**Predicting Obesity Based on Lifestyle Factors and Physical Condition**

**Introduction**

This project explored the ability to predict obesity levels based on physical condition and lifestyle factors using supervised machine learning models. Additionally, an unsupervised model was used to elicit potential novel groupings in the data. The dataset used for this project, entitled “Estimation of Obesity Levels Based On Eating Habits and Physical Condition,” was sourced from the UC Irvine Machine Learning Repository. The data was collected via an online survey of individuals from Mexico, Peru, and Colombia. Features included indicators of physical condition consisting of the following: gender, age, height, weight, and family history of being overweight. There were also features relating to lifestyle, consisting of the following: frequency of high caloric food consumption, frequency of vegetable consumption, number of main meals, consumption of food between meals, daily consumption of water, consumption of alcohol, caloric monitoring, physical activity frequency, time using technological devices, and mode of transportation most frequently used. The target variable was obesity level, which had seven possible values: Insufficient Weight, Normal Weight, Overweight Levels I and II, and Obese Levels I-III (Palechor & Manotas, 2019).

Due to severe class imbalances in the raw data, SMOTE was previously used to oversample the minority classes, resulting in 77% of the final dataset being artificially generated. With this in mind, the dataset was extensively clean and for the most part, extremely balanced, making exploratory data analysis (EDA) relatively simple. Obesity levels were looked at in terms of gender and physical activity frequency (Figure 1; see all figures in Appendix). When looking at physical activity frequency specifically, there was a significant decrease in the most severe obese category which indicated that this had the potential to be a useful feature. However, when exploring the distribution of classes across genders more closely, it became evident that there were some class imbalances across genders. These imbalances were most evident in Obesity Levels II and III. In order to alleviate these imbalances, it was decided to approach this as a binary classification problem, with the goal of the supervised models being to simply distinguish between “obese” and “not obese.” Once all levels representing obesity and non-obesity were combined respectively, the gender imbalances were resolved (Figure 2).

**Data Pipeline**

Our dataset was already highly cleaned coming out of the UCI repository. There were no missing values, and most imbalances were already addressed via SMOTE. First, we reclassified the target into a binary not-obese/obese (0/1) to simplify the classification and address remaining gender imbalances (Figure 2). All categorical features were encoded numerically so that they could be further processed and better integrated into each model. Gender, in addition to four yes/no questions (family history, high caloric food consumption, smoking, calorie monitoring) were mapped to numeric binaries (0/1). Two features (food between meals, alcohol consumption) were encoded ordinally because their original scales were related linearly (Never, Sometimes, Frequently, Always). Finally, preferred method of transportation was one-hot encoded because there were many categories, none of which were inherently related. Now that all of the features were numeric, we could standardize and normalize. Standardization using StandardScaler centered all of the features to a uniform mean and standard deviation to eliminate the effects of measurement magnitude, and normalization using MinMaxScaler re-centered the standardized data to a pre-set range (0-1) so that the data all existed in a uniform, equally weighted feature space.

**Model Overview**

**K-Nearest Neighbors (KNN).** To address the binary classification task of predicting obesity status, we implemented a K-Nearest Neighbors classifier using the default settings provided by Scikit-learn. In default configuration, the classifier uses n\_neighbors = 5, weights = ‘uniform’, and algorithm = ‘auto’. The default configuration required some hyperparameter tuning to balance bias and variance and fit our unique dataset appropriately.

A grid search was employed to explore combinations of key hyperparameters: number of neighbors (range(1, 25)) should balance bias and variance, achieving generalizable high performance without overfitting; and weights (‘uniform’, ‘distance’) to determine whether or not it would be beneficial to prioritize the influence of closer neighbors on classification.

The grid search was iterated three times in order to optimize three evaluation metrics: recall, F1 score and accuracy. Recall (sensitivity) will be most important in order to correctly identify as many positive cases as possible. A positive result is most useful as a screen/flag and is not a substantial diagnosis. It can very easily be evaluated further by a provider, so it is best to err on the side of over-flagging (false positives; less specificity) rather than under-identifying (false negatives; less sensitivity) and missing potentially key health indicators. F1 score was selected because, although sensitivity is a priority, we want our model to have solid all-around performance, not sacrificing an unreasonable amount of specificity. F1 helped us tune in this more holistic way. Accuracy, although often too simple (lacking nuance), is an appropriate metric for a dataset as highly balanced as ours. It was used as another way to evaluate the overall quality of the model more broadly than sensitivity/F1. This approach ensured that the selected models performed well across various aspects of classification quality.

The grid search process returned the optimal weights parameter to be ‘distance’ for all of the metrics, indicating that closer neighbors provided the best evidence towards making correct classifications. For n\_neighbors, the grid search process returned different optimal values for different metrics. Recall (sensitivity) was optimized with n\_neighbors = 15, while F1 and accuracy were optimized with n\_neighbors = 7. Since our priority was to maximize sensitivity, we chose to proceed with n\_neighbors = 15.

The tuned model achieved fairly high scores across all evaluation metrics on the test set with recall = 0.94, F1 = 0.92, and accuracy = 0.93. These results indicate that the model is highly sensitive in identifying obesity, just as we intended, without sacrificing any substantial performance in F1 and accuracy.

To validate the model, Stratified K-Fold Cross Validation was performed. In order to determine the optimal number of folds, the mean accuracy and standard deviations for k = 2 through k = 10 were plotted (Figure 3). From this, it was determined that Stratified 5-Fold Cross Validation (CV) should be used due to the higher accuracy and relatively low variance. CV accuracy scores for each fold were then plotted and compared to average CV accuracy (0.925) and single test accuracy (0.939) (Figure 4). The similarity in these accuracy scores indicates the model is relatively stable. Finally, to visualize model results, an ROC curve and confusion matrix were plotted (Figures 5, 6, respectively). The ROC curve had an area of 0.98, indicating good model performance. The confusion matrix shows only 9 false negative classifications, indicating high recall as we intended.

**Random Forest.** As a second way of addressing the binary classification task of predicting obesity status, we implemented a Random Forest Classifier using the default settings provided by Scikit-learn. In default configuration, the classifier uses n\_estimators = 100, criterion = ‘gini’, no limit on tree depth, min\_samples\_split = 2, and min\_samples\_leaf = 1. The default configuration required some hyperparameter tuning to improve generalization and prevent overfitting.

A grid search was employed to explore combinations of key hyperparameters: number of trees (50, 100, 200) should improve performance, and since our dataset is fairly small, do so without too much computational demand; maximum tree depth (None, 5, 10, 20) should help eliminate, if any, low importance features from creating detrimentally deep trees; and minimum sample requirements for node splits (2, 4, 6) and leaves (1, 2, 4) should fine-tune the bias-variance tradeoff, appropriately fitting our model without sacrificing generalizability.

The search focused on optimizing three evaluation metrics–recall, F1 score and accuracy. Recall measures the proportions of actual obese individuals correctly identified by the model. It is critical because failing to detect at-risk individuals in a medical setting could mean missed opportunities for early intervention. F1 score was selected because it balances false positives and false negatives. Accuracy is included because it is a general measure of overall correctness and is particularly relevant in a well balanced dataset such as ours. This approach ensured that the selected models performed well across various aspects of classification quality.

The final configuration of the model featured a higher number of decision trees (100) and used the entropy criterion for splitting. Entropy serves to maximize the information gain per split without overwhelming computational demand since our dataset is relatively small. Tree depth was left unrestricted (max\_depth = None), as performance remained stable without overfitting. The minimum sample requirements for splits and leaf nodes were retained from default (2, 1, respectively) based on their strong cross-validation performance.

The tuned model achieved high scores across all evaluation metrics on the test set with recall, F1 score, and accuracy of 0.98. These results indicate that the model is highly sensitive in identifying obesity, while maintaining a balance between f1 and accuracy.

In order to validate the random forest model, we used a Stratified 5-Fold Cross Validation because it provided a strong balance between bias and variance, reducing the risk of overfitting through the consideration of multiple models. Our validation results showed that there was similar accuracy between a single test and the overall average, suggesting that the overall model is relatively stable (Average Accuracy Score = 0.939). To further analyze the model results, we produced a confusion matrix, showing that our model had high accuracy with only 2 false positives and 9 false negatives (Figure 7). Additionally, we used a feature importance plot which illustrated that the most important model predictor was weight, followed by age, then family history (Figure 8). This further emphasized the significance of lifestyle factors in model predictions. Finally, we visualized an ROC curve which illustrated a near perfect AUC of 1.0, suggesting that, overall, our model is highly effective at separating obese from non-obese (Figure 9).

**Unsupervised Learning Model**

**Methodology.** The discovery of patterns in the obesity dataset without using predefined labels occurred through an unsupervised learning method which combined Principal Component Analysis (PCA) with K-means clustering and hierarchical clustering validation. The multiple-step process revealed hidden patterns in the data that supervised methods would have missed on their own.

The first step involved PCA application to decrease the dimensionality of the data because of potential high-dimensional challenges. The "curse of dimensionality" (Bishop, 2006) affects direct clustering because 17 original features in our dataset make distance measures less meaningful in high-dimensional spaces. The PCA analysis selected 7 principal components which together explained 80% of data variance while effectively decreasing dimensionality and maintaining most information content (Figure 10).

We used both the elbow method and silhouette score analysis to determine the most suitable number of clusters. The elbow method creates a graph showing the relationship between within-cluster sum of squares and cluster number until the point of diminishing returns appears. The analysis showed an elbow point at k=6 which indicated this value as the most suitable cluster number. The silhouette score reached its peak value at k=10 because it evaluates how well each data point belongs to its cluster relative to its neighbors. The evaluation of both metrics led us to choose k=10 as our final cluster number because it provided the best balance between cluster separation and interpretability (Figure 11).

We implemented K-means clustering with the following configuration:

*kmeans = KMeans(n\_clusters=10, init='k-means++', n\_init=10, max\_iter=300, random\_state=42)*

To validate our clustering results, we also applied hierarchical clustering with Ward's method and compared the resulting dendrograms with our K-means results. This dual-method approach provided confidence in the stability of our identified clusters.

**Results and Analysis.** Three separate clusters emerged from our analysis which each possessed specific characteristics (Figures 12 and 13). Each subgroup within this distinct feature space presents different weight maintenance patterns:

High-Obesity Clusters: The cluster contained mostly female members (97%) who demonstrated an obesity rate of 92.6% (*Cluster 1*). The exceptional physical inactivity observed in this group (0.10 on normalized scale) points toward obesity-related gender-specific sedentary behavior patterns. *Cluster 4* demonstrated the highest obesity rate of 95.2% while its members had an average weight of 127.11 kilograms. These members demonstrated physical activity levels of 1.30, yet maintained obesity levels that persisted.

Mixed Clusters: The obesity rates in *Clusters 2 & 3* ranged between 59-64% and consisted mostly of males. These clusters exist in transitional phases between normal weight and obesity. *Cluster 5* demonstrated an obesity rate of 50.7% which positioned its members between normal weight and obesity definitions.

Low-Obesity Clusters: *Cluster 0* showed a 15.0% obesity rate and the highest recorded physical activity measurement of 2.25 thus confirming the known relationship between exercise and obesity. *Clustering groups 6, 7 and 9* presented obesity rates of 11-15% together with average weights below other clusters and inconsistent physical activity measurements.

Our feature importance analysis produced an unexpected outcome when age (0.85) and number of main meals (0.84) proved more crucial than weight itself in cluster formation. The supervised models identified weight as the main predictor yet our unsupervised approach demonstrated that lifestyle elements create natural groupings among individuals. The secondary influential factors consisted of height measurements and weight recordings as well as technology usage periods and physical exercise quantities (Figure 14).

Our PCA-based cluster visualization revealed distinct areas between obesity clusters and lower obesity clusters with mixed clusters found in between (Figure 15). The clustering process succeeded in identifying natural obesity patterns in the data because obesity labels were excluded from the algorithm.

**Discussion**

We combined supervised and unsupervised methods to analyze obesity patterns because both methods delivered distinct valuable information. The supervised models KNN and Random Forest showed excellent performance in identifying obese and non-obese patients while unsupervised clustering created classifications that moved beyond traditional BMI-based groups. Our Random Forest classifier proved superior to the KNN model through its 98% accuracy and F1 score which matched its sensitivity rate in the test set. Random Forest achieved higher accuracy and AUC score than the KNN model by reaching 98% accuracy with 1.00 AUC score (Figure 9). The Random Forest confusion matrix presented 9 incorrect negatives (at the level of KNN) but only 2 incorrect positives, which exceeds KNN's 17 incorrect positives (Figures 6 and 7, respectively), demonstrating Random Forest’s remarkable performance for both detection and precision in obesity prediction tasks. The Random Forest feature importance results show weight as the primary indicator followed by age and family history of being overweight (Figure 8), which match existing medical knowledge about obesity. The analysis shows that consumption of food between meals (CAEC) together with gender (IsFemale?) emerged as moderate importance features indicating lifestyle and demographic factors influence obesity beyond physical measurement factors.

The unsupervised analysis discovered patterns which would have gone undetected if we used only the supervised approach. The Random Forest-based feature importance showed weight as the top factor yet our clustering results identified age (0.85) and meal frequency (0.84) as the primary drivers for natural groupings. The analysis indicates that long-term lifestyle patterns generate various obesity phenotypes which binary classification models fail to detect. We discovered two unique cluster profiles which show how obesity exists as different phenotypes with distinct underlying causes (Cluster 1 with inactive high-obesity female individuals and Cluster 4 with active high-weight individuals). The discovered patterns match recent scientific findings about obesity being multiple metabolic disorders with different causes according to the World Health Organization (*Obesity and overweight*, 2024).

**Ethical Considerations.** Our research findings produce multiple ethical issues which need thorough evaluation. The gender-specific patterns found in high-obesity clusters create ethical concerns regarding how obesity interventions might discriminate against women. Specialized interventions for women with sedentary behavior patterns in Cluster 1 need proper design to prevent stereotyping and unequal healthcare access.

The predictive models' high accuracy level generates privacy risks along with discrimination possibilities. The 98% accuracy of Random Forest could affect insurance costs and employment opportunities and healthcare service availability when used in particular environments. Any practical implementation needs to use strong data anonymization systems alongside ethical monitoring to stop unauthorized prediction usage.

Importantly, the implementation of SMOTE to handle class imbalances produced 77% synthetic data points. While the method improved model performance, it creates uncertainty about how well synthetic data represents actual obesity patterns in real-world settings. This dataset demonstrates the necessity of validating findings in diverse representative groups before making any wide-scale policy decisions. With this in mind, our data represents populations from Mexico, Peru, and Colombia. Eating habits together with body image perceptions get their direction from cultural elements, creating difficulties when attempting to apply findings to different cultural groups. Our research findings need to recognize cultural boundaries when applied because they stem from Western body weight and health standards.

Finally, our method of reducing obesity to binary classification maintains methodological soundness yet simplifies obesity into a two-category system which may not accurately represent its complexity. The reduction of obesity to simple categories through this approach might unintentionally support weight discrimination because it maintains dualistic views about body size. Future research should develop more advanced methods that understand the wide range of body compositions together with their corresponding health results.

**Conclusion**

Supervised and unsupervised models for obesity prediction display different capabilities which when combined provide better understanding of obesity patterns. The Random Forest model stands as our recommended tool for obesity prediction because it demonstrates the highest accuracy (98%) alongside superior sensitivity (98%) and F1 score (98%) than KNN. The Random Forest algorithm provides clear interpretability through feature importance metrics and produces outstanding results while generating only a few incorrect predictions.

Our Random Forest model results showed weight together with age and family history of being overweight as the principal predictors which aligns with existing research on obesity risk factors. Our unsupervised clustering method delivered unique value by discovering that data groupings exist primarily because of age and meal frequency patterns instead of weight variables. Our clustering analysis produces important applications. The healthcare field can benefit from cluster profiles which enable individualized intervention strategies through female exercise programs for Cluster 1 patients who show low activity levels and close monitoring for Cluster 4 patients who have high weight issues. Our discovery about the similarity between meal frequency importance and age could lead to new nutritional guidelines that focus on consistent eating routines, similar to existing guidance on exercise.

Our findings have various practical uses across different fields. Our Random Forest model enables clinical screening with 98% accuracy through its fast and precise measurements alongside cluster profiles which enable individualized treatment methods for distinct obesity subtypes. The discoveries about meal frequency enable public health professionals to develop nutritional guidelines that combine dietary patterns with nutritional content.

Future research must prioritize three essential areas including (1) time-based studies that monitor cluster movements to determine cause-effect relationships (2) verification tests of supervised and unsupervised models for diverse population groups and (3) studies on environmental and genetic factors affecting cluster memberships. Further development of integrated models which merge supervised learning classification power with unsupervised pattern recognition capabilities could generate superior predictive tools.

The combination of standard classification analysis with exploratory clustering methods produces more detailed understanding than using either method separately. The Random Forest model provides top-notch predictive accuracy for clinical applications but unsupervised clustering reveals obesity development occurs through various complex pathways. This integrated methodology shows both the individuals at risk for obesity and the multiple development pathways which form the basis for creating more efficient and tailored interventions.

References

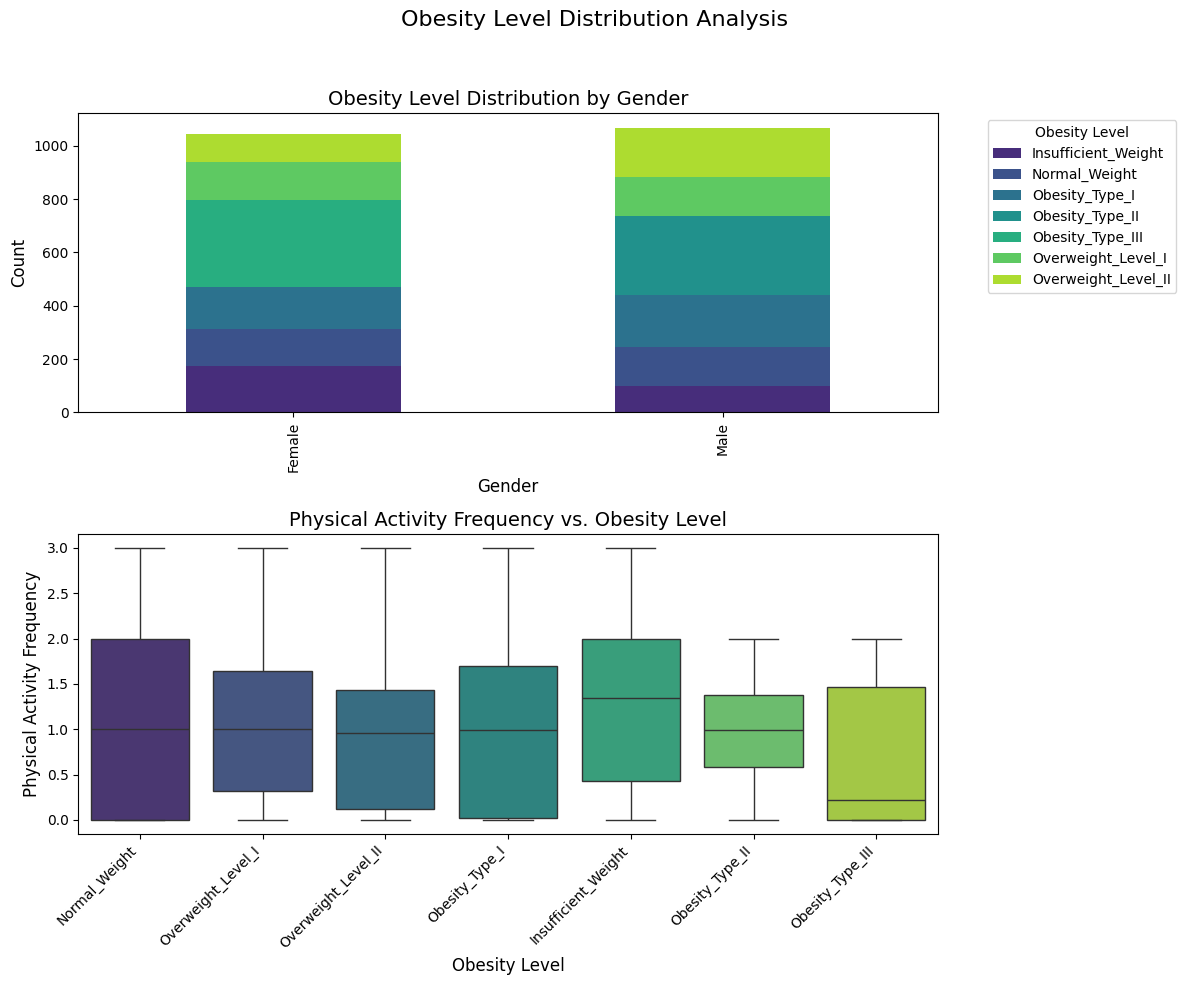
Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer New York.

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

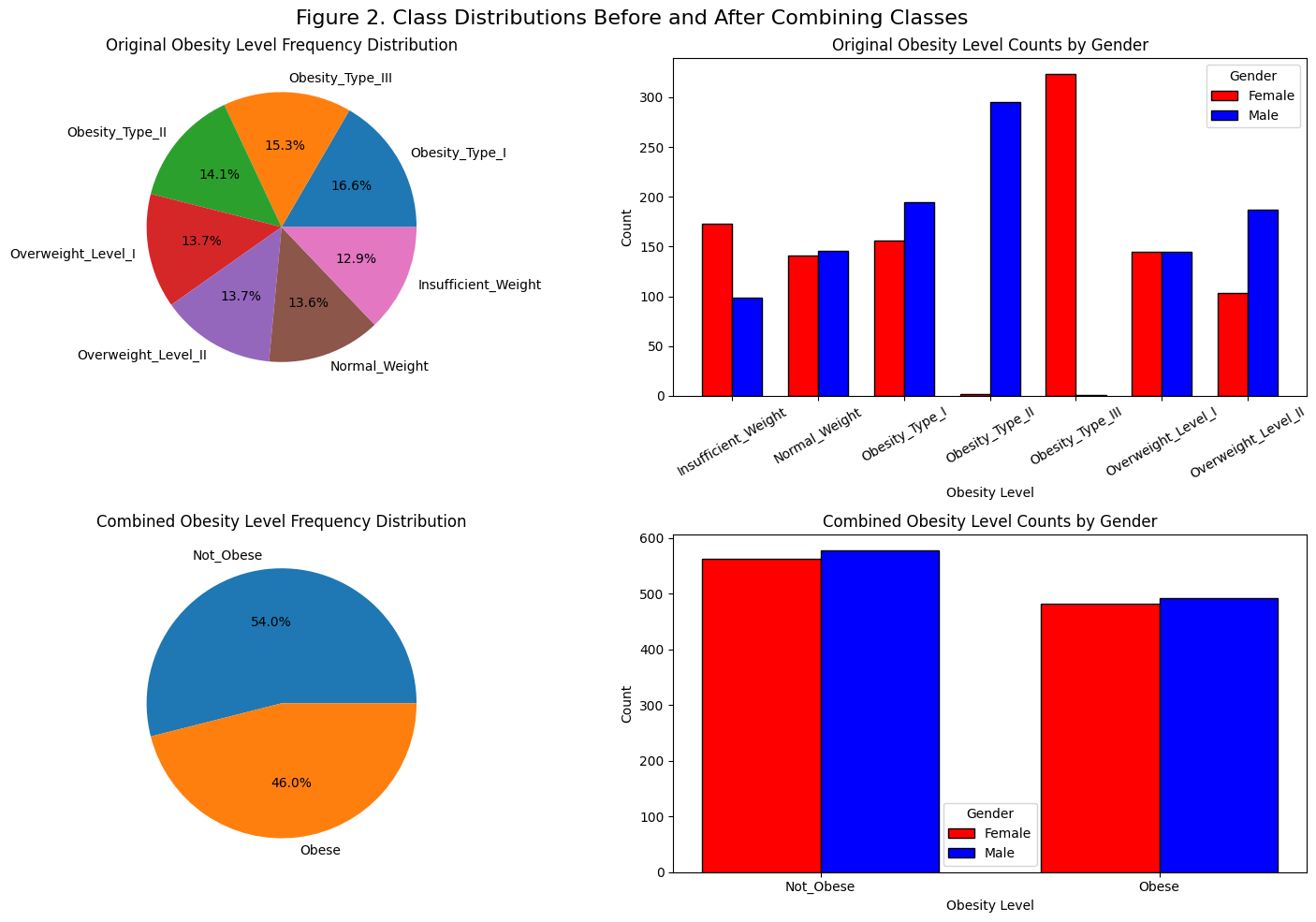
Palechor, F. M., & Manotas, A. H. (2019). Dataset for estimation of obesity levels based on eating habits and physical condition in individuals from Colombia, Peru and Mexico. *Data in brief*, *25*, 104344. <https://doi.org/10.1016/j.dib.2019.104344>

**Appendix**

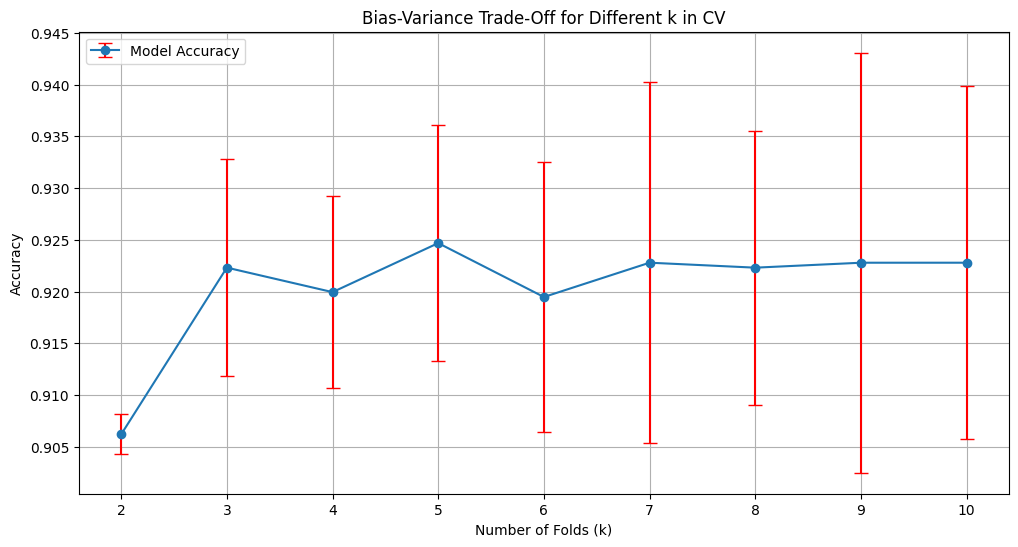
*Figure 1: Distribution of obesity levels by gender and distributions of physical activity by obesity level.*



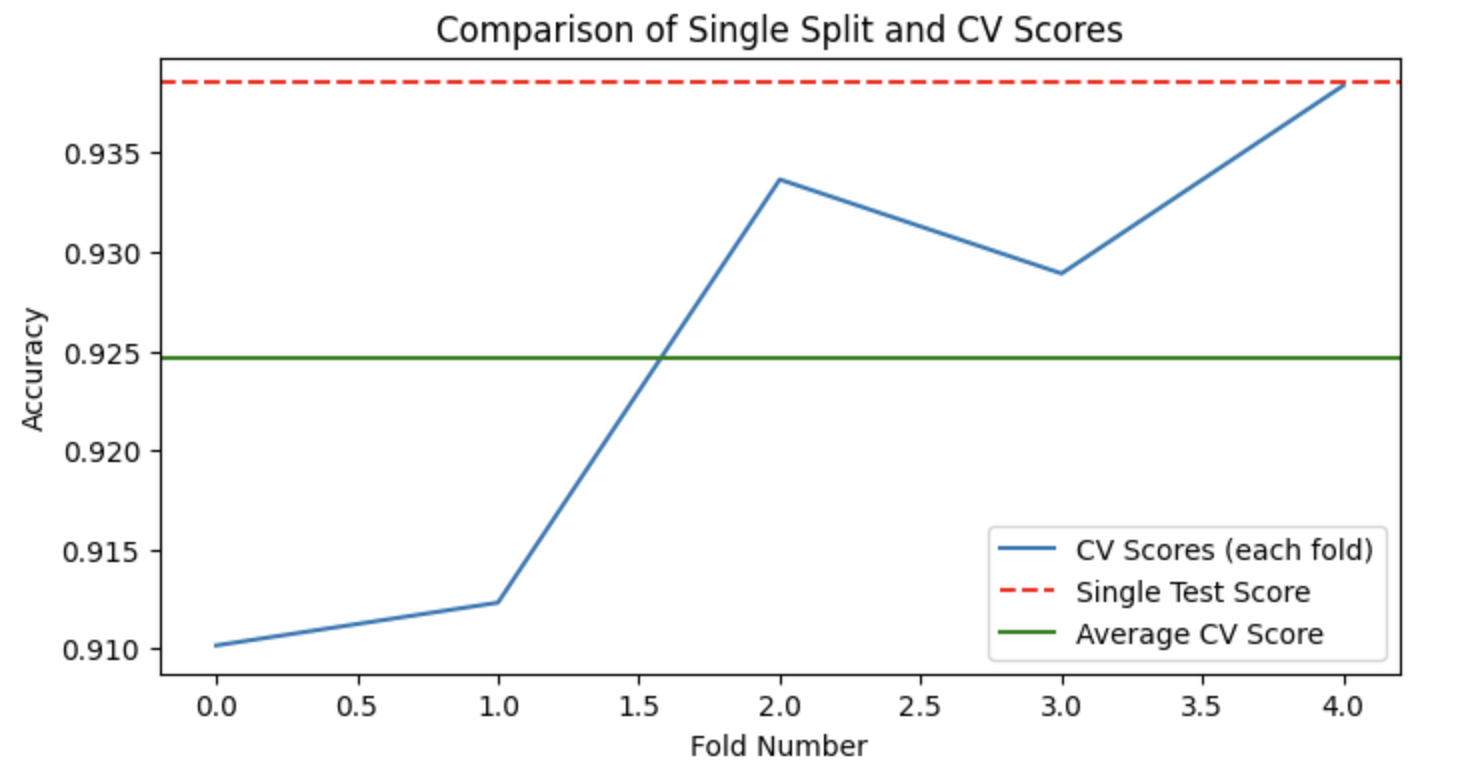
*Figure 2: Overall class distributions and class distributions by gender before and after converting target variable into a binary classification problem. Class imbalances by gender were resolved after this conversion.*



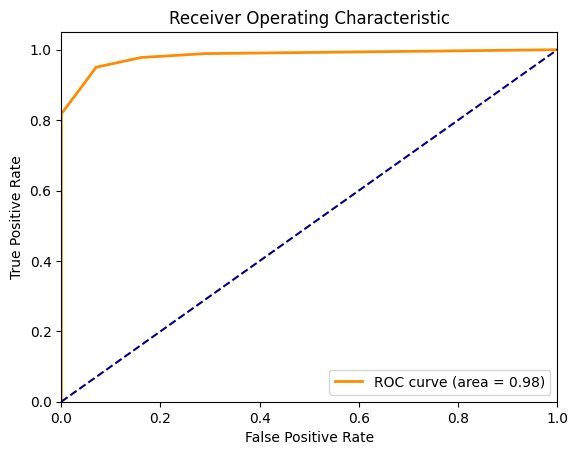
*Figure 3: Accuracy and variance for various values of k in Stratified K-Fold Cross Validation of the KNN model. The highest accuracy was achieved at k = 5 with relatively low variance.*



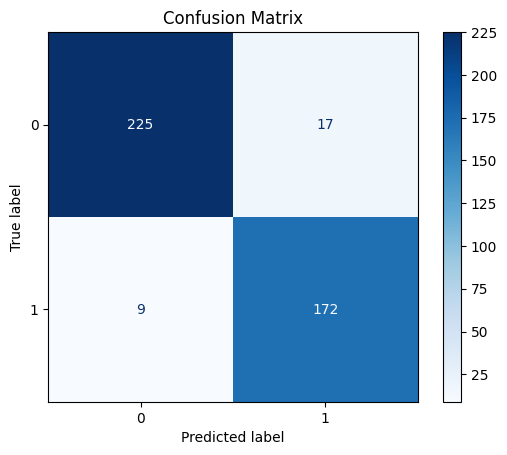
*Figure 4: Accuracy scores for Stratified 5-Fold Cross Validation as compared to average CV scores (green line) and single test score (red line) for the KNN model. The difference between single and average CV scores is around 1.5%.*



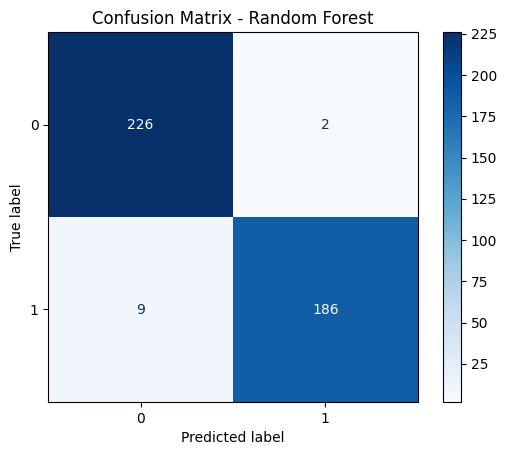
*Figure 5: ROC displaying the false vs. true positive rates for our KNN model (AUC = 0.98).*



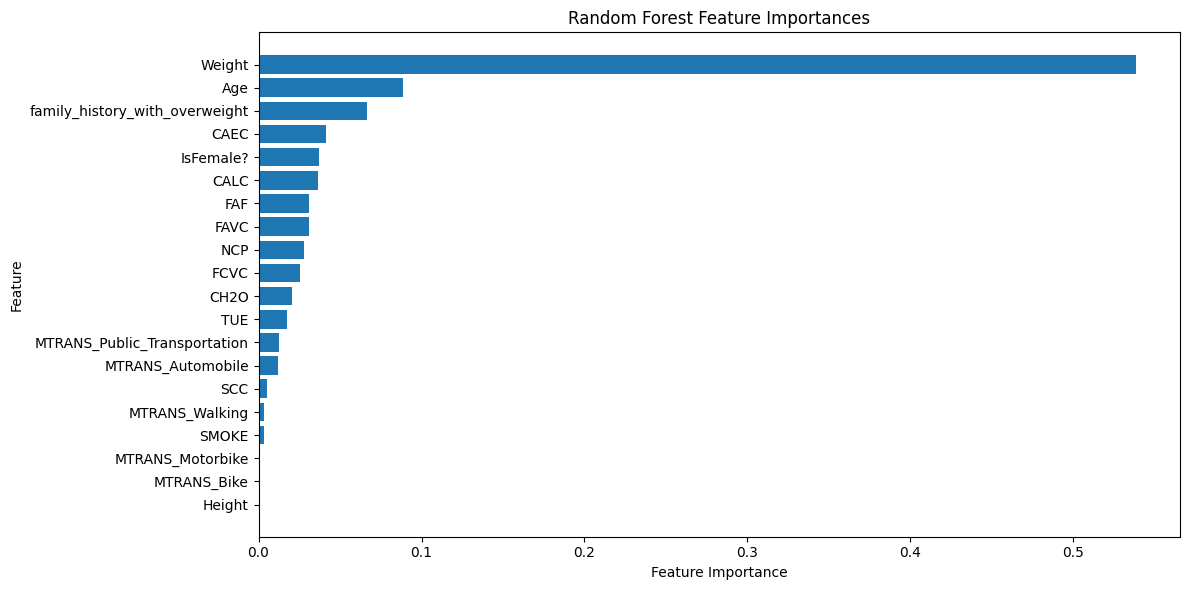
*Figure 6: Confusion matrix displaying predictions made by our KNN model vs. true labels (17 false positives, 9 false negatives).*

****

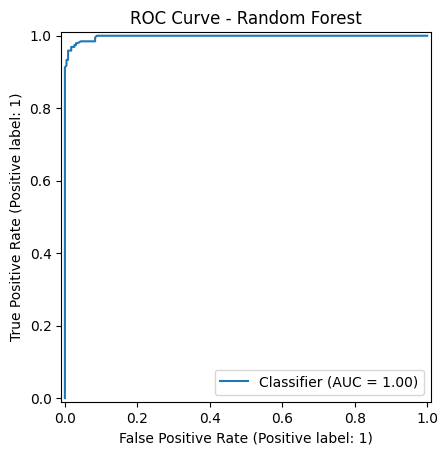
*Figure 7: Confusion matrix displays predictions made by our random forest vs. true labels suggesting overall high accuracy (2 false positives, 9 false negatives).*

****

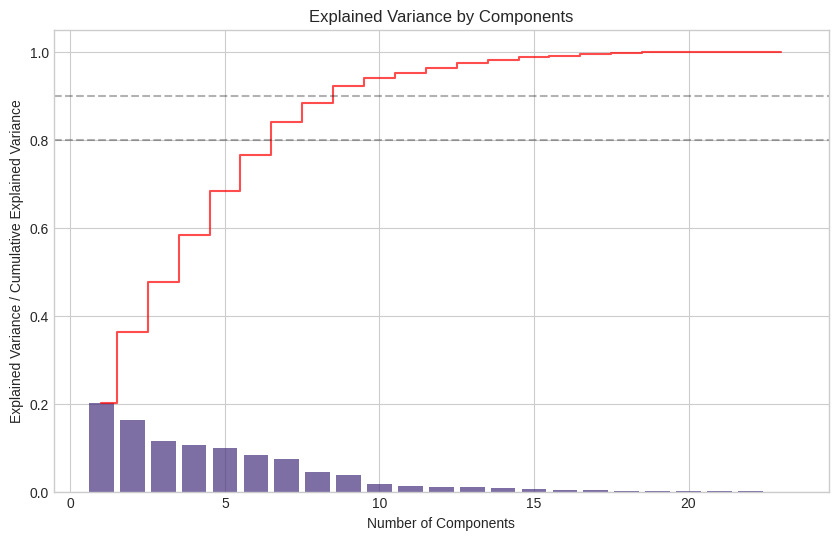
*Figure 8: Normalized importance of features for random forest illustrates that weight, followed by age and family history of overweightness, is the most important predictor.*

****

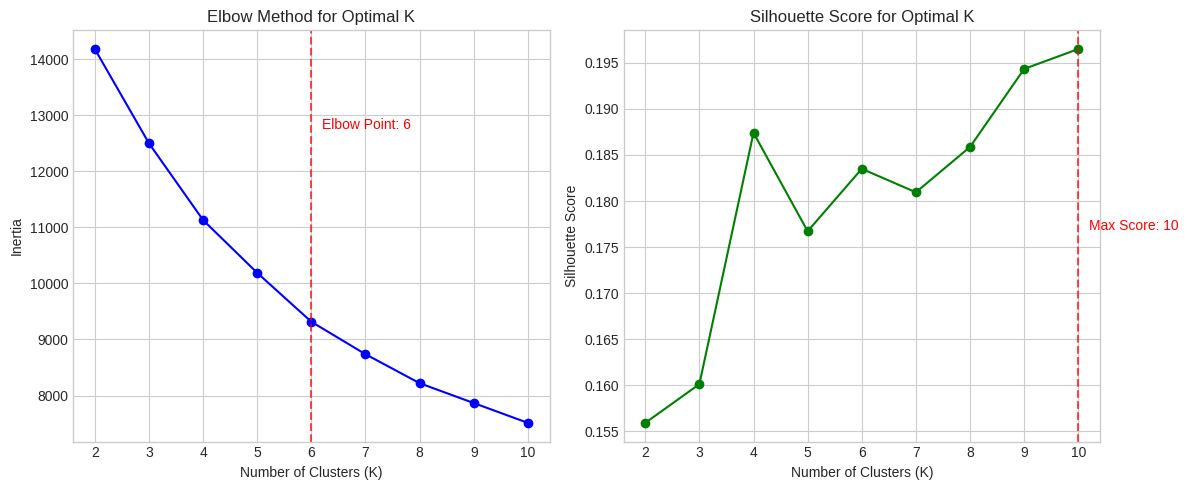
*Figure 9: ROC displaying the false v. true positive rates for our random forest model (AUC = 1.0) with near perfect accuracy at separating obese from non-obese.*



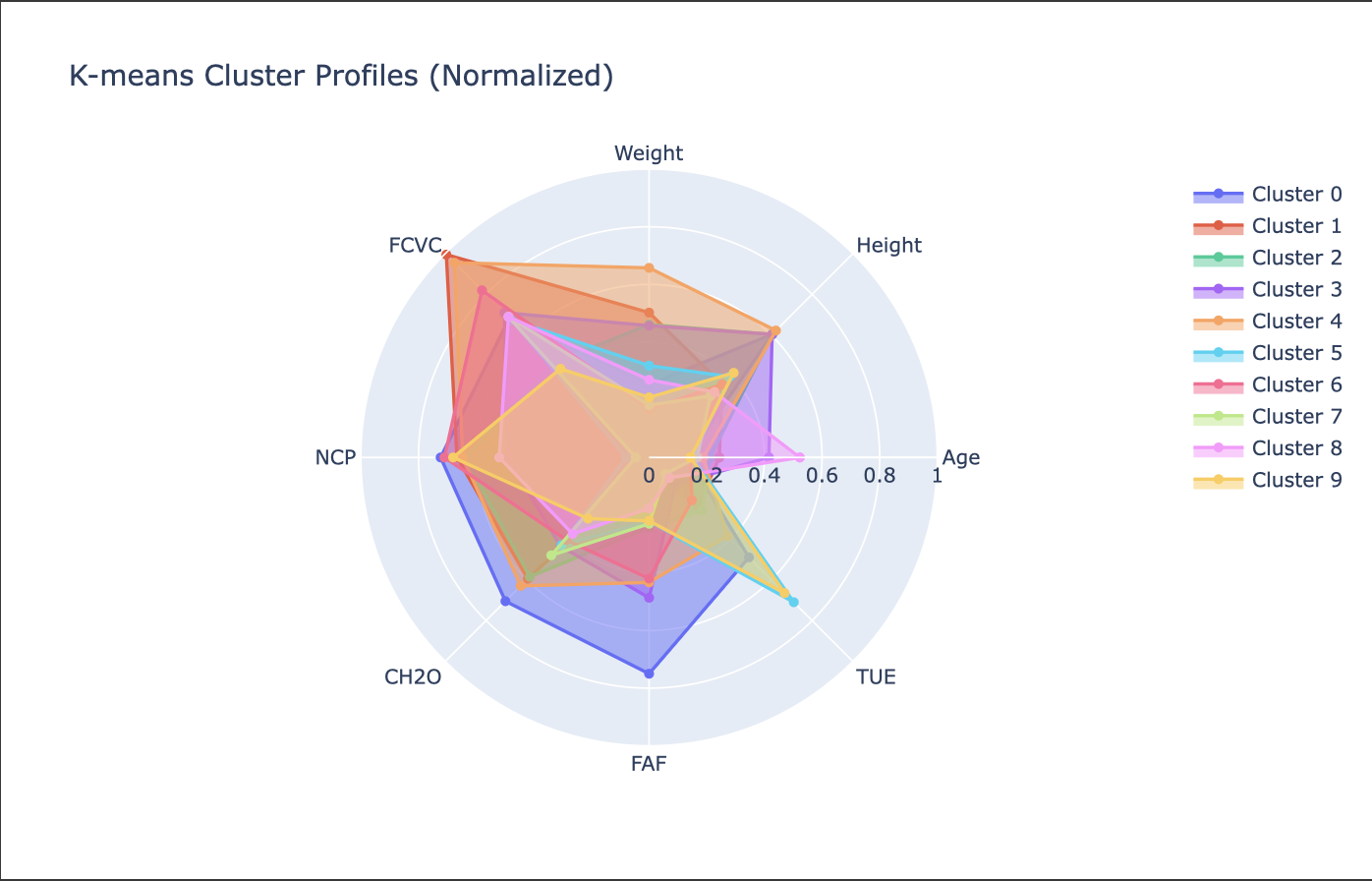
*Figure 10: Cumulative explained variance by principal components. The red line shows that 7 components capture 80% of total variance.*



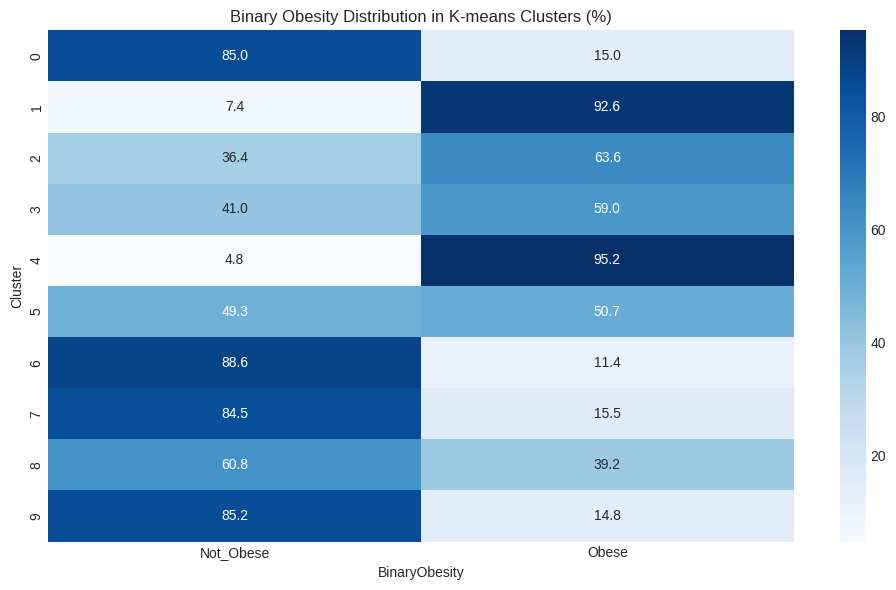
*Figure 11: Determination of optimal cluster number using elbow method (left) showing diminishing returns after k=6, and silhouette scores (right) reaching maximum at k=10.*



*Figure 12: Radar chart showing normalized feature profiles for each of the 10 clusters, highlighting distinct characteristics across weight, age, physical activity, and other key metrics.*



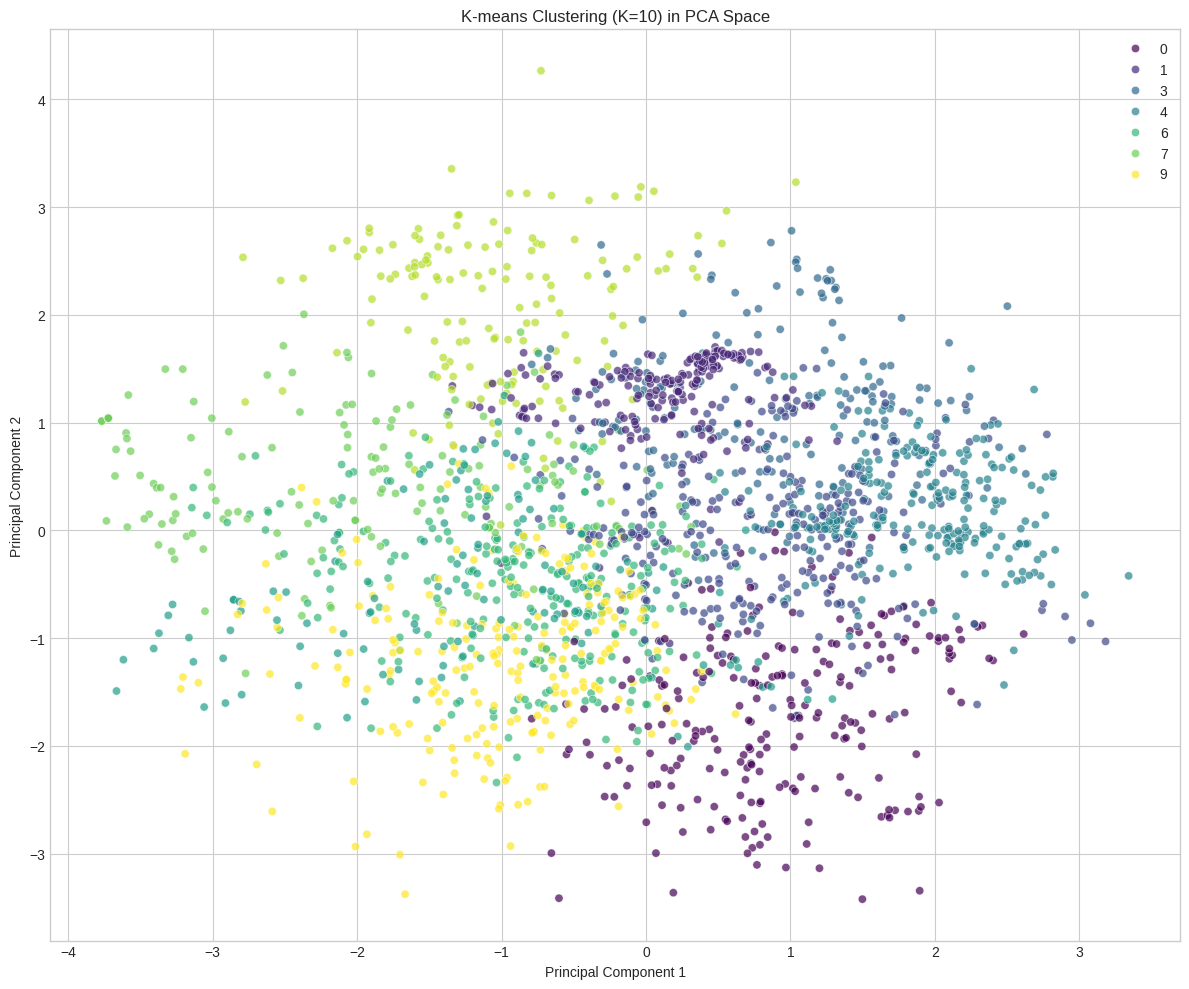
*Figure 13: Heatmap showing binary obesity distribution (%) across all 10 clusters, demonstrating clear separation between high-obesity, mixed, and low-obesity groups.*



*Figure 14: Normalized importance of features in cluster formation, showing age (0.85) and number of main meals (0.84) as the most influential factors, exceeding weight itself.*



*Figure 15: Visualization of clusters in 2D PCA space showing natural separation between high-obesity clusters (purple/blue) and low-obesity clusters (green/yellow).*

****