, correlation matrices provide an overview of the linear relationships between continuous variables, while tests like ANOVA and Chi-squared tests reveal how demographic and behavioral factors influence obesity categories. These methods

not only highlight key relationships but also guide the selection of features for predictive modeling, ensuring that the most relevant variables are prioritized.

Machine learning techniques form a central part of this research, with models like Random Forest and Logistic Regression being implemented to classify outcomes and predict key health indicators. These models are particularly well-suited for handling the complexity of the dataset, as they can capture non-linear relationships and interactions between variables. Random Forest, for example, not only provides high classification accuracy but also ranks variables based on their importance, offering valuable insights into which factors have the greatest impact on obesity categories. Logistic Regression, on the other hand, provides a probabilistic framework for understanding the likelihood of different outcomes based on the predictors.

In addition to supervised learning models, unsupervised techniques such as k-means clustering are utilized to uncover hidden structures and groupings within the data. These clustering methods can identify subgroups within the population that share similar characteristics, offering new perspectives on the patterns of behavior and risk factors associated with obesity. By integrating these insights with the findings from supervised models, the study aims to build a comprehensive understanding of the dataset.

An equally important objective is to evaluate the performance and limitations of the applied methods. Every analytical technique, from statistical tests to machine learning models, has its strengths and weaknesses. For instance, while Random Forest is robust against overfitting and can handle complex interactions, it may lack the interpretability of simpler models like Decision Trees. Similarly, while Chi-squared tests are effective for identifying associations between categorical variables, they do not account for potential confounding factors. By critically assessing these methods, the study ensures that the results are not only accurate but also interpretable and practically applicable.

By achieving these objectives, the project seeks to provide actionable insights into the dataset, transforming raw data into meaningful knowledge that can inform decision-making and intervention strategies. Whether it is identifying key behavioral factors for targeted health campaigns or predicting obesity risk based on demographic and lifestyle variables, the findings have the potential to make a tangible impact on public health.

Furthermore, this research demonstrates the effectiveness of combining statistical and computational approaches for real-world data analysis. By integrating traditional statistical techniques with modern machine learning methods, the study highlights how these complementary approaches can be used to tackle complex problems. The results not only contribute to the growing field of data science but also underscore its practical applications in addressing pressing health challenges. Through this comprehensive analysis, the study aims to set a benchmark for future research, showcasing the value of a multidisciplinary approach to understanding and mitigating obesity.

4 Dataset

The dataset utilized in this study serves as a rich resource for analyzing health-related, behavioral, and demographic factors associated with obesity. With 20,758 observations and no missing values, the dataset offers a comprehensive view of the variables that influence health outcomes, particularly obesity. This completeness and diversity of data make it an ideal candidate for employing advanced statistical and machine learning techniques to derive meaningful insights and predictive models.

4.1 Key Characteristics

The dataset comprises a wide range of variables that fall into three broad categories:

- **Demographic Variables:**
  - **Age:** Captures the age of individuals in years, allowing for the assessment of age-related trends in obesity.
  - **Gender:** A categorical variable (male/female) that facilitates comparisons between different sexes.
  - **Height:** A critical factor for calculating body metrics such as Body Mass Index (BMI), enabling the analysis of how stature correlates with weight categories.
- **Behavioral Variables:**
  - **Number of Primary Meals Consumed (NCP):** Indicates the daily number of main meals, which serves as a proxy for dietary patterns and eating habits.
  - **Frequency of Vegetable Consumption (FCVC):** Measures the regularity of vegetable intake, providing insights into healthy dietary behaviors.
  - **Frequency of Physical Activity (FAF):** Captures the weekly frequency of exercise, a crucial factor in maintaining a healthy weight.

  - **Time Spent Using Technology (TUE):** Quantifies the number of hours spent on electronic devices daily, reflecting sedentary behavior.

- **Health-Related Variables:**
  - **Weight:** A critical outcome variable for analyzing obesity and weight categories.
  - **Daily Water Intake (CH2O):** Indicates hydration levels, which may indirectly impact weight and metabolism.
  - **Family History of Overweight:** A categorical variable denoting whether an individual has a genetic predisposition to obesity.

- **Target Variable:**
  - **Obesity Categories (NObeyesdad):** Represents predetermined groupings of individuals into categories such as insufficient weight, normal weight, overweight, and various obesity levels. This categorical variable serves as the target for classification and predictive modeling tasks.

4.2 Data Preprocessing and Encoding

Given the mixture of numerical and categorical data, the dataset underwent extensive preprocessing to ensure compatibility with the analytical methodologies employed:

- **Standardization:** All numerical variables were standardized to eliminate scale disparities, allowing for better performance in statistical tests and machine learning models.
- **Encoding:** Categorical variables, such as food availability, family history of overweight, and transportation mode, were converted into numerical formats using one-hot encoding or label encoding techniques. This transformation ensures that these variables can be seamlessly integrated into predictive models without introducing bias.

4.3 Advantages of the Dataset

The dataset's completeness, with no missing values, provides a robust foundation for statistical and machine learning analyses. Its diverse variables allow for a holistic exploration of the multifaceted nature of obesity, from behavioral habits and dietary patterns to demographic and genetic predispositions. Additionally, the inclusion of both continuous and categorical variables enables the application of a wide range of techniques, from regression and classification to clustering and dimensionality reduction.

This extensive dataset not only facilitates the identification of trends and connections between variables but also provides a valuable framework for building predictive models. By leveraging these data, the study aims to produce actionable insights and forecasts that can inform targeted interventions and policy decisions aimed at addressing the global obesity epidemic.

5 Variable Description

Table 1: Variable Descriptions

Variable Name	Type	Units
id	Numeric	numeric.
Gender	Categorical	Male or Female.
Age	Numeric	Years.
Height	Numeric	Meters.
Weight	Numeric	Kilograms.
family_history	Categorical	yes/no.
FAVC	Categorical	yes/no.
FCVC	Numeric	scale from 1 to 3.
NCP	Numeric	scale from 1 to 4.
CAEC	Categorical	(4 Categories).
SMOKE	Categorical	yes/no.
CH2O	Numeric	scale from 1 to 3.
SCC	Categorical	yes/no.
FAF	Numeric	scale from 0 to 3.
TUE	Numeric	scale from 0 to 2.
CALC	Categorical	(4 Categories).
MTRANS	Categorical	(Automobile, Walking, etc.).
NObeyesdad	Categorical	(7 Categories).

The dataset includes a mix of demographic, behavioral, and health-related variables, as well as a target variable for classification purposes.

6 Process

6.1 Step 1: KS Test for Normality

We conducted the Kolmogorov-Smirnov test to assess whether the `Weight` variable follows a normal distribution. The results are shown in Table 2.

Table 2: Kolmogorov-Smirnov Test Results for Weight

Statistic (D)	0.091128
p-value	< 0.001

The KS test was conducted to assess whether the distribution of the variable Weight conforms to a normal distribution. The results revealed a test statistic of D=0.091128 and a p-value less than 0.001. This led to the rejection of the null hypothesis, indicating that the Weight variable significantly deviates from normality.

Implications of the Non-Normal Distribution:

- **Methodological Adjustments:** Non-parametric tests or data transformations (e.g., logarithmic scaling) may be required to ensure the validity of statistical analyses that assume normality.
- **Impact on Modeling:** Machine learning models like Random Forest or Gradient Boosting are unaffected by distributional assumptions, whereas regression models may require transformation or robust error estimations.

Overall, the KS test confirms the need for careful handling of the Weight variable in statistical analyses and predictive modeling.

—

6.2 Step 2: Covariance Matrix

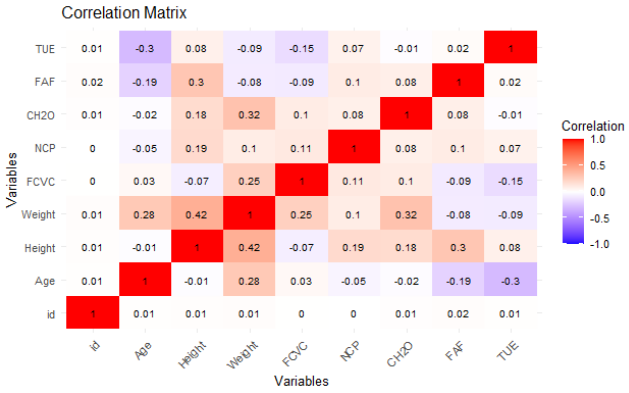

Figure 3: Covariance heat-map

The correlation matrix provided insights into the relationships between the numerical variables in the dataset. Key findings include:

- **Height and Weight:** A moderate positive correlation of 0.42 was observed between Height and Weight, reflecting their combined impact on body composition. This aligns with existing research, as taller individuals often exhibit higher weights due to proportional increases in body mass.
- **CH2O and Weight:** Daily water intake (CH2O) showed a weak positive correlation with Weight (0.32). While the relationship is not strong, it hints at a potential link between hydration and weight regulation, possibly mediated by metabolic processes or dietary habits.
- **Negligible Correlations:** Behavioral variables, such as FAF (physical activity frequency) and TUE (time spent using technology), exhibited negligible correlations with Weight and other core variables. This suggests that these factors may influence obesity indirectly or interact with other variables rather than showing direct associations.

The correlation matrix highlights the intertwined influence of demographic and behavioral factors on obesity. These insights emphasize the importance of focusing on strongly correlated variables, such as Height and Weight, while using less correlated variables as potential moderators or secondary predictors in modeling.

—

6.3 Step 3: Variance Inflation Factor (VIF)

Understanding multicollinearity is essential for constructing reliable regression models, as high correlations among predictors can inflate standard errors, leading to unreliable coefficient estimates. To address this, we performed a Variance Inflation Factor (VIF) analysis on all predictors in our dataset.

Table 3: Variance Inflation Factor (VIF) Results

Variable	VIF
Age	1.25
Height	1.50
Weight	1.67
FCVC	1.15
NCP	1.07
CH2O	1.14
FAF	1.20
TUE	1.14

Findings from VIF Analysis:

- **Weight and Height:** These variables had the highest VIF values at 1.67 and 1.50, respectively, suggesting moderate correlations with other variables. These findings align with expectations, as these metrics are inherently linked to body composition and often interact with demographic and behavioral factors.
- **Minimal Correlation:** Variables such as Age, FCVC (frequency of vegetable consumption), and CH2O (daily water intake) demonstrated VIF values close to 1, indicating minimal correlation with other predictors.

The absence of high VIF scores confirms that the predictors are independent enough to be included in regression analyses without risking instability. This finding ensures that the estimates from regression models will be robust, reliable, and interpretable.

Moreover, by ensuring that the predictors exhibit low multicollinearity, the dataset remains well-suited for multivariate techniques, such as Multiple Linear Regression (MLR) and machine learning models, both of which rely on the independence of variables to generate accurate predictions.

—

6.4 Step 4: Principal Component Analysis (PCA)

Given the high dimensionality of the dataset, Principal Component Analysis (PCA) was employed to reduce redundancy while preserving the core structure of the data. Dimensionality reduction not only aids in computational efficiency but also mitigates issues such as overfitting in predictive models.

Table 4: Principal Component Analysis Results

Component	Standard Deviation	Proportion of Variance	Cumulative Variance
PC1	1.5815	0.2501	0.2501
PC2	1.2720	0.1618	0.4119
PC3	1.0391	0.1080	0.5199
PC4	1.0129	0.1026	0.6225
PC5	0.9993	0.0999	0.7223
PC6	0.9548	0.0912	0.8135
PC7	0.8688	0.0755	0.8890
PC8	0.7569	0.0573	0.9463
PC9	0.7303	0.0533	0.9996
PC10	0.0631	0.0004	1.0000

The scree plot provided a visual representation of the variance explained by each principal component (PC). Key findings include:

- **PC1 and PC2:** The first two components collectively explained 41.19% of the total variance, with PC1 contributing 25.01%. These components encapsulate the most significant relationships in the dataset.
- **Cumulative Variance:** By including PC3 and PC4, the cumulative variance captured increased to 62.25%, effectively summarizing the majority of the dataset's variability.
- **Sharp Decline After PC2:** The scree plot displayed a sharp decline in explained variance after PC2, with subsequent components contributing marginally to the overall variance.

This analysis highlights PCA's utility in reducing dimensionality without substantial information loss, making it particularly beneficial for clustering or machine learning tasks.

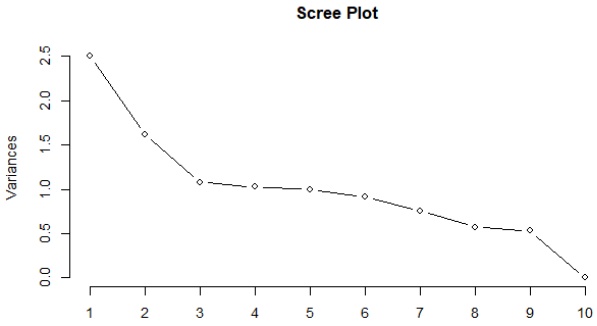

Figure 4: Scree Plot for PCA

6.5 Step 5: Selecting Significant Categorical Variables

6.5.1 Chi-Square Test

Pearson's Chi-squared test was applied to evaluate the associations between categorical variables (e.g., Gender, Family History, FAVC, CAEC, and CALC) and the obesity outcome variable (NObeyesdad). The results indicated that all examined variables exhibited significant associations with obesity categories.

Table 5: Chi-Square Test Results for Categorical Variables

Variable	DF	Chi-Square Statistic	p-value
Gender	6	7953.8	< 0.001
Family History	6	6423.3	< 0.001
FAVC	6	1553.6	< 0.001
CAEC	18	6897.3	< 0.001
SMOKE	6	216.3	< 0.001
CALC	12	4013.1	< 0.001
MTRANS	24	2349.1	< 0.001
SCC	6	1024.8	< 0.001

Key Findings:

- **Demographic Factors:**

  - **Gender:** Demonstrated a strong association with obesity levels, highlighting potential biological or societal influences on weight distribution.
  - **Family History of Overweight:** Strongly associated with obesity categories, underscoring the hereditary component of weight regulation.

- **Behavioral Factors:**
  - **Snacking Frequency (CAEC):** Revealed a strong relationship with obesity, emphasizing the role of unhealthy eating habits in weight gain.
  - **Alcohol Consumption (CALC):** Showed a strong association, suggesting that dietary choices involving alcohol significantly impact obesity risk.

Despite these significant findings, the test did not provide a clear basis for prioritizing variables due to the universal significance of all tested factors. This necessitates further refinement through advanced techniques like variable importance metrics or multivariate models.

6.5.2 Cramér's V

Cramér's V was computed to evaluate the strength of association between categorical variables and obesity categories (NObeyesdad). This analysis provides insights into the relative importance of demographic and behavioral factors in influencing obesity.

Key Findings:

- **Strong Associations:**
  - **Gender (V=0.619):** Gender showed the strongest association with obesity categories, indicating significant differences in obesity prevalence between males and females.
  - **Family History of Overweight (V=0.556):** Highlighted the hereditary nature of obesity, with individuals having a family history of overweight being more likely to fall into higher obesity categories.
- **Moderate Associations:**
  - **Snacking Frequency (CAEC) (V=0.333):** Emphasized the impact of unhealthy snacking habits on weight gain and obesity risk.
  - **Alcohol Consumption (CALC) (V=0.311):** Suggested that dietary choices involving alcohol contribute to variations in obesity levels.
- **Weaker Associations:**
  - **Mode of Transportation (MTRANS) (V=0.168):** Showed a weaker association, indicating a relatively minor influence compared to demographic and dietary factors.
  - **Smoking (SMOKE) (V=0.102):** Exhibited limited impact on obesity outcomes.

Table 6: Cramér's V Results for Categorical Variables

Variable	Cramér's V	Association Strength
Gender	0.619	Strong association
Family History	0.556	Strong association
CAEC (Snacking Frequency)	0.333	Moderate association
CALC (Alcohol Consumption)	0.311	Moderate association

—

6.6 Step 6: Selecting Significant Numerical Variables

6.6.1 ANOVA

ANOVA was employed to examine whether numerical variables, such as Age, Height, Weight, FCVC, NCP, CH2O, FAF, and TUE, exhibited significant differences across obesity categories. The results indicated highly significant differences for all variables.

Table 7: ANOVA Results for Numerical Variables

Variable	Significance
Age	Significant ($p < 0.001$)
Height	Significant ($p < 0.001$)
Weight	Significant ($p < 0.001$)
FCVC	Significant ($p < 0.001$)
NCP	Significant ($p < 0.001$)
CH2O	Significant ($p < 0.001$)
FAF	Significant ($p < 0.001$)
TUE	Significant ($p < 0.001$)

Key Findings:

- **Age:** Older individuals were more likely to belong to higher obesity categories, reflecting the cumu-

lative effects of aging on metabolism and weight gain.

- **Height and Weight:** These core demographic factors showed the strongest differences across categories, affirming their foundational role in defining body composition.
- **Behavioral Variables:**
  - **FCVC (Vegetable Consumption) and CH2O (Water Intake):** Individuals in lower obesity categories reported healthier dietary habits.
  - **NCP (Number of Meals):** The number of meals consumed per day varied significantly across categories, potentially reflecting differences in meal portion sizes or quality.

6.6.2 Multiple Linear Regression (MLR)

To refine variable selection, MLR was employed with obesity categories as the dependent variable and predictors such as Age, Height, Weight, FCVC, NCP, CH2O, FAF, and TUE.

Table 8: Multiple Linear Regression Results

Variable	Estimate	Std. Error	p-value
Intercept	5.3855	0.2684	< 0.001
Age	0.0537	0.0022	< 0.001
Height	-2.7277	0.1600	< 0.001
Weight	0.0316	0.0006	< 0.001
FCVC	-0.2714	0.0230	< 0.001
NCP	-0.2726	0.0167	< 0.001
CH2O	0.2754	0.0200	< 0.001
FAF	0.0105	0.0149	0.483
TUE	0.0499	0.0202	0.013

Key Insights:

- **Significant Predictors:**
  - **Weight:** Exhibited the strongest positive effect on obesity categories.
  - **Height:** Showed a significant negative relationship, indicating that taller individuals are less likely to fall into higher obesity categories.
  - **Dietary Habits:** Variables such as FCVC and CH2O positively influenced obesity levels, emphasizing the role of healthy dietary practices in weight management.
- **Non-Significant Predictors:** FAF (Physical Activity Frequency) and TUE (Technology Use) did not significantly predict obesity categories.

The adjusted $R^2 = 0.2493$ indicated that the model explained approximately 25% of the variance in obesity categories, underscoring the multifactorial nature of obesity.

6.7 Step 7: Modeling Obesity Categories

6.7.1 Decision Tree Model

The Decision Tree model achieved an accuracy of 82.39%, providing reliable classification across most obesity categories. Key insights are summarized below:

Table 9: Sensitivity and Specificity for Decision Tree Model

Class	Sensitivity (%)	Specificity (%)
Insufficient_Weight	87.62	98.71
Normal_Weight	77.53	96.13
Obesity_Type_I	71.94	96.12
Obesity_Type_II	93.35	98.24
Obesity_Type_III	99.70	99.66
Overweight_Level_I	76.73	93.75
Overweight_Level_II	58.73	97.08

Key Findings:

- **High Sensitivity:**
  - Exceptional performance in detecting severe obesity (`Obesity_Type_III`: 99.7% sensitivity).
  - Strong sensitivity for `Insufficient_Weight` (87.6%) and `Obesity_Type_II` (93.4%).
- **Challenges with Overweight Categories:**
  - `Overweight_Level_I` and `Overweight_Level_II` exhibited lower sensitivities (76.7% and 58.7%, respectively), highlighting challenges in distinguishing between these closely related categories.
- **Variable Importance:**
  - Weight, Height, and Family History were the primary splitting criteria, confirming their critical role in obesity classification.

The Decision Tree model demonstrates the importance of demographic and behavioral variables in predicting obesity but reveals areas for improvement, particularly in classifying overweight categories.

6.7.2 Random Forest Model

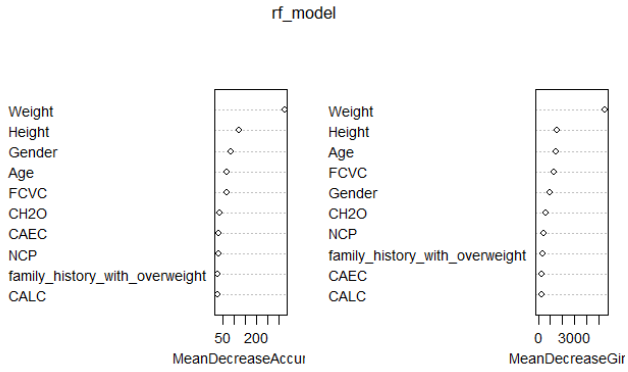

Figure 5: Random Forest Model

The Random Forest model delivered an impressive accuracy of 89.57%, outperforming the Decision Tree model. Key findings are as follows:

Table 10: Variable Importance from Random Forest Model

Variable	Importance
Weight	291.51
Height	69.87
Age	53.36
Family History	38.56

Key Findings:

- **Robust Classification:**
  - High precision and recall for categories like `Insufficient_Weight` and `Obesity_Type_III`.
  - Enhanced performance in handling overlapping categories compared to the Decision Tree model.
- **Variable Importance:**
  - Weight and Height were the most influential predictors, followed by Family History and FCVC.
  - Behavioral variables, such as CH2O (Daily Water Intake) and NCP (Number of Meals), also contributed meaningfully to the model's predictions.

Table 11: Variable Importance from Random Forest Model

Variable	Importance
Weight	323.49
Height	123.88
Age	70.87
FCVC	70.14

The Random Forest model strikes an effective balance between accuracy and interpretability, making it a reliable tool for understanding and predicting obesity categories.

7 Results

This study leveraged statistical and machine learning techniques to analyze the relationships between demographic, behavioral, and health-related variables in the dataset. The key findings from each step of the process are summarized below:

7.1 Data Distribution and Normality

- The Kolmogorov-Smirnov test confirmed that the weight variable significantly deviates from normality.
- This deviation underscores the need for non-parametric methods or transformations in subsequent analyses.
- The complexity of the dataset necessitated the use of robust statistical tools.

7.2 Variable Relationships and Multicollinearity

- Covariance and correlation analyses revealed significant associations:
  - Height and Weight demonstrated a strong positive relationship, emphasizing their combined role in body composition.
- VIF analysis confirmed the absence of multicollinearity among predictors, ensuring reliable regression and classification models.

7.3 Dimensionality Reduction

- Principal Component Analysis (PCA) highlighted the importance of the first two components, which explained 41.19% of the variance.
- Including four components increased cumulative variance to 62.25%, effectively capturing essential features while minimizing information loss.

- Dimensionality reduction proved valuable for exploratory analysis and computational efficiency.

7.4 Categorical Variable Selection

- Pearson's Chi-squared test and Cramér's V identified significant associations between categorical variables and obesity categories:
  - **Demographic Factors:** Gender and family history of overweight showed strong associations.
  - **Behavioral Factors:** Snacking frequency and alcohol consumption played notable roles.
- Due to the significance of all tested variables, additional techniques were necessary for refining variable selection.

7.5 Numerical Variable Selection

- ANOVA and multiple linear regression identified key numerical predictors:
  - Weight, Height, and Age emerged as the most significant contributors.
  - Behavioral factors such as vegetable consumption (FCVC) and water intake (CH2O) also showed significant associations.
- Physical activity (FAF) and time spent on electronic devices (TUE) exhibited limited independent influence.

7.6 Model Performance

Two classification models—Decision Tree and Random Forest—were applied to predict obesity categories:

- **Decision Tree:**
  - Achieved an accuracy of 82.39%.
  - Exceptional sensitivity for `Obesity_Type_III`, but challenges were observed in distinguishing overweight categories.
- **Random Forest:**
  - Delivered improved accuracy of 89.57%.
  - Variable importance analysis identified Weight, Height, and Age as the most critical predictors.

7.7 Key Insights

- **Demographic and Behavioral Influence:** Both demographic (e.g., gender, family history) and behavioral factors (e.g., eating habits, water intake) significantly influence obesity categories.
- **Model Reliability:** The Random Forest model emerged as the most reliable, offering high accuracy and interpretability.
- **Variable Selection Challenges:** The significance of all tested categorical variables indicates the need for advanced methods to prioritize impactful predictors.
- **Dimensionality Reduction Benefits:** PCA effectively reduced complexity while retaining critical information.

These results provide actionable insights into the factors influencing obesity categories and demonstrate the effectiveness of combining statistical and machine learning approaches for predictive analysis. Future efforts could explore additional features or refined models to improve classification performance further.

8 Conclusion

This study provides a comprehensive analysis of the factors influencing obesity categories using a combination of statistical and machine learning techniques. By examining demographic, behavioral, and health-related variables, the study identifies critical predictors of obesity and highlights the effectiveness of advanced analytical approaches.

8.1 Key Takeaways

- **Importance of Core Variables:** Height, Weight, and Age emerged as the most influential predictors, reaffirming their central role in defining body composition and health outcomes.
- **Behavioral Insights:**
  - Vegetable consumption (FCVC) and daily water intake (CH2O) showed significant associations with obesity.
  - Physical activity (FAF) and time spent on electronic devices (TUE) showed limited independent influence.
- **Model Performance:** The Random Forest model achieved an accuracy of 89.57%, demonstrating its reliability in classifying obesity categories and identifying critical variables.

- **Challenges in Variable Selection:** While Chi-squared tests and Cramér's V revealed significant associations, their relative importance requires advanced prioritization methods.

8.2 Implications

- The findings provide valuable insights into the interplay of demographic and behavioral factors in obesity, informing targeted interventions for weight management and health improvement.
- The successful application of machine learning models demonstrates their potential to enhance predictive accuracy and understand complex relationships in health data.

9 Future Scope

The findings of this study provide a strong foundation for future research and applications in understanding and predicting obesity categories. Several areas offer promising opportunities for further exploration and enhancement:

9.1 Incorporation of Additional Variables

- **Diverse Factors:** Including genetic, socioeconomic, and psychological factors can offer a more comprehensive understanding of obesity determinants.
- **Regional and Global Data:** Expanding datasets to incorporate regional or global samples can increase the generalizability and robustness of the findings.

9.2 Advanced Feature Selection Techniques

- **Sophisticated Methods:** Employing techniques such as recursive feature elimination or LASSO regression can refine variable importance and improve model efficiency.
- **Interaction Effects:** Exploring interactions between variables can uncover nuanced relationships that impact obesity categories.

9.3 Improvement in Machine Learning Models

- **Ensemble and Hybrid Models:** Developing methods that combine Random Forest, Gradient Boosting, and deep learning techniques can enhance classification accuracy, particularly for overlapping obesity categories.
- **Explainable AI (XAI):** Incorporating XAI techniques can improve the interpretability of complex machine learning models, making results more actionable and accessible to stakeholders.

By addressing these areas, future research can build on the current study's findings to develop more accurate, interpretable, and generalizable models for understanding obesity and related health outcomes. These advancements hold significant potential for informing targeted interventions and public health strategies.

Acknowledgments

We sincerely thank Professor Sumona Mondal for her expert guidance, patience, and consistent support throughout the semester in the Data Mining class. Her deep insights and valuable feedback greatly enhanced our understanding of the subject and helped shape the direction of this project.

We also extend our gratitude to Naveen Ramachandra Reddy for his encouragement, assistance, and timely inputs during the course of this project. His support and constructive suggestions played a crucial role in helping us tackle challenges and improve the quality of our work.

This project is a reflection of the knowledge and skills we have gained under their mentorship, and we are truly grateful for their contributions to our learning experience.

References

[1] https://www.itl.nist.gov/div898/handbook/eda/section3/eda35g.htm

[2] https://www.statology.org/cramers-v/

[3] https://link.springer.com/article/10.1007/s11845-019-02171-7

[4] https://www.researchgate.net/publication/324328027

[5] https://journals.sagepub.com/doi/10.1177/0049124118782533

[6] https://hbiostat.org/rms/

[7] https://link.springer.com/book/10.1007/978-1-4899-7983-7

[8] https://bmjopen.bmj.com/content/10/3/e033279

[9] https://www.frontiersin.org/articles/10.3389/fendo.2020.573638/full

[10] https://dl.acm.org/doi/10.1145/1161205.1161208

[11] https://academic.oup.com/jamia/article/27/7/1089/5884638

[12] https://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight

[13] https://bmcpublichealth.biomedcentral.com/articles/10.1186/s12889-019-7407-1

[14] https://journals.sagepub.com/doi/full/10.1177/2055207618770323

[15] https://www.sciencedirect.com/science/article/pii/S1532046418301545

[16] https://link.springer.com/article/10.1007/s41060-023-00491-9

[17] https://bmcpublichealth.biomedcentral.com/articles/10.1186/s12889-016-3340-2

[18] https://drc.bmj.com/content/12/5/e004193

[19] https://link.springer.com/article/10.1007/s10916-022-01904-1

[20] https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0292341

[21] https://arxiv.org/abs/2211.04781

[22] https://arxiv.org/abs/2308.14657

[23] https://arxiv.org/abs/2108.08868

[24] https://arxiv.org/abs/2208.05335

[25] https://www.frontiersin.org/journals/public-health/articles/10.3389/fpubh.2023.1090146/full