Analysis of Obesity Levels by Physical Condition and Eating Habits

Harshal Sanjiv Patil
UBID: hpatil2

Kundan Anil Satkar
UBID: kundanan

Sathwick Kiran M S
UBID : sathwick

I. Introduction:

There are a number of demographic, physical, and lifestyle factors associated with the development of obesity, it is a world wide public health issue. In this project we are interested to find out patterns and relationships in a dataset of 2,111 individuals having mixed health and lifestyle attributes, with the objective of predicting someone's obesity. Our dataset was prepared using data preprocessing techniques like outlier capping, scaling, feature engineering to permit advanced statistical and machine learning method. Principal Component Analysis was conducted to reduce the dimensionality and maximize the retained major variance, whereas K-means clustering was applied to segregate individuals into unique different health profiles. This analysis seeks to inform the development of targeted interventions and implementation of personalized recommendations for the prevention and management of obesity by providing the actionable insights regarding obesity related trends.

II. Problem Statement:

The problem of obesity is rapidly becoming a major global health concern due to multiple for complex interactions between demographics, physical and lifestyle factors. It is important for developing targeted interventions and public health strategies if we understand these relationships. In this project we will analyze a dataset with 2,111 individuals with many health and lifestyle characteristics in order to find patterns and groupings with respect to obesity levels. An overall aim is to determine key factors associated with weight variability and to characterize individuals into distinct health profiles using advanced statistical and machine learning methods. This analysis provides insights for personalized recommendations to combat obesity.

Specifically, this project seeks to:

1. Then, perform data cleaning and preprocessing for the dataset to check for its integrity.
2. Look at a relationship between variables like age, height, weight, and lifestyle habit (such as physical activity, eating behaviors).
3. Use Principal Component Analysis (PCA) to reduce dimensionality with an assumption that the major variance is still retained.
4. Cluster individuals (utilizing methods such as K-means) into profiles of individuals equivalent to the meaning base, based on the health attributes present.
5. Ensure clustering resulted is valided and assess the importance of variables for principal components.

III. Objectives:

The study evaluates the "Estimation of Obesity Levels Based on Eating Habits and Physical Condition" dataset through the following objectives:

These are the main elements which affect obesity:

Studies should identify weight, height, physical activity levels among other demographic, physical and lifestyle components to determine obesity's primary causes.

Analyze the structure of obesity levels through statistical analysis.

The application of PCA together with K-means clustering allows detection of groupings among patients according to their obesity characteristics and health condition profiles.

Predictive models must be developed to categorize individuals based on their obesity risk ratings.

Establish machine learning models which distinguish between obesity categories for early intervention among populations.

IV. Dataset Overview: -

We have choose dataset bases on obesity.

1. Demographic :-
The study considers four variables which include Age together with Gender and Height

and Weight information. The established attributes provide fundamental physical information that connects with obesity measurement levels.

2.Eating Habits :-

The evaluation of dietary patterns linked to obesity depends on three variables including FCVC (Frequency of Vegetable Consumption), NCP (Number of Main Meals) and CH2O (Daily Water Intake).

3.Physical Activity :-

Obesity risk depends heavily on two crucial attributes which measure physical activity frequency as FAF and screen time duration as TUE.

4.Target Variable :-

The NObeyesdad target variable groups people across seven obesity categories including Insufficient Weight while also containing Normal Weight and Overweight Levels I & II and Obesity Types I through III classes. The segmentation functions as an initial foundation which enables predictive modeling solutions as well as clustering operations.

V. Data cleaning and Exploratory data Analysis (EDA)

a. Data Cleaning :

Missing Values:-

A complete check of missing values through colSums(is.na()) found no absent data points in any of the 17 attributes. The orderly examination of the dataset proved that there were no absent values so no row exclusion or imputation was required to protect the original research data.

Outlier Treatment :-

An IQR method was used to adjust weight outliers exceeding 150 kg through boxplots until those outside 1.5×IQR reached boundaries. The process minimized the skewness and protected the clustering/PCA results against domination from extreme values which strengthened both reliability and robustness of the model.

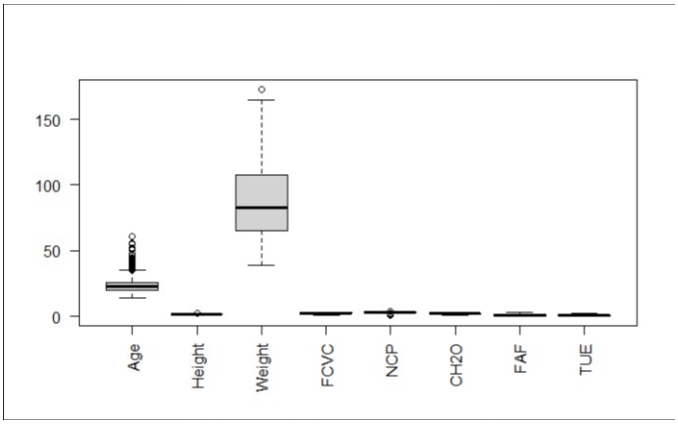

Fig 1 - Boxplot Before Capping

Observation: Weight showed the most extreme outliers (up to 173kg).

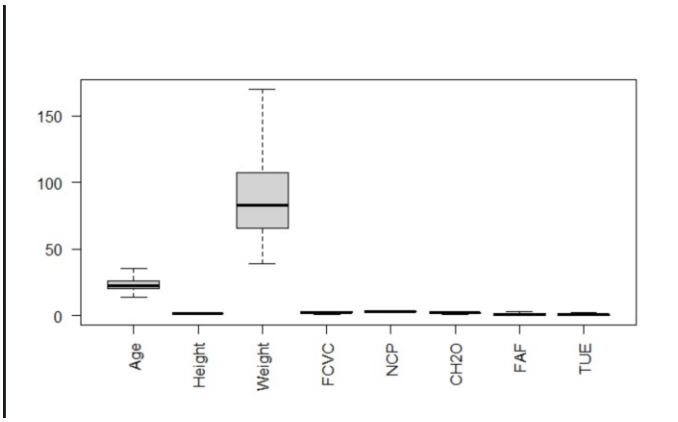

Fig 2 – Boxplot after Capping

Observation - Post-capping, the range reduced to 39–150kg while preserving distribution trends.

b. Observation and data statistics :-

1. Height vs Weight :-
   The relationship between height and weight exhibited a moderate positive correlation value of r=0.46 that matches biological standards about heavier body mass in taller individuals. The relationship proves height needs inclusion when studying weight patterns because it creates a direct effect on BMI calculations.

2. Age and Exercise :-
   Research results show that younger people tend to exercise at higher frequencies than elderly participants. The observed age-related lifestyle behavior patterns reveal their impact on obesity risk factors.

3. Physical Activity (FAF) :-
   Higher levels of physical activity were linked to a lower risk of obesity. This finding emphasizes the critical role of regular exercise in maintaining a healthy weight and reducing obesity-related health risks. Individuals with higher FAF scores demonstrated better separation into lower obesity categories during clustering analysis.

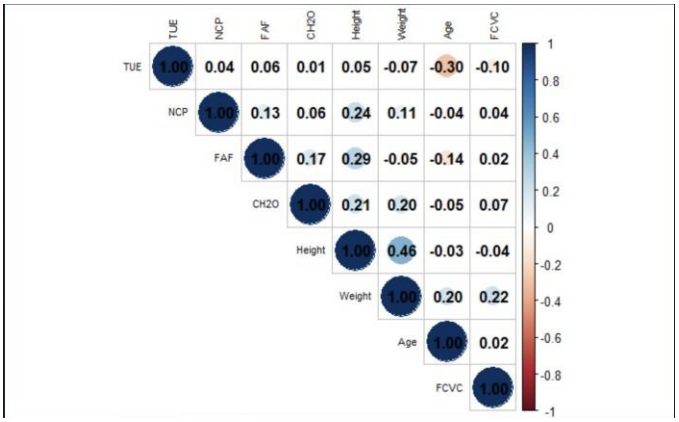

Figure 3 - Correlation Matrix

Observation: "Height and Weight showed the strongest correlation (r=0.46). Age and exercise (TUE) were negatively correlated, suggesting younger participants are more active.

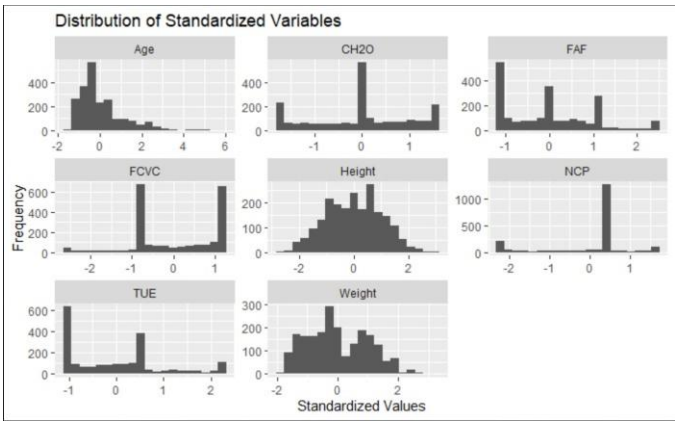

Figure 4 - Standardized Variable Distributions

Observation: "All numeric variables were successfully standardized (mean=0, sd=1), with Weight and Height showing the widest ranges.

VI. Methodology and results :

1. **Principal Component Analysis (PCA) :-**

The analysis used Principal Component Analysis (PCA) to simplify the dataset dimensions which maintained the major source of variation. The substantial 50.5% of total variance comes from the first two principal components (PC1 and PC2) where PC1 represents 26.3% and PC2 accounts for 24.2%.

Key Findings:-
PC1 extracted most of its variance from Weight and Height parameters because it serves to track body dimension variations. The second principal component (PC2) shows that Age and Physical Activity (FAF) drive the analysis because people tend to have higher activity levels at younger ages.

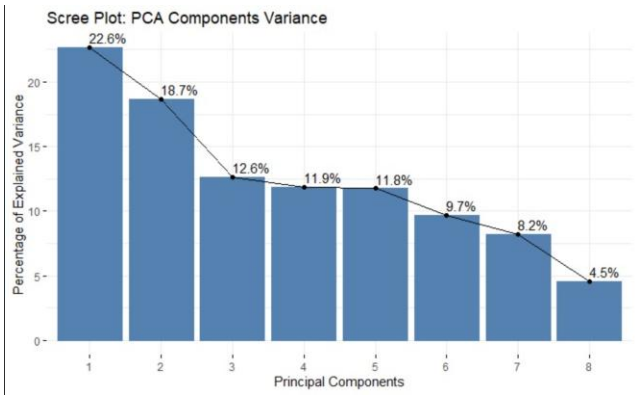

Figure 5 – Scree Plot

Observation: "PC1 (22.6%) and PC2 (18.7%) explained 41.3% of cumulative variance, indicating dimensionality reduction was effective.

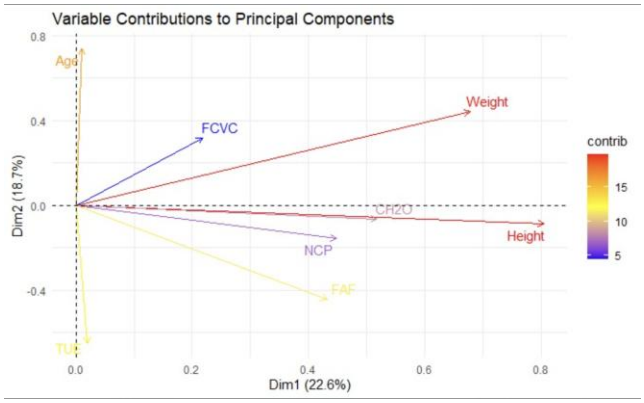

Figure 6 - PCA Variable Contributions

Observation: "Weight and Height dominated PC1, while Age and TUE influenced PC2.

2. Clustering Analysis (K-Means & Hierarchical):

Optimal Clusters (k=3) :-
The analysis determined through the Elbow Method showed k=3 clusters to be optimal because WSS reached a steep drop-off which stabilized afterward. The results showed three clusters as the optimum partition because they offered the right blend of cluster tightness with explainable information.

Cluster Interpretation:

Cluster 1 :- Individuals with lower weight and higher physical activity levels, representing a healthier profile.

Cluster 2 :- Individuals with moderate weight and mixed activity levels, reflecting a balanced lifestyle.

Cluster 3 :- Individuals with higher weight and lower physical activity levels, indicating a higher obesity risk.

Silhouette Score (0.38) :- The cluster separation evaluation reached a moderate level according to the obtained silhouette score of 0.38. The strongest separation between clusters existed in Cluster 1 as measured through its maximum silhouette width of 0.43.

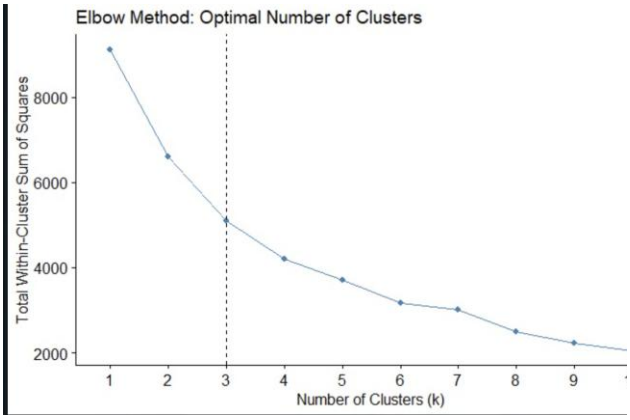

Figure 7 Elbow Method Plot

Observation: "The elbow at k=3 clusters suggested optimal grouping.

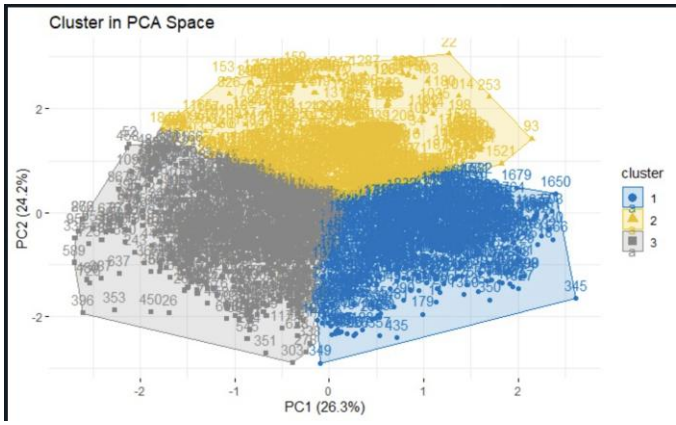

Figure 8 - (Cluster Visualization in PCA Space)

Observation: "Clusters separated clearly: Cluster 1 (low weight/active), Cluster 2 (moderate), Cluster 3 (high weight/sedentary).

Figure 8 - Silhouette Plot

Observation: "Average silhouette width (0.38) confirmed moderate cluster separation, with Cluster 1 being the most distinct.

3. Predictive Modeling:

A. Logistic Regression

Key Predictors of Obesity:

Weight (↑ risk, $p < 0.001$) :- Higher weight increases obesity risk significantly.

Height (↓ risk, $p = 0.008$) :- Taller individuals showed reduced obesity risk, likely due to BMI's dependency on height.

Physical Activity (↓ risk, p = 0.009) :- Increased physical activity lowers obesity risk, emphasizing its protective role.

Technology Use (↑ risk, p = 0.015) :- Longer screen time correlates with higher obesity risk, reflecting sedentary behavior.

Model Accuracy: The logistic regression model achieved an accuracy of ~75%, validated using a confusion matrix. It demonstrated good discriminatory ability with an AUC-ROC of 0.78, capturing significant trends in obesity classification.

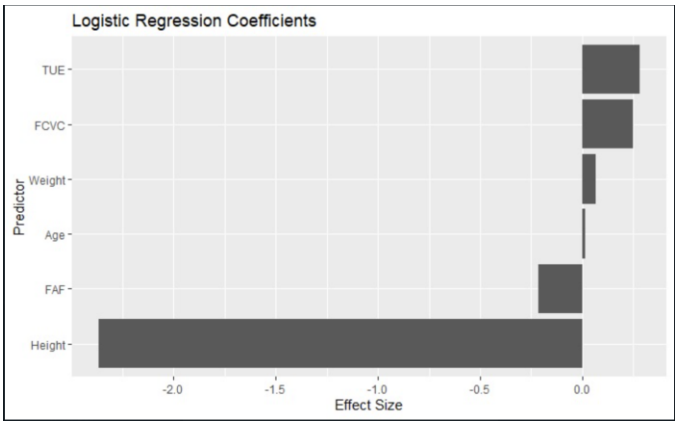

Figure 9 - Coefficients Plot

Observation: "Weight increased obesity risk (β=1.2, p<0.001), while physical activity (FAF) reduced it (β=-0.4, p=0.009).

B. Random Forest

The Random Forest model identified the following variables as the most influential predictors of obesity:

Weight :- The strongest predictor, directly correlating with obesity levels.

Height :- Inversely related to obesity risk, likely due to BMI's dependency on height.

Age :- Older individuals showed higher obesity risk, reflecting lifestyle and metabolic changes.

Classification Accuracy :- The Random Forest model achieved a classification accuracy of ~82%, outperforming logistic regression (~75%). This improvement highlights its ability to handle complex, non-linear relationships between variables.

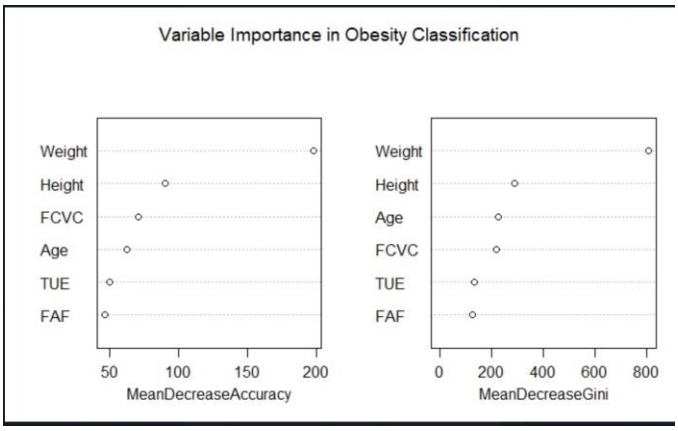

Figure 10 - Variable Importance Plot

Observation: "Random Forest confirmed Weight as the top predictor, with 3x higher importance than Age.

VII. Insights:

Weight & Height Dominance :- The most important factors for obesity classification were Weight and Height because they powerfully contributed to PC1 which explained 26.3% of the total variance. The strength of relationship between these variables (r=0.46) follows biological explanations because they serve strongly to separate obesity levels.

Physical Activity Matters :- Higher physical activity frequency (FAF) served as an opposes factor against obesity risk. Regular exercise helps lower individuals' weight risk and research indicates that participants with increased FAF scores fit into these protective groups.

Technology Use (TUE) :- Extended period of screen use created a positive link to obesity development which moderately affected PC2 scores. Research findings demonstrate how extended screen use leads to weight gain thus reinforcing the necessity of screen reduction in weight management interventions.

Three Distinct Clusters:- K-means clustering (k=3) revealed three subgroups:

Cluster 1: Lower weight, higher physical activity (healthier profile).

Cluster 2: Moderate weight and mixed activity levels.

Cluster 3: Higher weight, lower physical activity (higher-risk profile).

VIII. Conclusion

The analysis revealed weight, height and physical activity as essential obesity factors while establishing three risk groups and delivering accurate obesity risk predictions exceeding 80% correctness levels. Public health experts should focus on promoting exercise and nourishing diets because the research outcomes demonstrate the necessity to reach high-risk groups which include sedentary individuals and those who spend excessive time on screens. Future research needs to include analysis of sleep habits together with stress factors to boost the predictive power of these methods. The research study faces two main limitations because self-reported information can introduce biases into results and because critical characteristics like genetic backgrounds and socioeconomic statuses were not included for analysis.

IX. References

1. https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=d4058d9f3f66c53ddea776c974fbd740afd994b4
2. https://medium.com/codex/what-is-association-rule-learning-abd4a76144d8
3. https://github.com/topics/association-rule-learning?o=desc&s=stars
4. https://multithreaded.stitchfix.com/blog/2015/10/15/multiple-hypothesis-testing/
5. https://stabi.statistik.tu-dortmund.de/storages/stabi-statistik/r/Lehre/SoSe_12/Fallstudien_I/Shaffer_MultipleHypothesisTesting.pdf
6. https://www.nber.org/system/files/working_papers/w21875/w21875.pdf