Obesity levels based on eating habits and physical condition

In this project, it used the *Obesity Levels Based On Eating Habits and Physical Condition* dataset from UCI Machine Learning Repository (2019). The research question focus on:

- To what extent can individual eating patterns, physical activity, and family history predict obesity levels among adults in Latin America?

EDA analysis

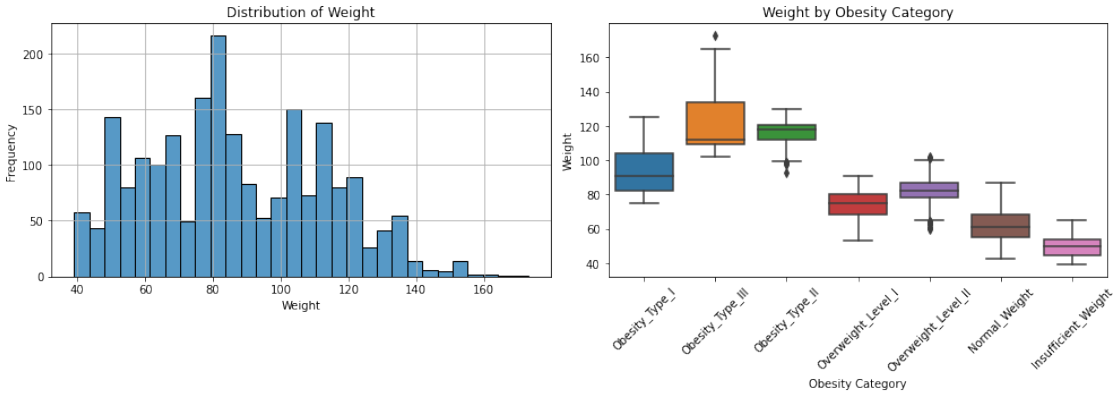

Figure 1: Distribution of weight and box plot by Obesity Category

We conducted an exploratory analysis of the dataset, which contains 2,111 records with 16 variables. Features include demographics (gender, age, etc), behaviours (food consumption frequency, physical activity, etc) from individuals in Mexico, Peru, and Colombia, and a target variable (`NObeyesdad`) representing one of seven obesity classes:

- Insufficient Weight; Normal Weight; Overweight Level I/II; Obesity Type I/II/III

All features are clean with no missing values, the class distribution (Figure 3) is fairly balanced, with each class comprising approximately 13–16% of the dataset. Thus, it is suitable for multiclass classification tasks without major imbalance corrections. Although Obesity Type I and Type III represent the largest classes, each class has sufficient representation for supervised learning. This mitigates the risk of extreme class imbalance and supports the use of classification-based approaches.

Histograms and boxplots of weight distributions (Figure 1) showed significant variation across obesity categories. Median weight increased monotonically from Insufficient Weight to Obesity Type III, validating the internal consistency of class labels and suggesting that weight is a strong discriminative feature. Additionally, wider inter-quartile ranges were observed among higher obesity classes, indicating greater within-group heterogeneity among obese individuals.

Pairwise Pearson correlation coefficients among features were visualised using a correlation heatmap (Figure 4). Generally, weak correlations between features were observed, with most absolute values falling well below 0.3. This low multicollinearity is desirable, particularly for logistic regression, which assumes that predictor variables are not strongly linearly correlated. The highest correlation between weights and height ($r = 0.46$) reflects expected anthropometric relationships, while other features such as `FCVC`, `CH2O` and `FAF` only show minimal associations with one another. This weak interdependence supports the stability and interpretability of subsequent multivariate modeling.

Table 1: Cramér's V between categorical features and obesity level

Feature	Gender	family_history	FAVC	CAEC	SMOKE	SCC	CALC	MTRANS
Cramér's V	0.5558	0.5403	0.3282	0.3523	0.1113	0.2355	0.2251	0.1785

Table 2: ANOVA F-statistics and p-values for numeric features

Feature	Age	Height	Weight	FCVC	NCP	CH2O	FAF	TUE
F-stat	77.95	38.43	1966.52	112.32	26.81	16.17	17.48	7.88
p-value	3.6×10^{-88}	1.7×10^{-44}	0	3.7×10^{-123}	6.3×10^{-31}	2.8×10^{-18}	7.7×10^{-20}	2.1×10^{-8}

Table 1 summarises the strength of association between each categorical feature and the obesity class labels using Cramér's V statistic. It can observe that `Gender` and `family overweight history` showed moderate associations with the target variable, whereas lifestyle-related variables such as `SMOKE` and `MTRANS` exhibited weaker relationships.

Table 2 presents the ANOVA results for numeric features, which indicate statistically significant differences across obesity levels for all features ($p < 0.001$). `Weight` demonstrated by far the largest F-statistic, highlighting its strong discriminative power, followed by `FCVC` and `Age`. These results support the relevance of both anthropometric and behavioral features in explaining obesity class variation.

Classification of Obesity levels

To evaluate how well lifestyle patterns predict obesity category, we built three classifiers **Logistic Regression (LR)**, **Decision Tree (DT)** and **Random Forest (RF)** , all of which were trained and evaluated by using `PySpark` pipelines.

All models used a combined feature vector of scaled numeric variables and indexed categorical variables. Data was split into 80% training and 20% testing. To optimize model performance and prevent overfitting, we performed hyperparameter tuning using 5-fold cross-validation across all classifiers. For LR we applied multinomial classification in `PySpark` as the target variable is multi-class categorical with 7 obesity classes. Regarding the cross-validation, the average accuracy across the folds was used to select the best parameter combination to train the final model on the entire training set. The best hyperparameters selected for each model were shown in Table 3

Table 3: Best hyperparameters selected via cross-validation

Model	**optimal parameters after tunning**
Logistic Regression	regParam = 0.01, elasticNet = 1.0, maxIter = 50
Random Forest	maxDepth = 15, numTrees = 100, impurity = gini
Decision Tree	maxDepth = 5, impurity = gini

Interpretation:

To assess the predictive performance of the three classifiers, which were evaluated by two key metrics: Accuracy and Macro-Averaged F1 Score as summarised in Table 4. Accuracy reflects overall correctness, while the F1 score balances performance across all classes, which accounts for both precision and recall per class. It shows that RF achieves the highest accuracy (95.1%), followed by DT (90.2%) and LR (81.9%). The F1 scores also show that RF outperformed the others, indicating robustness in handling class imbalance and preserving per-class performance. This performance trend aligns with the complexity and flexibility of the models. It can be concluded that tree-based models (especially RF) are well-suited for

this task due to their ability to capture complex decision boundaries. RF's strong performance is largely attributed to effective handling of non-linear relationships and robustness to noise and feature interactions. The worse performance of LR is likely due to its linear decision boundaries and less flexibility in capturing complex feature interactions.

Table 4: Classification accuracy and macro-averaged F1 scores for the three models

Model	Accuracy	F1 Score
Logistic Regression	0.8186	0.8131
Decision Tree	0.9020	0.9014
Random Forest	0.9510	0.9516

Additionally, RF feature importance visualisation for each class (Figure 5) identified key predictors driving model decisions. Features such as `Weight`, `FCVC` (frequency of vegetable consumption), `CH2O` (water consumption), and `NCP` (number of meals) were among the top predictors, which align well with known dietary and physical activity influences on obesity.

Figure 2 presents the confusion matrices for the LR, DT, and RF models, respectively. Classes 0 through 6 correspond to increasing levels of obesity, ranging from Insufficient Weight (class 0) to Obesity Type III (class 6) as shown in 3

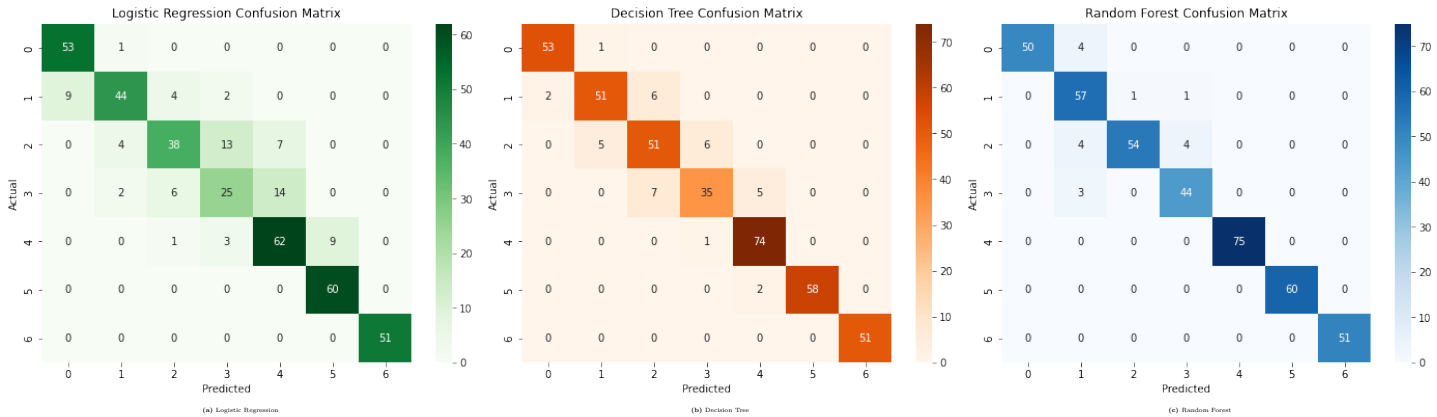

Figure 2: Confusion matrices for three classifiers

- **Logistic Regression (Figure 2a)** shows reasonable separation across classes, with some obvious misclassifications between adjacent categories, especially in class 2 (Overweight Level I), class 3 (Overweight Level II), and class 4 (Obesity Type I). The model achieves strong performance on extremes like class 0 and class 6, but struggles to cleanly distinguish adjacent obesity categories in the middle BMI spectrum. It might be due to the soft category boundaries (BMI or behavioral thresholds are arbitrary).
- **Decision Tree (Figure 2b)** improves class separation over Logistic Regression, especially for classes 2 and 5. The matrix is more diagonally dominant, indicating fewer major confusions, and a very slight overlap exists in overweight obesity levels.
- **Random Forest (Figure 2c)** shows the best separation with minimal off-diagonal values, which reflects high precision and recall for nearly all classes. Misclassification rates were minimal, typically fewer than 5 instances per class(From RF classification report). The confusion matrix is strongly diagonally dominant and indicates good class prediction performance and generalization. And it can be observed that most confusions occurred between adjacent obesity categories (Overweight and Obese), which means that the model captures the meaningful transitions in BMI-based categories effectively.

In conclusion, the high classification performance, particularly from Random Forest demonstrates that behavioural and familial features are highly predictive of obesity categories. These patterns provide strong evidence that individual lifestyle behaviours are informative and can be used for risk stratification.

References

UCI Machine Learning Repository (2019), 'Estimation of obesity levels based on eating habits and physical condition'.
URL: *https://archive.ics.uci.edu/dataset/544/estimation+of+obesity+levels+based+on+eating+habits+a*

Appendices

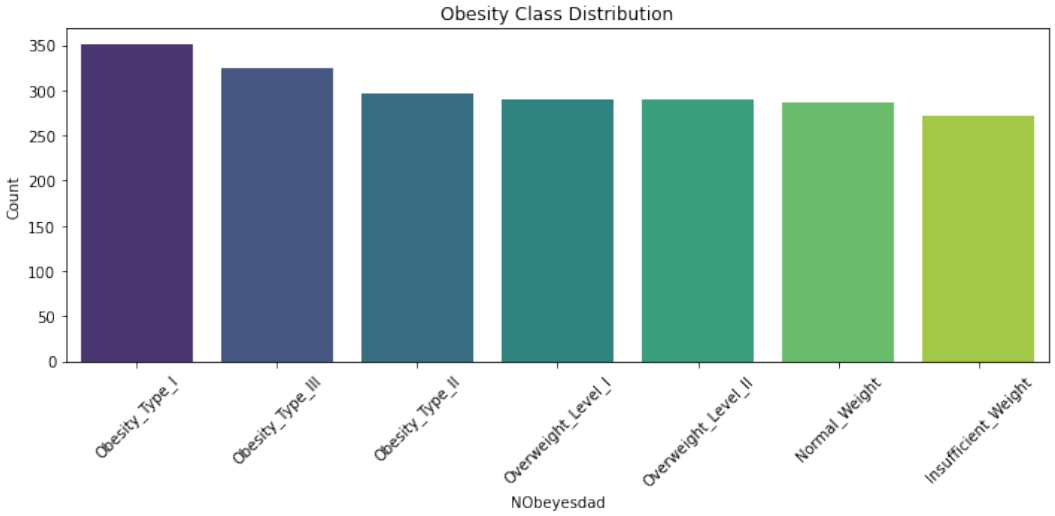

Figure 3: Obesity Class Distribution

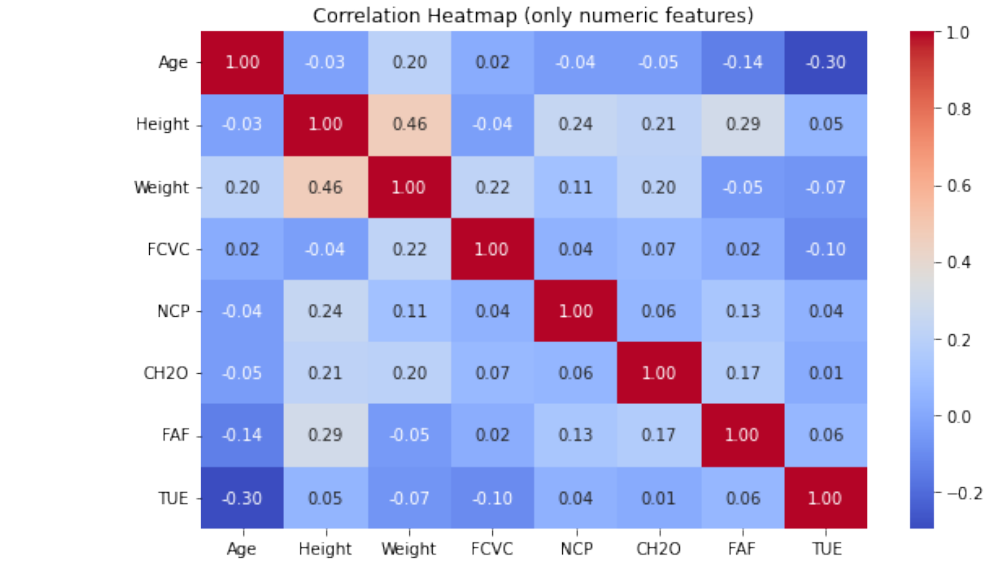

Figure 4: Correlation Heatmap (only numeric features)

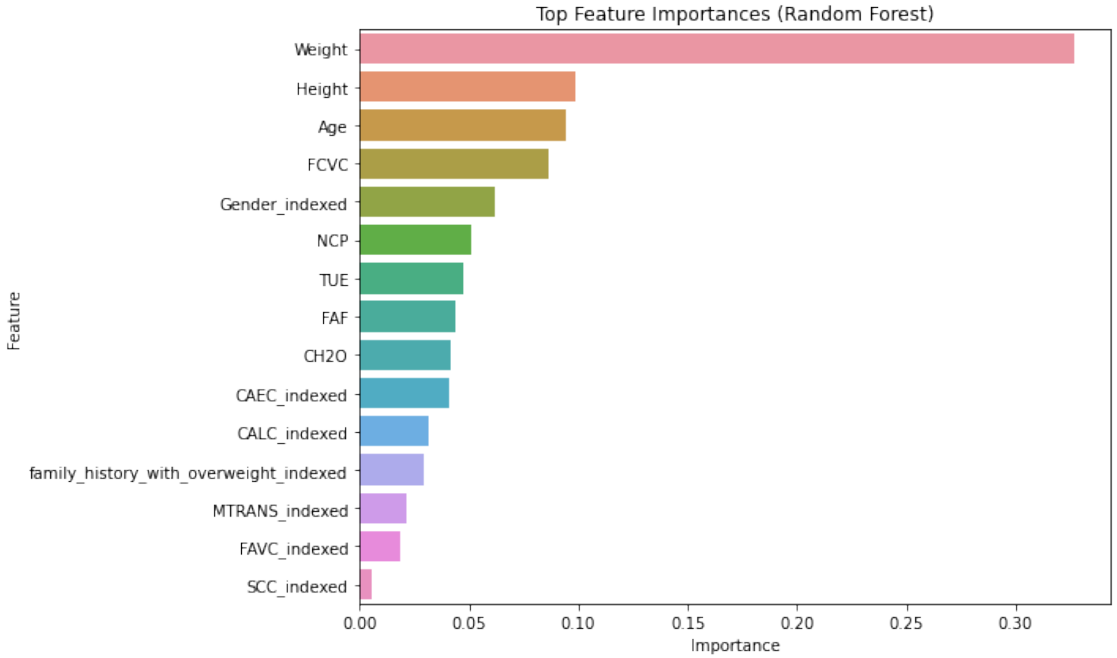

Figure 5: Top Feature Importances (Random Forest)

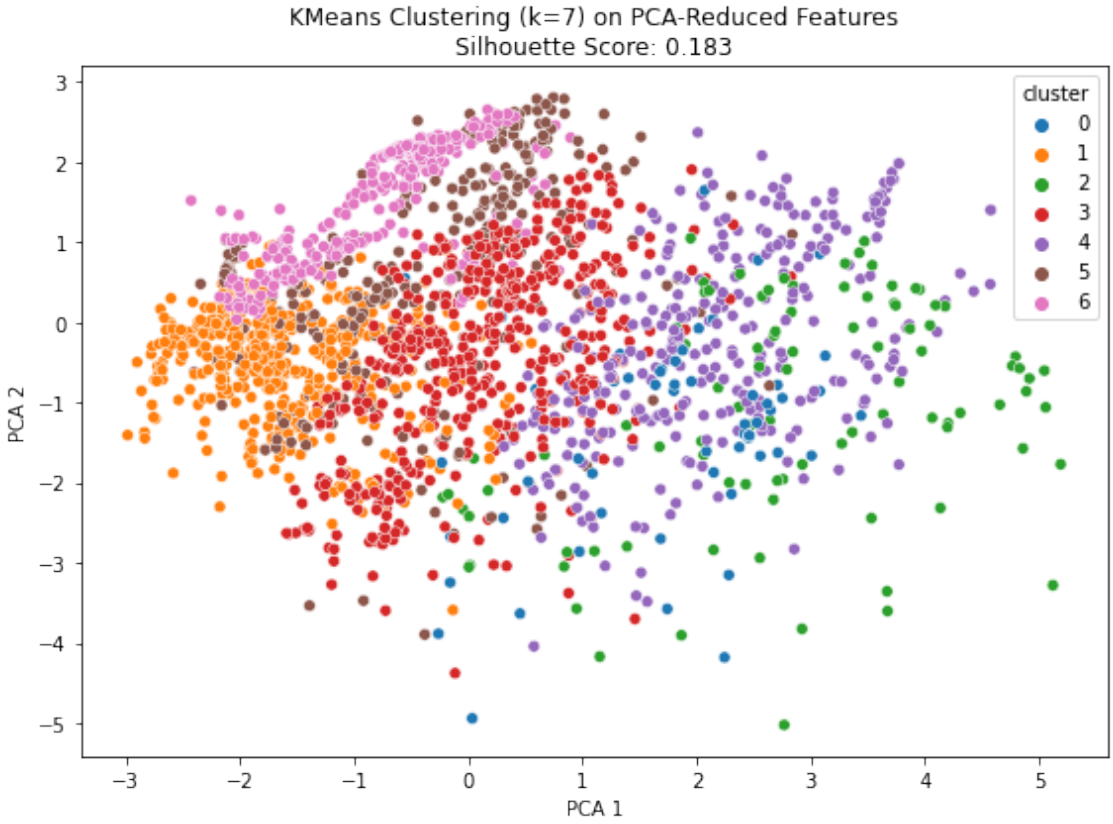

Figure 6: KMeans Clustering (k=7) on PCA-Reduced Features

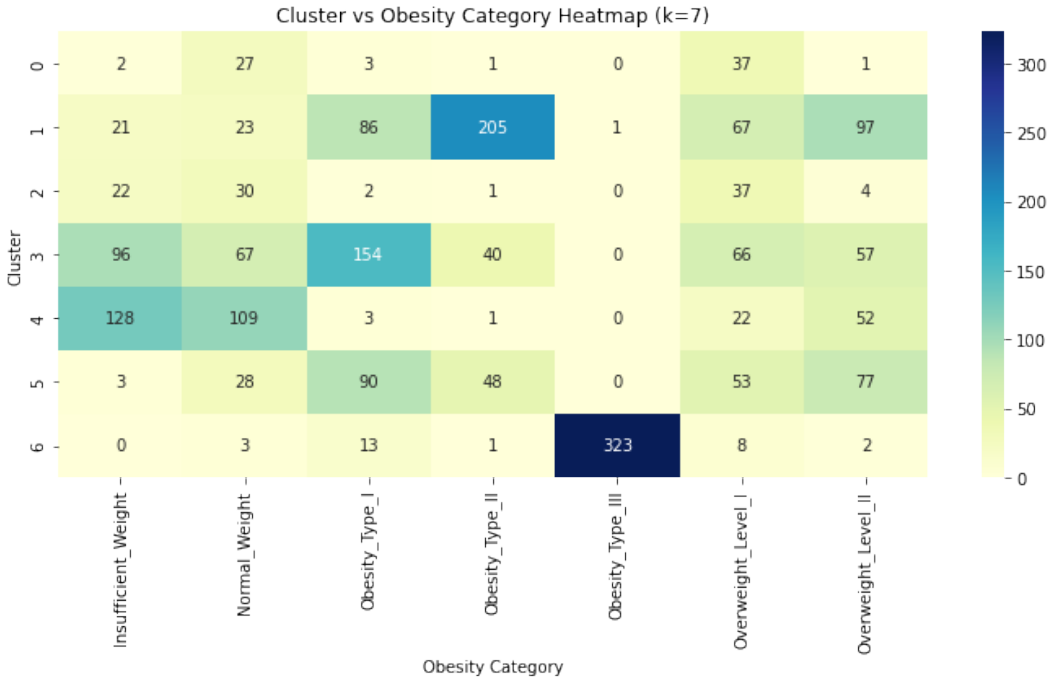

Figure 7: Cluster vs Obesity Category Heatmap (k=7)

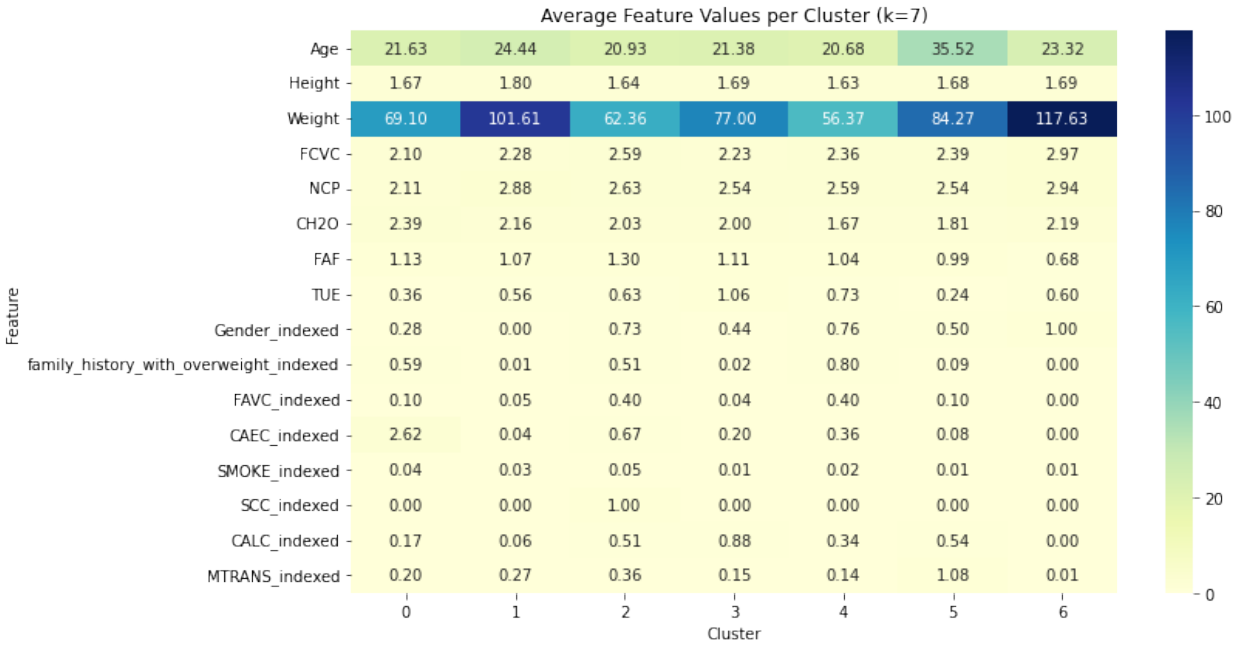

Figure 8: Average Feature Values per Cluster (k=7)