CHAPTER ONE

1.0 Introduction

Obesity continues to be a pressing and multifaceted public health concern, affecting millions of individuals globally. The rise in obesity rates across various age groups and regions has significant implications, including elevated risks of developing chronic illnesses such as type 2 diabetes, cardiovascular conditions, musculoskeletal disorders, and certain types of cancer (World Health Organization, 2023). In addition to its physiological burden, obesity also contributes to psychosocial stress, stigma, and reduced quality of life. Understanding, diagnosing, and mitigating obesity through data-driven approaches has become an essential goal for public health agencies, clinicians, and researchers alike.

While traditional epidemiological methods offer valuable insights into obesity trends and risk factors, modern advancements in data science now provide an opportunity to enhance predictive capabilities through machine learning (ML). ML models, especially supervised learning algorithms, are capable of learning complex patterns from historical data to classify individuals based on their risk level. This predictive capability not only improves the efficiency of diagnosis but also enables proactive intervention strategies (Rajkomar, A. et.al, 2019).

In this study, I develop a supervised ML model to classify individuals into various obesity categories using a broad array of features encompassing demographic information, lifestyle behaviors, and physiological metrics. my aim is to investigate the effectiveness of the Random Forest algorithm in providing accurate predictions and interpretable insights. Additionally, we employ SHAP (SHapley Additive exPlanations) values to interpret the model's decision-making process and identify key risk factors.

CHAPTER TWO

2.0 Dataset Overview

2.1 Source: UCI Machine Learning Repository — Obesity Levels Dataset (https://archive.ics.uci.edu/dataset/544/estimation+of+obesity+levels+based+on+eating+habits+and+physical+condition)

2.2 Dataset Size: The dataset comprises 2,111 individual records. Each entry contains 17 variables, including demographic, physiological, and behavioral attributes that collectively influence obesity.

2.3 Target Variable: NObeyesdad, the target variable represents categorical obesity classifications such as Insufficient Weight, Normal Weight, Overweight, and various grades of obesity (Type I, II, and III). This makes the problem a multiclass classification task.

2.4 Key Features:

- **Demographic Information**: Gender, Age
- **Physiological Data**: Height (in meters), Weight (in kilograms)
- **Behavioral Patterns**: Smoking habits, family history with overweight
- **Lifestyle Choices**: Vegetable intake frequency, daily water consumption, alcohol use, number of meals, snack habits, physical activity (FAF), tech device usage, and transportation mode

This dataset was chosen for its rich and diverse representation of contributing factors to obesity, as well as its sufficient sample size for building robust machine learning models. It allows the integration of multiple dimensions of human behavior and biology in predicting a complex health condition.

Table 1. Key features of the Dataset used.

Column Name	Meaning
Gender	Male or Female
Age	Age of the person
Height	Height (in meters)
Weight	Weight (in kg)
Family_history_with_overweight	Yes/No
FAVC	Frequent high-calorie food (Yes/No)
FCVC	Frequency of vegetable consumption
NCP	Number of main meals per day
CAEC	Consumption of food between meals
SMOKE	Smoker (Yes/No)
CH2O	Water consumption (liters/day)
SCC	Calories monitor app user (Yes/No)
FAF	Physical activity frequency
TUE	Time spent using tech devices
CALC	Alcohol consumption
MTRANS	Mode of transportation
NObeyesdad	Target variable (Obesity level)

CHAPTER THREE

3.0 Research Methodology

3.1 Data Cleaning & Preparation

Effective machine learning begins with robust data preprocessing. The dataset was examined for inconsistencies and missing values, of which none were present. Categorical features such as 'Gender', 'Smoking', and 'Transportation Method' were transformed into binary variables using label encoding to facilitate model training. Numerical features including Age, Height, and Weight were standardized using StandardScaler to normalize feature scales and prevent bias in distance-based learning algorithms.

The dataset was subsequently divided into training and test sets using an 80/20 ratio, ensuring an unbiased evaluation of model performance. This stratification also maintained proportional class distribution across both sets, preventing class imbalance issues in testing.

3.2 Exploratory Data Analysis (EDA)

Exploratory data analysis was conducted to better understand the structure and distributions within the dataset:

1. **Grouped bar charts**

The Grouped bar charts provided insights into how categorical features (e.g., alcohol use, gender, transportation) relate to obesity levels.

- **Obesity Level Distribution by Gender:** The analysis of obesity distribution across genders revealed several key patterns. Obesity_Type_III recorded the highest count overall, with females significantly more represented than males. Similarly, Obesity_Type_II had a high frequency in both genders, though females maintained a slight lead. In contrast, Obesity_Type_I and Overweight_Level_II were more prevalent among males. The categories Overweight_Level_I and Normal_Weight showed a nearly equal gender distribution. Notably, females were more represented at the extremes—in both Insufficient_Weight and Obesity_Type_III—while males were more dominant in mid-level categories

such as Overweight_Level_II and Obesity_Type_I. This suggests a potential gender-based variation in the distribution of obesity severity levels.

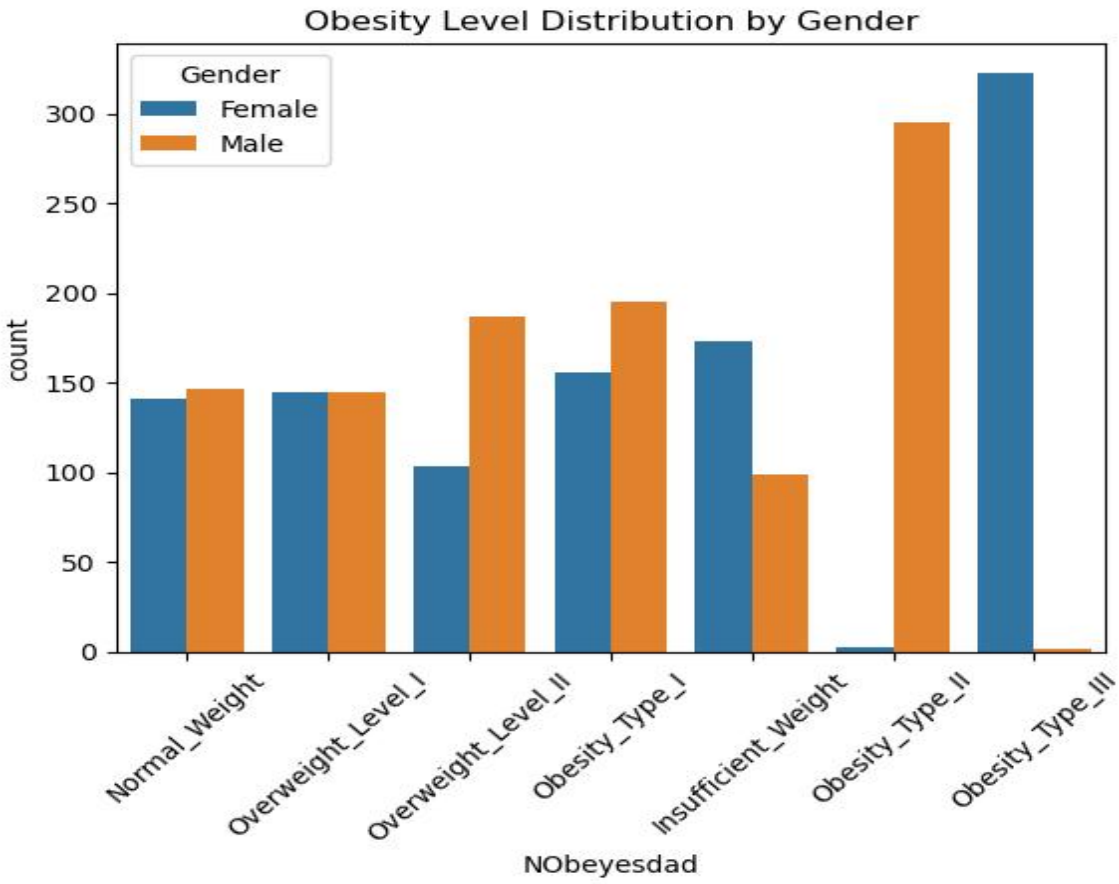

Figure 1: Obesity Level Distribution by Gender

- **Alcohol Consumption Patterns Across Obesity Levels:** The analysis of alcohol consumption patterns across obesity levels reveals that moderate drinking ("Sometimes") is the most common behavior, especially among individuals classified as Obesity Type III, Obesity Type II, and Overweight Level I. In contrast, non-drinkers ("No") were most prevalent in Obesity Type I, followed by Overweight Level II, Normal Weight, and even Insufficient Weight, suggesting a notable presence of non-alcohol users across a range of weight categories. Frequent and daily drinkers ("Frequently" and "Always") were rare overall, with only minor representation in Normal Weight and Obesity Type I groups. Particularly within Obesity Type III, the dominance of “Sometimes” drinkers suggests a possible association between moderate alcohol use and extreme obesity. Meanwhile, the Insufficient Weight group showed a higher proportion of

non-drinkers, indicating that individuals with lower body weight are generally less likely to consume alcohol.

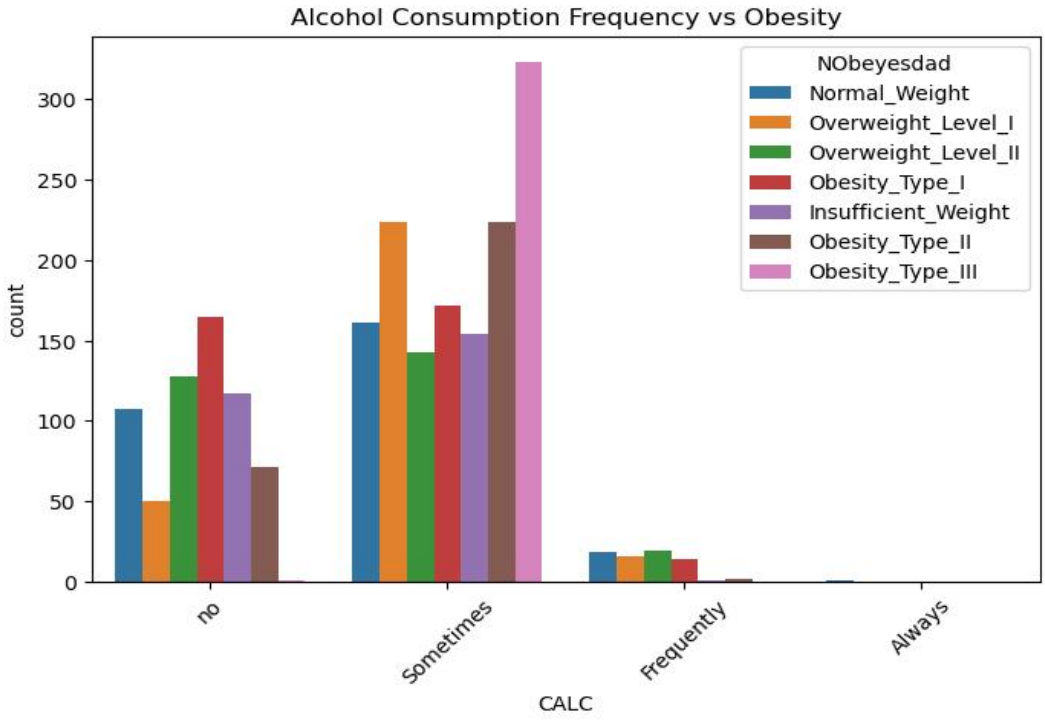

Figure 2: Alcohol Consumption Frequency vs Obesity

- **Transportation Patterns and Their Link to Obesity Levels:** The analysis of transportation patterns reveals that public transportation is the most widely used mode across all obesity levels, with particularly high usage among individuals in Obesity Type III, Obesity Type I, and Overweight Level I. This widespread reliance may be influenced by accessibility or socioeconomic factors. Automobile usage was notably higher among Obesity Type I and II individuals, suggesting a possible link between frequent car use and moderate obesity, whereas normal weight individuals were least likely to use cars. In contrast, walking was more common among those with Normal or Insufficient Weight, and rare among higher obesity categories, especially in Obesity Type III, indicating reduced physical activity may contribute to weight gain. Bike and motorbike usage was extremely rare overall, with slightly more presence among normal-weight individuals. For those in Obesity Type III, reliance on public transport was nearly exclusive, with minimal walking or car usage, potentially reflecting mobility limitations or lifestyle constraints. Overall, these findings suggest that active transport methods

such as walking and cycling are more associated with lower obesity levels, and promoting them may support healthier urban populations.

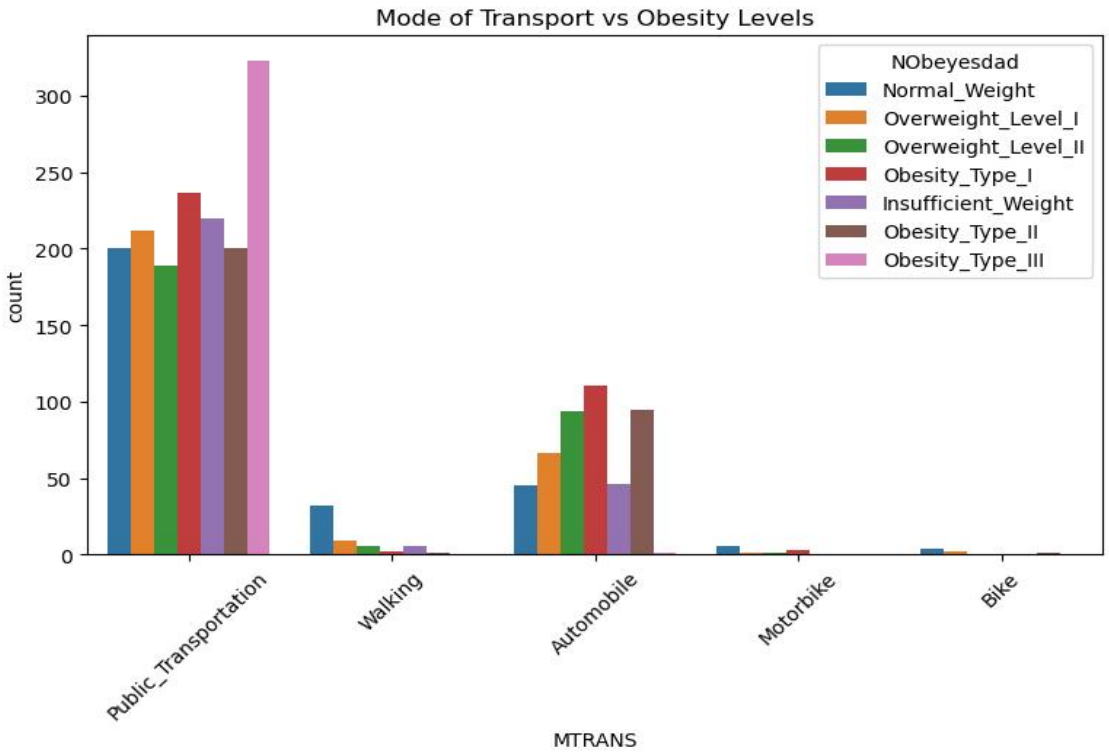

Figure 3: Mode of Transport vs Obesity Levels

2. **Boxplots**

In this project, boxplots were used to explore the distribution of numerical features, detect the presence of outliers, and visually compare how various metrics (such as age, height, and physical activity level) varied across different obesity classes. This helped in understanding data spread, identifying potential anomalies, and assessing the central tendencies of variables.

- **Body Mass Index (BMI) Distribution Patterns Across Obesity Classes:** The distribution of BMI values across obesity categories revealed a clear, stepwise progression from Insufficient Weight to Obesity Type III, with each class occupying a distinct BMI range and exhibiting minimal overlap. Normal Weight served as the healthy baseline, while Overweight Levels I and II represented transitional zones leading into higher risk categories. A significant observation was the sharp jump in BMI between Overweight Level II and Obesity Type I,

marking a critical threshold in obesity progression. Notably, Obesity Type III showed the widest BMI range, indicating high variability in extreme obesity, whereas Insufficient Weight had the narrowest spread. These patterns affirm BMI's reliability as a classification metric and underscore the need for stage-specific intervention strategies, particularly during the shift from overweight to obesity, where early prevention efforts may be most impactful.

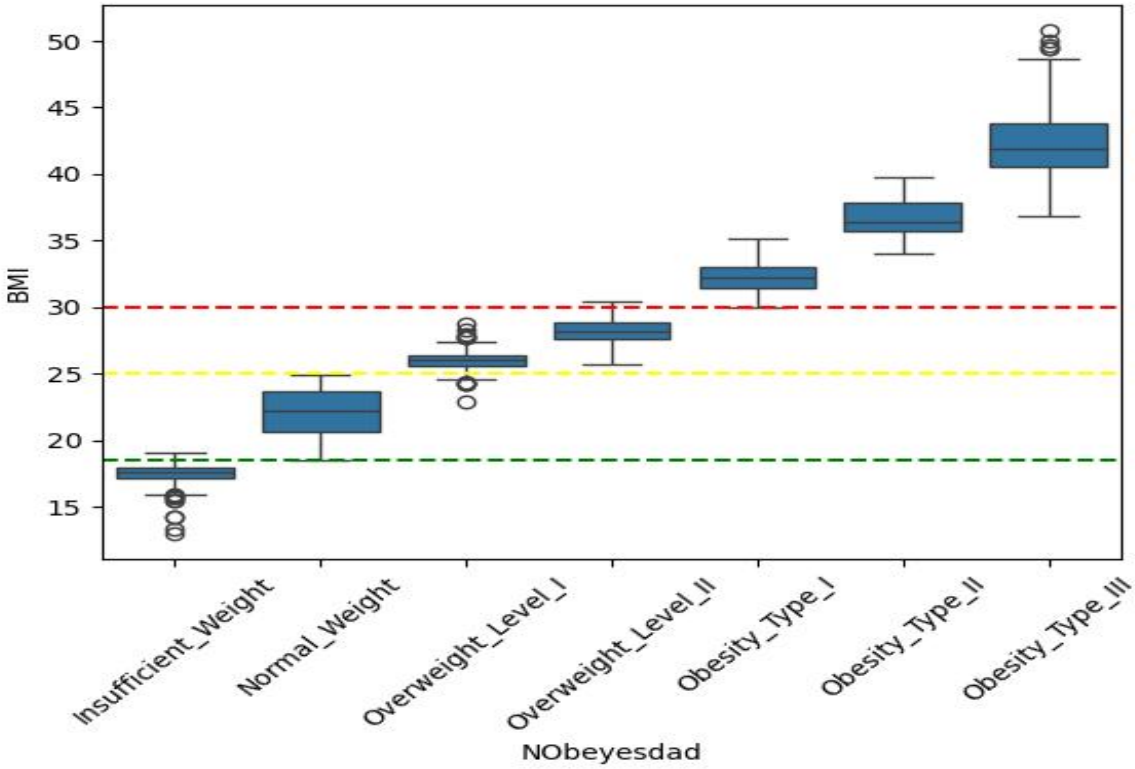

Figure 4: Body Mass Index (BMI) Distribution Patterns Across Obesity Classes

- **Physical Activity Trends Across Obesity Levels**: An inverse relationship was observed between physical activity levels and obesity severity. Normal Weight individuals exhibited the highest levels of physical activity, while activity steadily declined as obesity severity increased, reaching the lowest levels in Obesity Type III. A particularly sharp drop in physical activity was noted between Overweight Level II and Obesity Type I, marking a critical transition point. Interestingly, the Insufficient Weight group showed moderate activity levels, suggesting that factors other than physical inactivity such as nutritional deficiencies or health conditions may contribute to low weight. Additionally, a subset of Obesity Type I individuals maintained moderate activity, which may point to underlying metabolic factors affecting weight independent of physical exertion. These

findings underscore the importance of tailored interventions that consider both behavioral and physiological contributors to obesity.

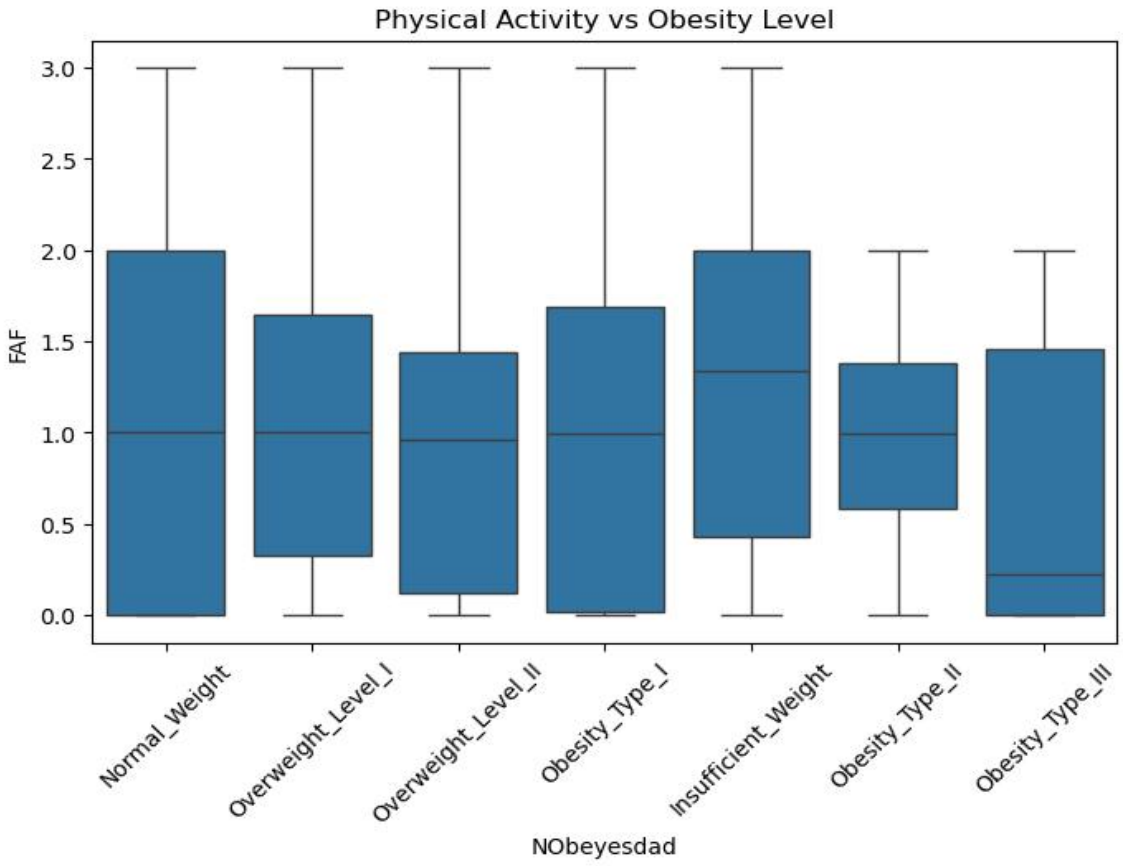

Figure 5: Physical Activity vs Obesity Level

- **Hydration Patterns and Their Link to Obesity Severity:** The analysis of water intake across obesity levels revealed a clear hydration gradient, with Normal Weight individuals consuming the highest amounts of water. A steady decline in hydration was observed from Overweight Level I through to Obesity Type III, where intake was lowest, potentially indicating dehydration risks in individuals with severe obesity. The largest drop in water consumption occurred between Overweight Level II and Obesity Type I, suggesting this as a critical threshold for lifestyle intervention. Interestingly, the Insufficient Weight group did not follow this trend, maintaining moderate water intake despite low body weight, a finding that may point to non-hydration-related causes of underweight, such as nutritional deficiencies or medical conditions. These patterns highlight the need for targeted hydration interventions, especially among Overweight Level II individuals, and

suggest that monitoring hydration in Normal Weight individuals could serve as a preventive measure for future weight gain.

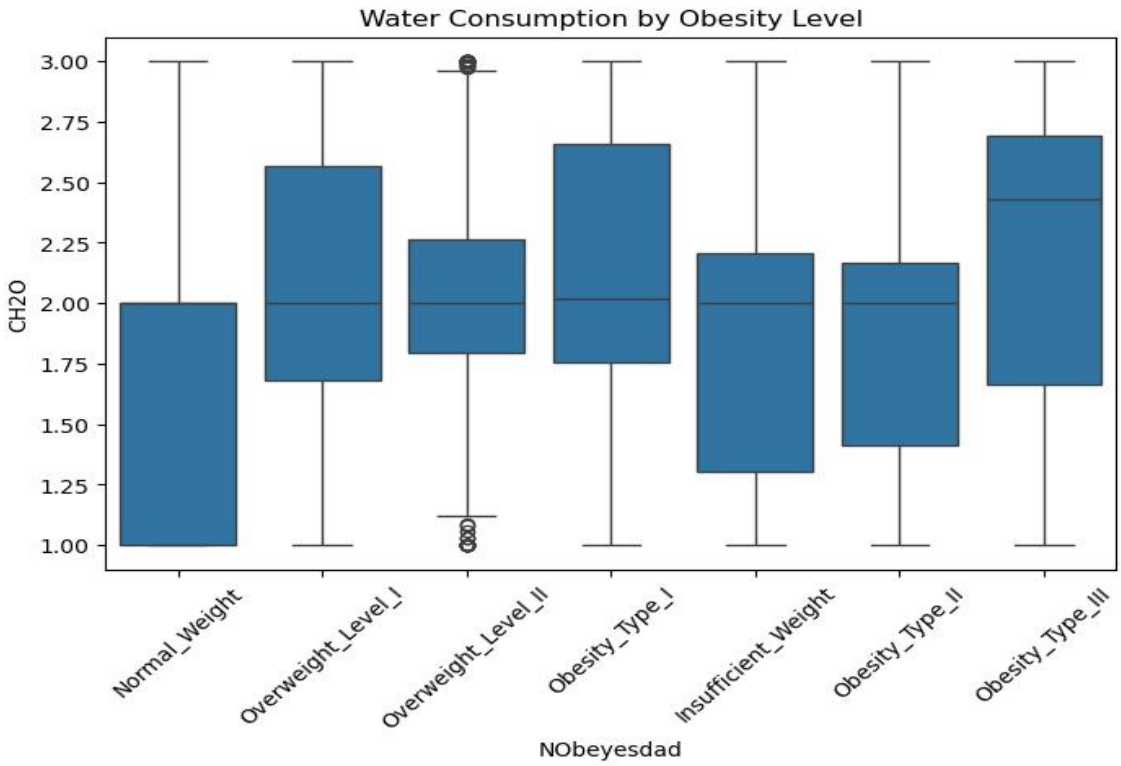

Figure 6: Water Consumption by Obesity Level

- **Outlier Analysis Across Key Features:** Outlier analysis revealed several important patterns in the dataset. Age displayed outliers on the higher end, particularly above 40 years, which may be due to the dataset targeting a younger population—such as students—while older entries could represent parents, instructors, or input errors. In the Height feature, a few outliers above 1.9 meters were observed, likely due to natural height variation or unit conversion issues (e.g., inches mistaken for meters). Weight showed clear outliers beyond 150 kg, possibly caused by severe obesity cases, data entry errors (like extra zeros), or confusion between kilograms and pounds. For NCP (number of meals per day), there were consistent outliers at 1 and 4 meals, which may reflect inconsistent data input or the use of float values instead of whole numbers. On the other hand, features such as CH2O (water intake), FAF (physical activity), TUE (technology use), FCVC (vegetable consumption), and BMI did not show any significant or

visible outliers. These findings helped guide the decision to cap extreme values where necessary, and to retain the integrity of features with clean distributions.

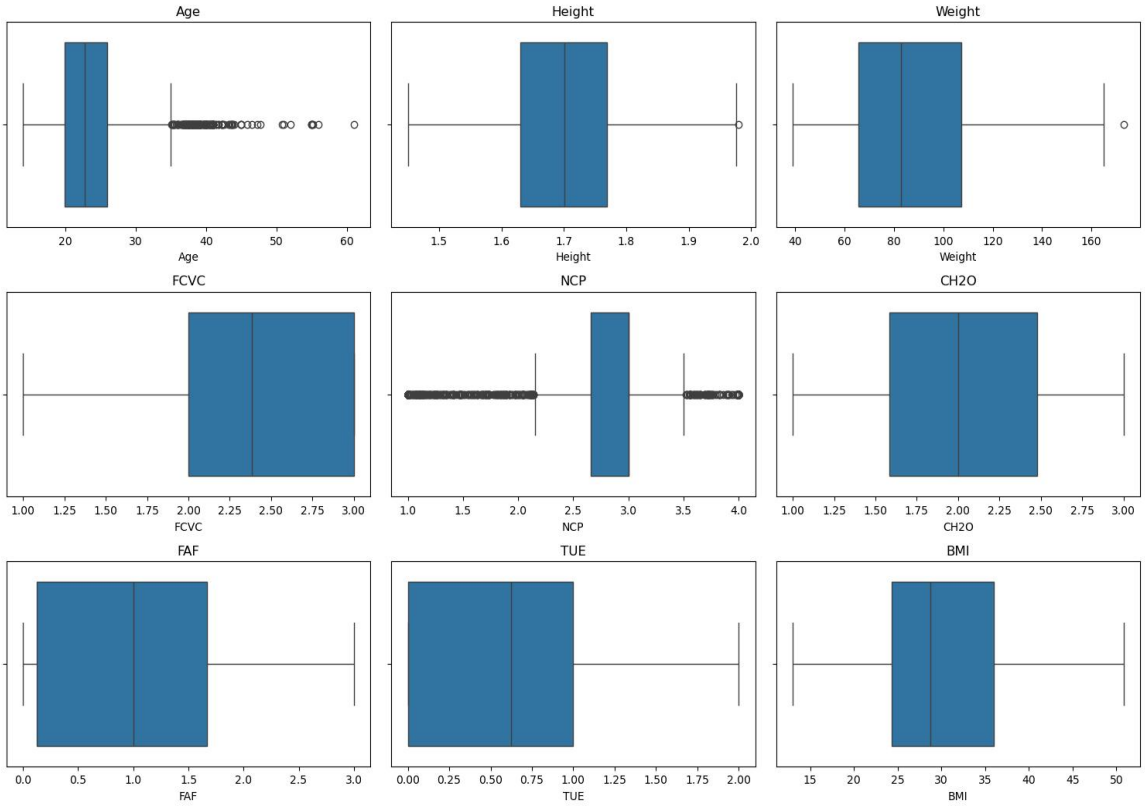

Figure 7: Outlier Analysis Across Key Features

3. Scatterplots

Scatterplots were used to visually examine the relationships between numerical variables such as height and weight, highlighting patterns and clustering across different obesity categories. These plots provided intuitive insights into how features like body dimensions interact, and helped identify separable groups or overlaps between weight classes, supporting both feature understanding and model development.

- **Insights from Height-Weight Scatterplot Analysis:** The scatterplot analysis of height versus weight revealed distinct clustering patterns across obesity categories, illustrating a clear progression in weight from Normal Weight to Obesity Type III. Overweight groups served as transitional clusters between the

normal and obese classes. Notably, taller individuals (above 1.8m) were more concentrated in the higher obesity categories, suggesting that standard BMI thresholds may require adjustment based on height. The 1.6–1.7m height range emerged as a critical transition zone, where weight distributions began to diverge significantly into obesity levels. Interestingly, Insufficient Weight cases appeared across all height groups, indicating that height alone does not determine underweight status, and other factors such as metabolism or diet may be involved. A few outliers, likely muscular individuals, were identified within the Normal Weight class despite high body weight. These findings suggest that obesity screening and weight management strategies may benefit from height-specific considerations, especially for individuals within the 1.6–1.8m range.

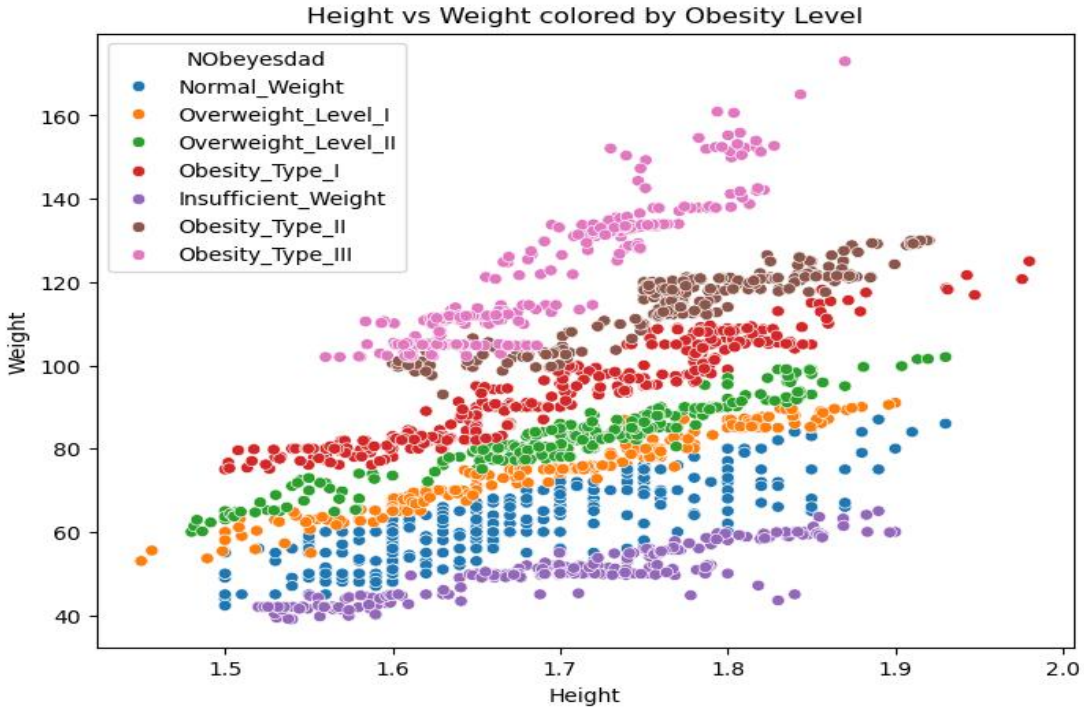

Figure 8: Height vs Weight colored by Obesity Level

- **Critical Height-Weight Transition Zone Analysis:** A more focused analysis within the critical transition zone (1.6m–1.8m) revealed that weight thresholds for obesity vary significantly by height. Individuals between 1.6m and 1.7m typically fall within the normal weight range at 50–70kg, while Obesity Type I begins above 100kg. For those between 1.7m and 1.8m, the normal weight shifts upward to 60–80kg, with obesity onset above 110kg, suggesting a 10kg tolerance shift per 0.1m of height. The range of 1.65m to 1.75m emerged as a high-risk cluster,

marking the steepest gradient of category transitions and serving as a tipping point—particularly around 1.7m, where all obesity subtypes become visible. Outlier patterns also revealed that individuals classified as Insufficient Weight displayed a flat distribution, typically weighing between 40–60kg regardless of height, while Obesity Type III was only present among individuals taller than 1.65m. These findings suggest the importance of height-specific obesity thresholds and targeted intervention for individuals in the critical transition range.

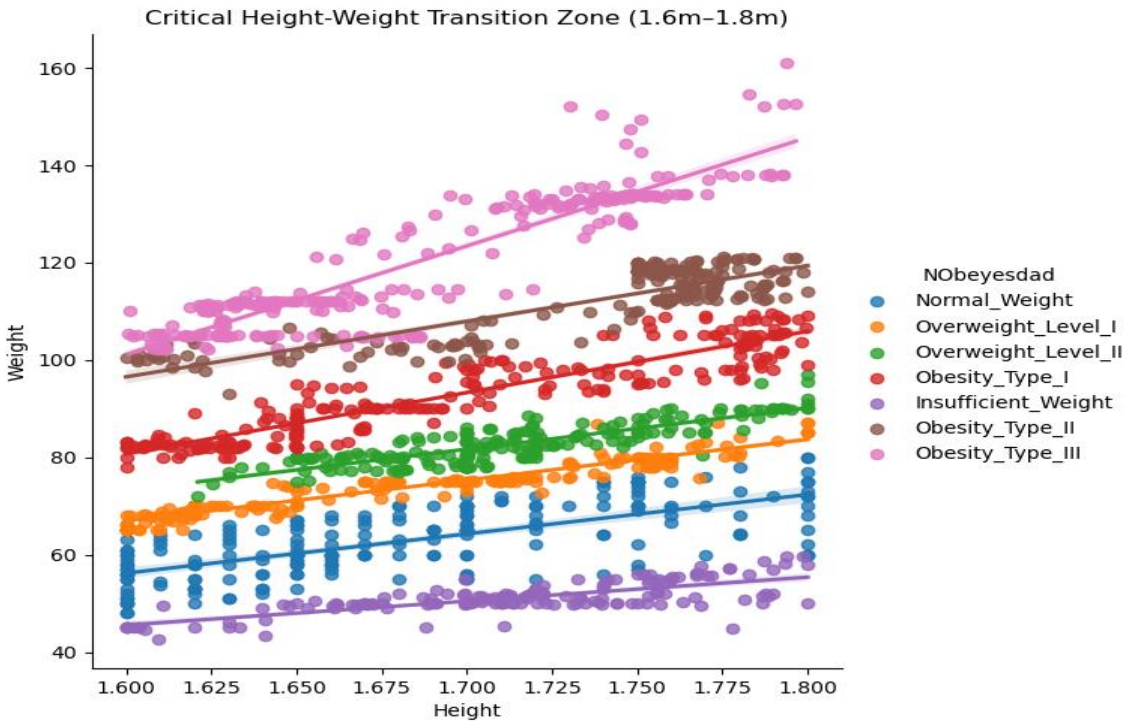

Figure 9: Critical Height-Weight Transition Zone (1.6m–1.8m)

4. Heatmaps

The Heatmap visualized the correlations among continuous variables, showing that weight and height were highly correlated with the target.

- **Correlation Matrix Insights:** The correlation matrix analysis revealed several meaningful relationships among the variables. Weight and BMI exhibited the strongest positive correlation (0.94), which aligns with the fact that BMI is directly derived from weight and height (Nuttall, F.Q. (2015)). A moderate correlation (0.45) was observed between weight and height, indicating that taller individuals tend to weigh more, while weight and family history of overweight

also showed a moderate link (0.48), reinforcing the potential genetic influence on obesity. Interestingly, gender and vegetable consumption frequency (FCVC) had a moderate negative correlation (-0.37), possibly reflecting gender-based differences in eating habits. Age and mode of transportation (MTRANS) showed a relatively strong correlation (0.63), suggesting lifestyle shifts with age. Additional moderate correlations were found between frequent consumption of high-calorie food (FAVC) and weight (0.30), as well as water intake (CH2O) and weight (0.24). Most other variables displayed weak correlations (< 0.3), including smoking habits, physical activity levels, and dietary patterns. Overall, BMI, weight, and family history emerged as key variables with the strongest associations, while gender and age demonstrated targeted interactions with specific behavioral features.

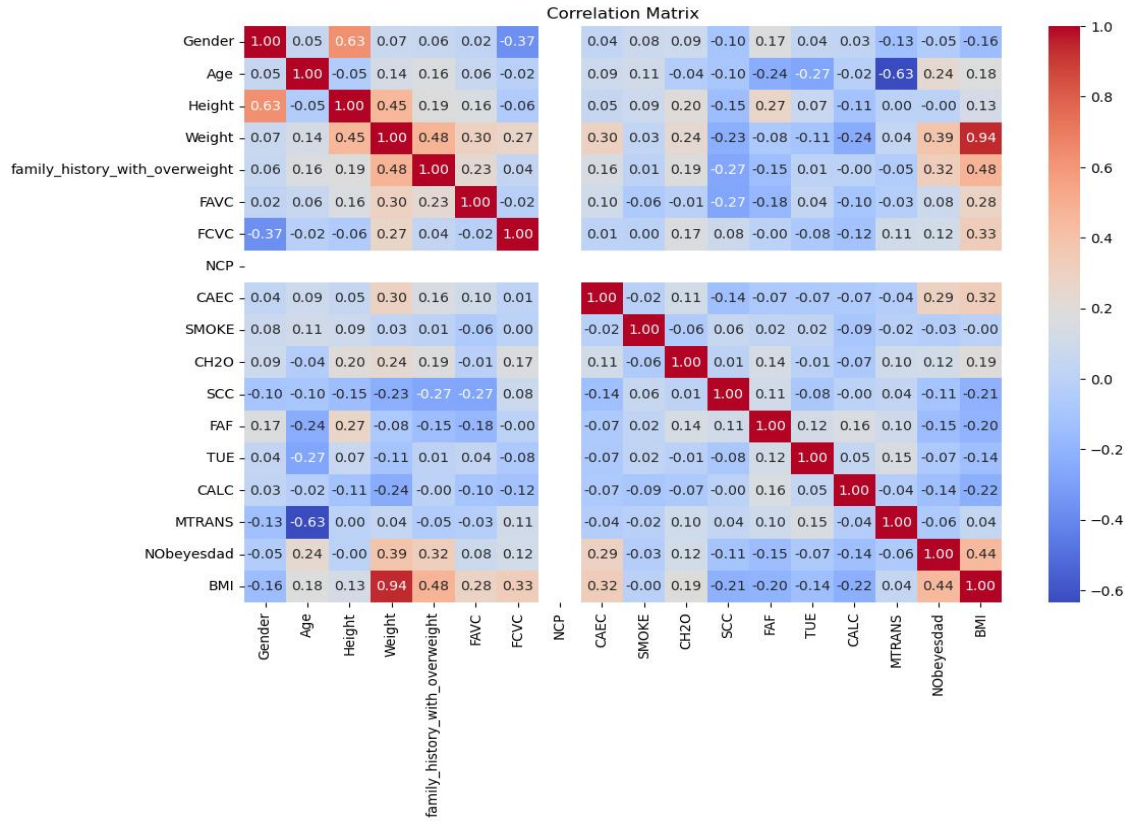

Figure 10: Correlation Matrix

5. **Countplots**

Countplots were used to visualize the frequency of each obesity category

- **Distribution of Obesity Levels:** This chart shows the distribution of obesity levels among participants. It helps to identify class imbalance. Most participants fall under Obesity_Type_1.

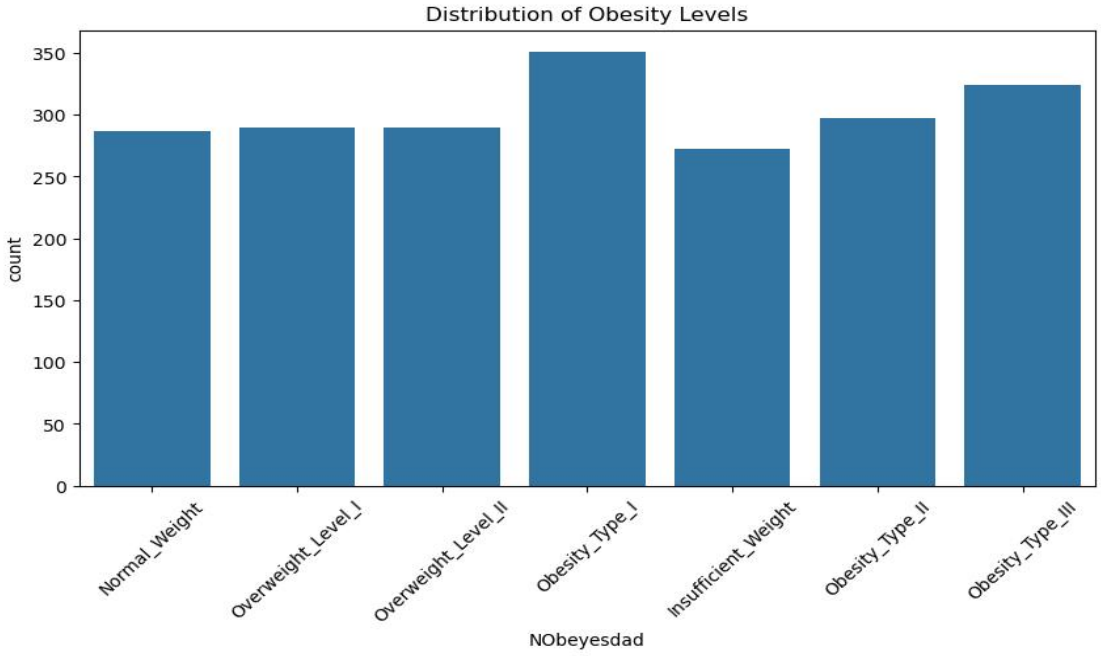

Figure 11: Distribution of Obesity Levels

These visualizations helped uncover non-linear relationships and confirmed that multiple lifestyle and physiological variables influence obesity, aligning with existing literature.

3.3 Dimensionality Reduction

To assess potential redundancy in features and explore dimensionality reduction, Principal Component Analysis (PCA) was performed. A cumulative explained variance plot was generated, illustrating that a small number of components could explain the majority of variance in the dataset. Although PCA was not integrated into the final model to preserve interpretability, this step reinforced that our feature set was both comprehensive and informative.

3.3.1 PCA and Dimensionality Reduction Insights: The PCA explained variance analysis provided valuable insight into dimensionality reduction. The first 10 principal components accounted for approximately 90% of the total variance, indicating that the dataset can be effectively compressed to just 10 features without significant loss of information. This is particularly beneficial for simplifying models while maintaining predictive performance. An elbow point was observed around components 10 to 12, beyond which the explained variance plateaued, suggesting diminishing returns in model complexity beyond that range. For an optimal tradeoff between simplicity and data retention, using 7–10 components offers a balance with around 85–90% variance explained, while 12+ components may capture more detail but at the cost of increased complexity. Overall, PCA effectively compressed correlated features, making subsequent tasks such as clustering, classification, and regression more efficient and interpretable.

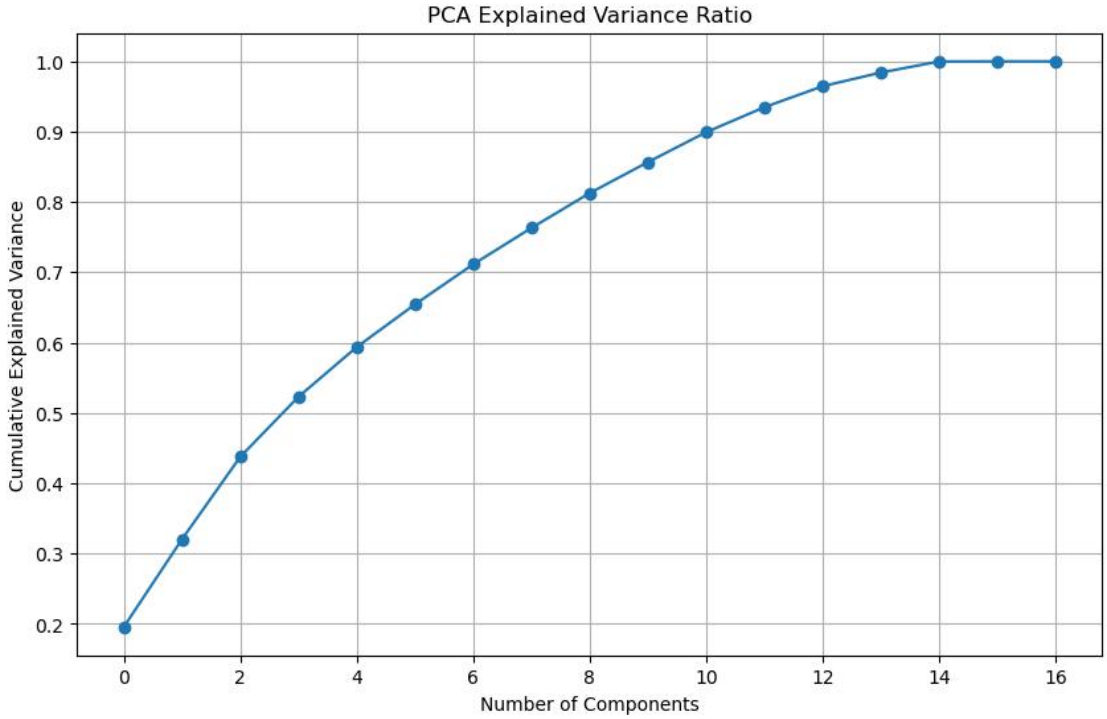

Figure 12: PCA Explained Variance Ratio

CHAPTER FOUR

4.0 Modeling & Evaluation

4.1 Model Selection

A Random Forest Classifier was chosen due to its high accuracy, robustness to noise, ability to manage both categorical and continuous variables, and resistance to overfitting. Moreover, it supports feature importance metrics, making it compatible with interpretability tools like SHAP.

4.2 Hyperparameter Tuning

To enhance model performance and prevent underfitting or overfitting, hyperparameter tuning was conducted using GridSearchCV. Several parameters were explored:

i. n_estimators: Number of trees (100, 200)
ii. max_depth: Tree depth (5, 10, None)
iii. min_samples_split: Minimum samples for node split (2, 5)
iv. min_samples_leaf: Minimum samples at a leaf node (1, 2)
v. bootstrap: Sampling strategy (True, False)

A 5-fold cross-validation was employed to validate each parameter combination. The best model was selected based on accuracy scores.

4.3 Model Performance

The final model achieved the following performance metrics on the test set:

- ✓ **Test Accuracy**: 99%
- ✓ **Macro F1-Score**: 0.99
- ✓ **Confusion Matrix**: Near-perfect predictions, with only 2 misclassified cases

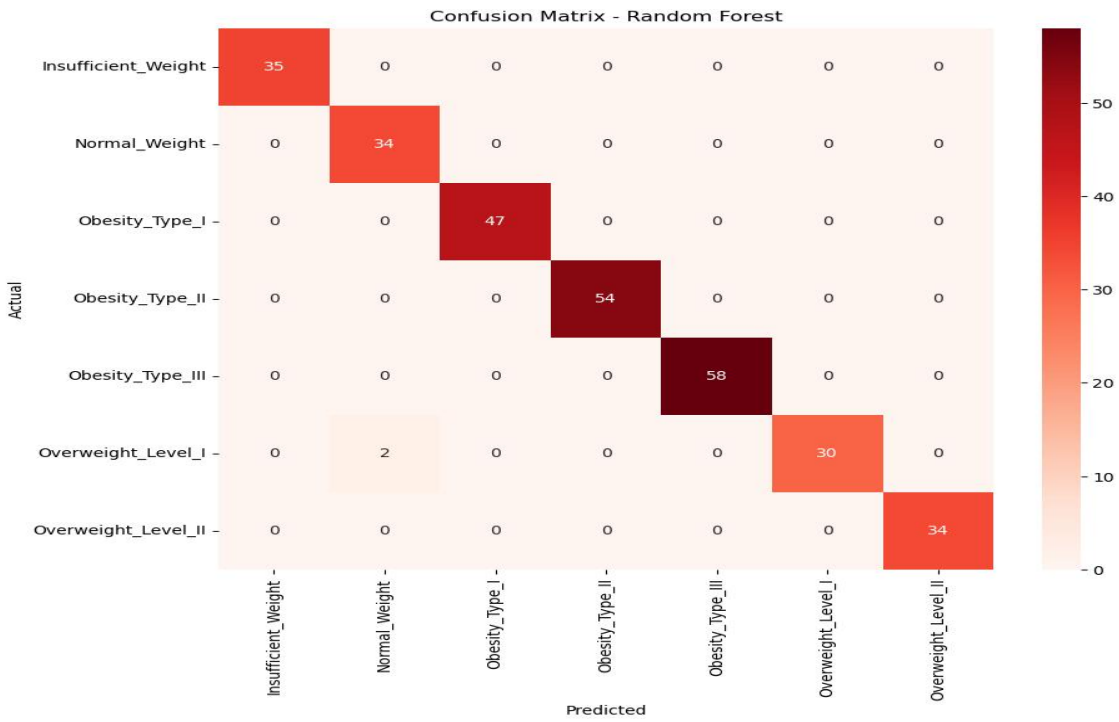

Figure 13: Confusion Matrix - Random Forest Accuracy

This high level of accuracy and minimal misclassification suggest the model generalizes well across multiple obesity classes, including less frequent ones.

4.4 Model Interpretability: SHAP Analysis

To improve model transparency, SHAP values were used to quantify the contribution of each feature to the model's predictions. The SHAP summary plot highlighted the most influential features:

- Weight: Strong positive influence on higher obesity categories,
- FAF: Inversely related to obesity.
- Height: Negatively associated with obesity classification,
- Age: Moderate but consistent influence across classes,

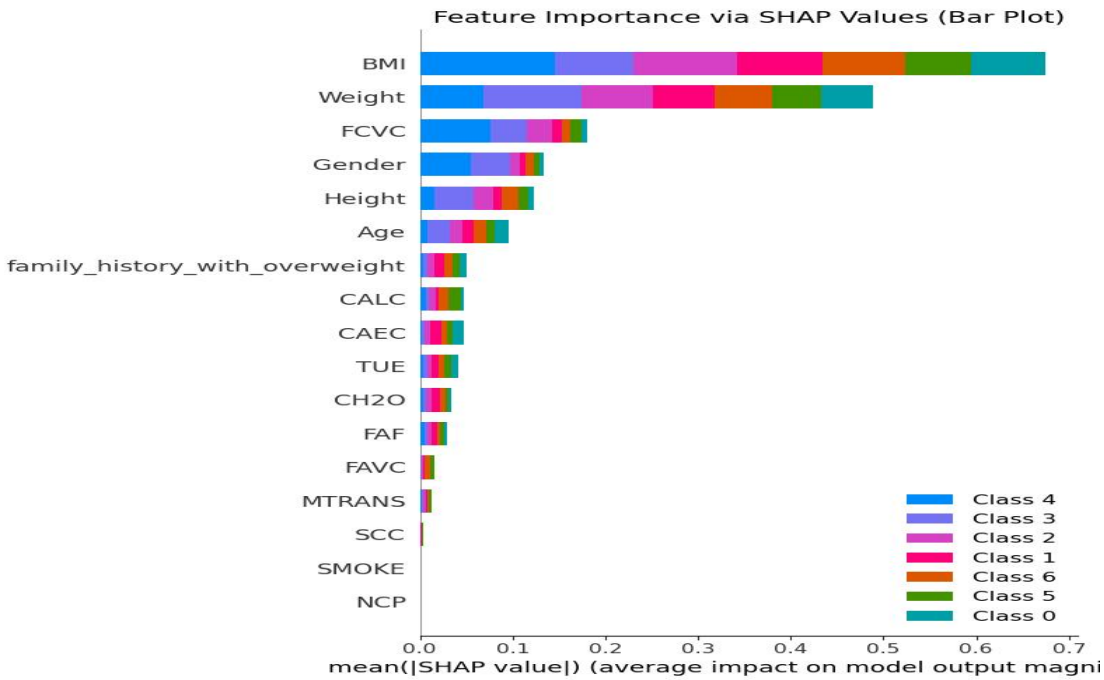

Figure 14: Feature Importance via SHAP Values (Bar Plot)

These results were consistent with medical research, lending further credibility to the model's logic. SHAP made the model's decision-making process accessible and explainable, a key requirement in healthcare-related applications (Lundberg, S.M. et.al, 2017).

CHAPTER FIVE

Conclusion

This project demonstrates the successful application of machine learning in predicting obesity levels using demographic, behavioral, and physiological data. Through a carefully structured pipeline that included preprocessing, EDA, modeling, hyperparameter tuning, and interpretability analysis, the study achieved high prediction accuracy and valuable health insights. The Random Forest classifier proved effective in both performance and explainability. SHAP analysis validated that the model's internal logic was consistent with known health risk factors.

The findings of this work have real-world applicability, especially in the design of digital health platforms aimed at monitoring obesity risks. By integrating these predictive insights into healthcare workflows, early intervention strategies can be implemented, potentially reducing the long-term health burden of obesity.

The successful implementation and performance of the Random Forest classifier, alongside other supervised learning models in this study, were largely supported by tools from the Scikit-learn library, an essential framework for building and evaluating machine learning models efficiently in Python (Pedregosa et al., 2011).

REFERENCES

1. Lundberg, S.M. and Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. In: Advances in Neural Information Processing Systems. Available at: https://proceedings.neurips.cc/paper_files/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html [Accessed 2025].
2. Nuttall, F.Q. (2015). *Body Mass Index: Obesity, BMI, and Health: A Critical Review*. Nutrition Today, 50(3), pp.117–128. Available at: https://doi.org/10.1097/NT.0000000000000092 [Accessed 2025].
3. Pedregosa, F. et al. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12, pp.2825–2830. Available at: http://jmlr.org/papers/v12/pedregosa11a.html [Accessed 2025].
4. Python Libraries: Pandas, Seaborn, Matplotlib — for data visualization and manipulation.
5. Rajkomar, A., Dean, J. and Kohane, I. (2019). Machine Learning in Medicine. New England Journal of Medicine, 380(14), pp.1347–1358. Available at: https://doi.org/10.1056/NEJMra1814259 [Accessed 2025]
6. Scikit-learn Documentation. Available at: https://scikit-learn.org [Accessed 2025].
7. SHAP Documentation. SHapley Additive Explanations. Available at: https://shap.readthedocs.io [Accessed 2025].
8. UCI Machine Learning Repository. Obesity Levels Dataset. Available at: https://archive.ics.uci.edu/dataset/544/estimation+of+obesity+levels+based+on+eating+habits+and+physical+condition [Accessed 2025].
9. World Health Organization (WHO) (2023) Obesity and overweight. Available at: https://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight [Accessed 2025].