**Analysis of Estimation of Obesity Levels Based on Eating Habits and Physical Condition Dataset**

**Abstract**

This report is prepared for the COMP4462 Data Mining course. We have applied association rule mining, classification, and clustering tasks to the Estimation of Obesity Levels Based on Eating Habits and Physical Condition dataset obtained from the UCI ML repository.

1. **Introduction**

This dataset include data for the estimation of obesity levels in individuals from the countries of Mexico, Peru and Colombia, based on their eating habits and physical condition. The data contains seventeen attributes and 2111 records, the records are labeled with the class variable NObesity (Obesity Level), that allows classification of the data using the values of Insufficient Weight, Normal Weight, Overweight Level I, Overweight Level II, Obesity Type I, Obesity Type II and Obesity Type III. 77% of the data was generated synthetically using the Weka tool and the SMOTE filter, 23% of the data was collected directly from users through a web platform.  
  
It has 17 attributes: Gender , Age , Height , Weight , family\_history\_with\_overweight, FAVC(high caloric food) , FCVC(vegetables) , NCP(number of meals) , CAEC(any food between meals) , SMOKE , CH2O(daily water intake) , SCC(monitoring calories) , FAF(physical activity) , TUE(usage of electronical devices) , CALC(frequency of drinking alcohol) , MTRANS(type of transportation) , NObeyesdad(obesity level).

***Objective of the Project***

The objective of this project is to perform data analysis, clustering, association rule mining, and supervised classification to gain insights and evaluate model performance for a given dataset.

The overarching goal is to understand data patterns, cluster observations, and classify outcomes effectively, while evaluating and comparing different machine learning techniques for their effectiveness.

**Structure of the Report**

* The Introduction provides information about the dataset (page 1).
* Data Preprocessing to explain steps taken to prepare data (pages 2-3).
* Association Rule Mining to discover patterns and rules (pages 3-5).
* Clustering for grouping data and evaluating cluster quality (pages 5-7).
* Classification for explaining the chosen model (pages 7-10).
* The Conclusion summarizes outcomes, challenges, solutions (pages 10-11)
* Report sources (page 12)

1. **Data Preprocessing**

Missing Values Check:

The dataset checked for missing values. There were no missing values in the dataset.

Feature Encoding:

Categorical features , such as MTRANS , NObeyesdad , were encoded to be able to usable for machine learning algorithms. Label encoding was applied.

Over Sampling and Class Imbalance:

Due to 77% of the data was generated synthetically. Imbalance increasing the minority class replication was expected. To manage them, SMOTE-NC was applied.

Outlier Detection:

Used box plots to check for outliers. No significant anomalies were found.

*Justification for the Methods Used*

Label Encoding:

Label Encoding is a technique that is used to convert categorical columns into numerical ones so that they can be fitted by machine learning models which only take numerical data. It is an important pre-processing step in a machine-learning project.

SMOTE-NC:

This algorithm extends the SMOTE algorithm to manage both nominal and continuous features appropriately. Manages mixed data types without distorting the categorical feature space.

Outlier Detection:

diyagram, çizgi, ekran görüntüsü, sayı, numara içeren bir resim

Açıklama otomatik olarak oluşturuldu

Missing Value Detection:

metin, ekran görüntüsü, menü, yazı tipi içeren bir resim

Açıklama otomatik olarak oluşturuldu

1. **Association Rule Mining**

The Apriori algorithm was used for association rule mining due to its efficiency in identifying frequent itemset (Agrawal and Srikant, 1994). Its approach reduces computational complexity, making it well-suited for rule mining over dataset.

Key metrics of Apriori Algorithm

* Support: Measures how frequently an item appears in the dataset relative to the total number of transactions.
* Confidence: Tells us how often items go together.
* Lift: Shows how strong the connection is between items.

The values for each of them

* Support > 0.05
* Lift > 1.0
* Confidence > 0.8

*Frequent Itemset and Association Rules Discovered*

**Frequent itemset:** {Age=high, Weight=high, FCVC=low, Gender=male, FamilyHistory=present}

**Rule:** If Age is high, Weight is high, FCVC is low, Gender is male, and FamilyHistory is present, then Height is high, NCP is high, and CAEC is high (Support: 0.10, Confidence: 0.83).

**Frequent itemset:** {Age=high, Weight=high, FCVC=low, NCP=high, Gender=male}

**Rule:** If Age is high, Weight is high, FCVC is low, NCP is high, and Gender is male, then Height is high, CAEC is high, and FamilyHistory is present (Support: 0.10, Confidence: 0.83).

**Frequent itemset:** {Age=high, Weight=high, FCVC=low, NCP=high, Gender=male}

**Rule:** If Age is high, Weight is high, FCVC is low, NCP is high, and Gender is male, then Height is high and CAEC is high (Support: 0.11, Confidence: 0.85).

**Frequent itemset:** {Age=high, Weight=high, FCVC=low, NCP=high, Gender=male, FamilyHistory=present}

**Rule:** If Age is high, Weight is high, FCVC is low, NCP is high, Gender is male, and FamilyHistory is present, then Height is high and CAEC is high (Support: 0.10, Confidence: 0.85).

**Frequent itemset:** {Age=high, FCVC=low, Weight=high, Gender=male}

**Rule:** If Age is high, FCVC is low, Weight is high, and Gender is male, then Height is high, NCP is high, and CAEC is high (Support: 0.11, Confidence: 0.82).

**Frequent itemset:** {Age=high, FCVC=low, NCP=high, Gender=male, FamilyHistory=present}

**Rule:** If Age is high, FCVC is low, NCP is high, Gender is male, and FamilyHistory is present, then CAEC is high, Height is high, and Weight is high (Support: 0.10, Confidence: 0.81).

**Frequent itemset:** {Age=high, FCVC=low, NCP=high, Gender=male, FamilyHistory=present}

**Rule:** If Age is high, FCVC is low, NCP is high, Gender is male, and FamilyHistory is present, then Height is high and CAEC is high (Support: 0.11, Confidence: 0.84).

**Frequent itemset:** {Age=high, NCP=high, Weight=high, Gender=male}

**Rule:** If Age is high, NCP is high, Weight is high, and Gender is male, then Height is high, CAEC is high, and FamilyHistory is present (Support: 0.13, Confidence: 0.81).

**Frequent itemset:** {Age=high, FCVC=low, Gender=male, FamilyHistory=present}

**Rule:** If Age is high, FCVC is low, Gender is male, and FamilyHistory is present, then Height is high, NCP is high, and CAEC is high (Support: 0.11, Confidence: 0.81).

**Frequent itemset:** {Age=high, Weight=high, FCVC=low, Gender=male, FamilyHistory=present}

**Rule:** If Age is high, Weight is high, FCVC is low, Gender is male, and FamilyHistory is present, then Height is high and CAEC is high (Support: 0.10, Confidence: 0.84).

*Interpretation of the results:*

Age, weight, and gender: Older, heavier individuals, particularly males, tend to have higher height, NCP, and CAEC, especially when they consume fewer vegetables. Family history with overweight: A family history of overweight is frequently associated with higher height, NCP, and CAEC, suggesting a genetic influence. Confidence and support: High confidence values (0.81 to 0.85) indicate strong reliability in these relationships, while support values (0.10 to 0.13) show that these patterns are significant but observed in a moderate proportion of the dataset.

In general, the combination of genetic (family history), lifestyle (weight, vegetable consumption), and demographic factors (age, gender) plays a key role in predicting outcomes such as height, NCP, and CAEC.

1. **Clustering**

For clustering, we used K-Means and Hierarchical clustering

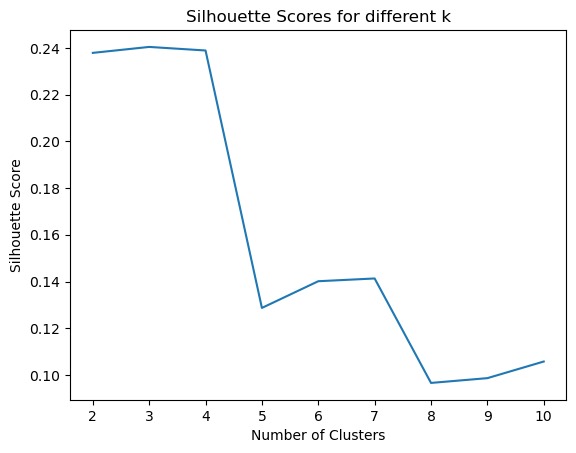
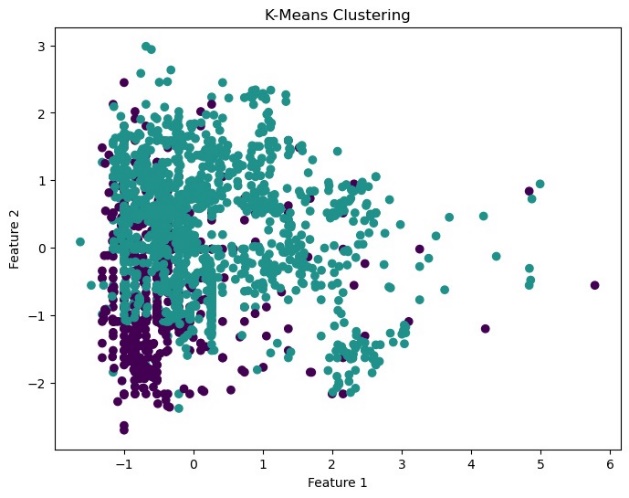
K-means clustering is a technique used to organize data into groups based on theirsimilarity and divides into K distinct group (K = number of cluster). To calculate that similarity, it uses Euclidean distance as a measurement.

Hierarchical clustering is a connectivity-based clustering model that groups the data points together that are close to each other based on the measure of similarity or distance. A dendrogram produced by hierarchical clustering shows the hierarchical relationships between groups. Individual data points are located at the bottom, while the largest clusters, which include all the data points, are located at the top.

We applied feature selection to achieve a higher Silhouette Score.

Before feature selection:

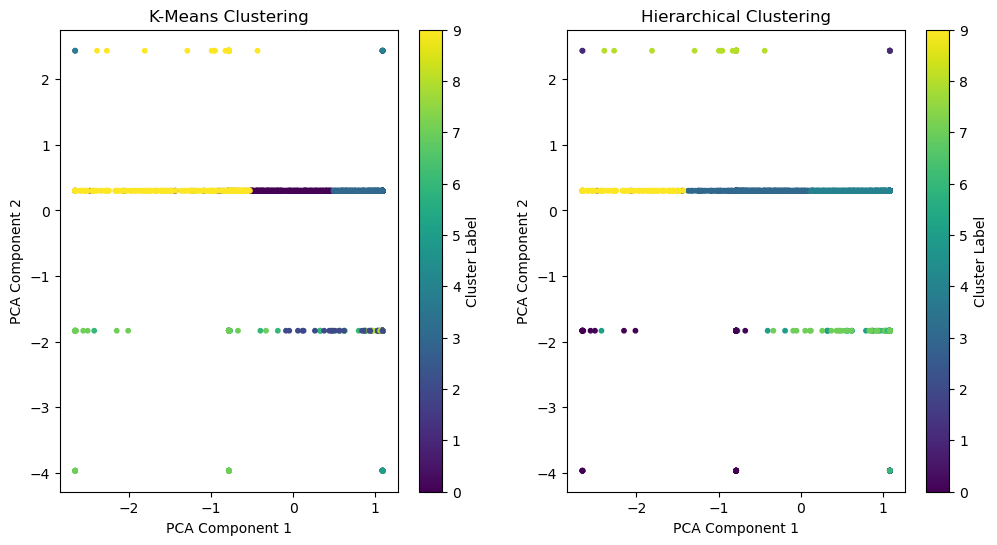
The Silhouette Score for K-Means and Hierarchical Clustering was 0.24. We checked the elbow chart to obtain suitable number of clusters.

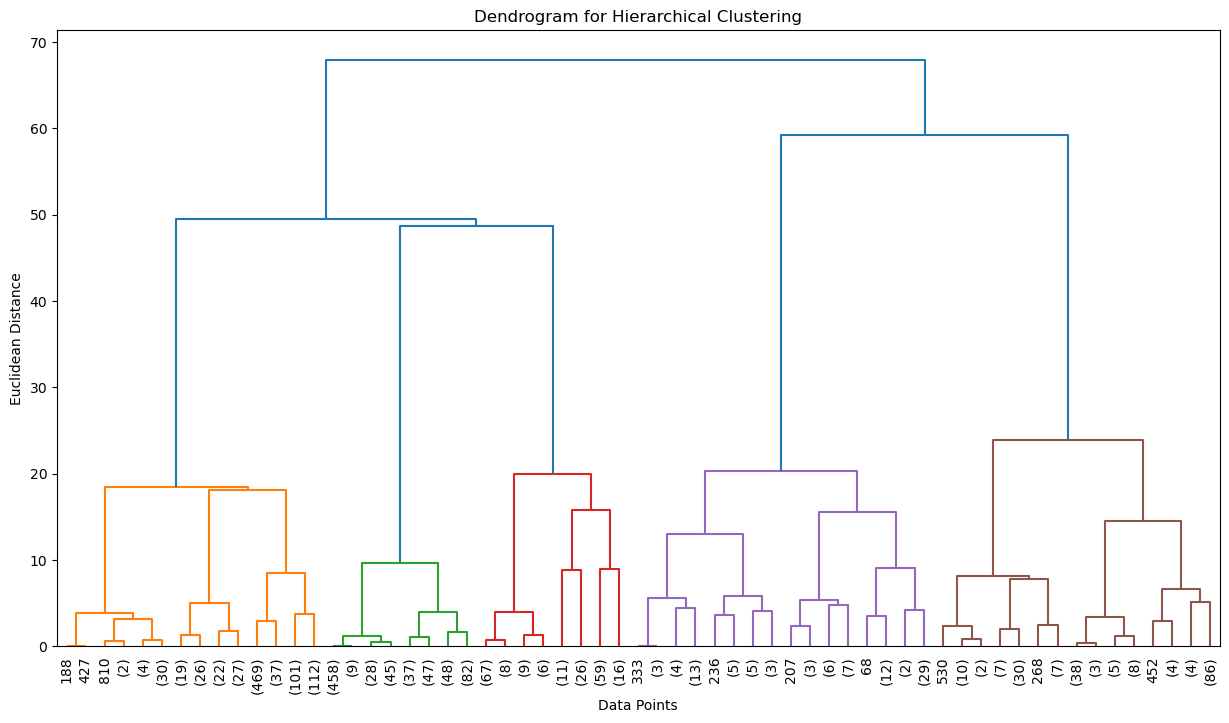
After feature selection:

* The Silhouette Score for K-Means clustering increased to **0.610**.
* The Silhouette Score for Hierarchical Clustering increased to **0.721**.

Feature selection applied by focusing eating habits and monitoring calories. The selected features are FCVC (eating vegetable), CAEC (eating between meals), SCC (monitoring calories), FAVC (high caloric food). We selected our number of clusters as 10 according to Silhouette Score of each cluster.



The plot displays two distinct clustering outcomes: one from K-Means clustering and the other from Hierarchical clustering. Both methods aim to group data points into clusters based on their similarity.



The dendrogram shows five distinct clusters formed by hierarchical clustering with similar eating habits and monitoring calorie intake. Smaller clusters represent specific lifestyles, while larger clusters indicate more general behavior. Since the dataset includes data from Colombia, Peru and Mexico, the clusters may reflect regional lifestyle differences.

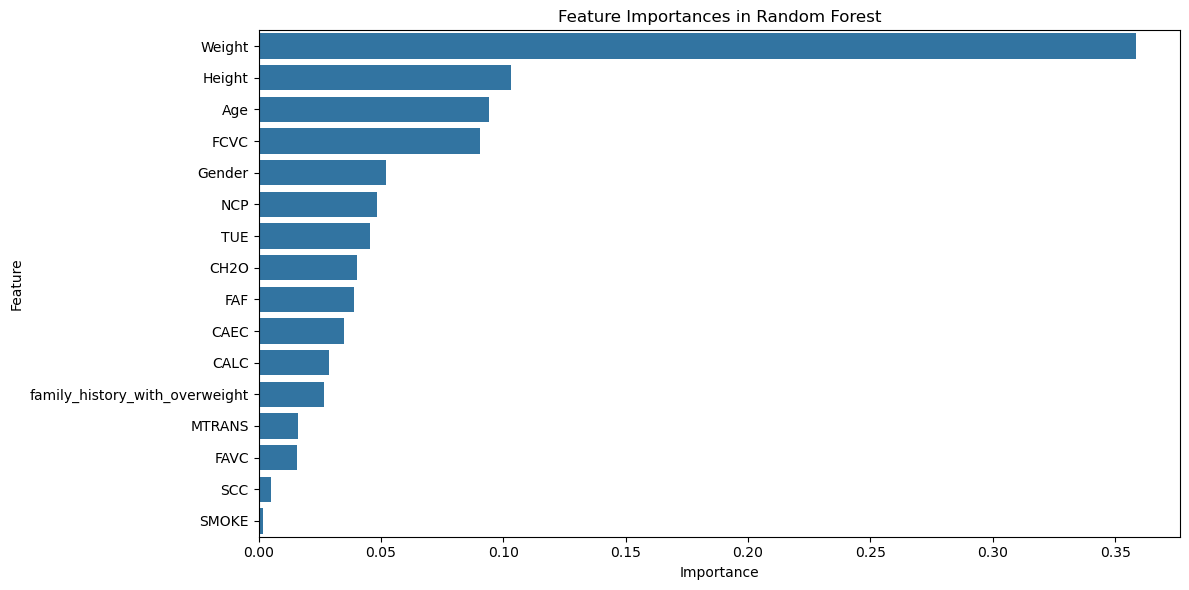
1. **Classification**

Algorithm Selection and Implementation:

For classification, we choose the Random Forest algorithm. The Random Forest is a supervised Machine learning algorithm used for classification, regression, and other tasks using decision trees. Random Forests are particularly well-suited for managing large and complex datasets, dealing with high-dimensional feature spaces, and providing insights into feature importance.

To evaluate its performance, we split our dataset into training and testing sets, allocating 80% of the data for training and 20% for testing.

Feature Importance Evaluation:



The Random Forest model identifies Weight, Height, and Age as the most critical features for predicting obesity levels, with lifestyle and dietary habits such as fast-food consumption (FCVC) and meal frequency (NCP) also playing a significant role. As least crucial features for predicting obesity levels, with monitoring calorie intake (SCC) and smoking (SMOKE).

In obesity calculation, height and body mass index are the most significant factors due to their direct correlation. Additionally, age and vegetable consumption have been found to play a crucial role in maintaining a healthy lifestyle.

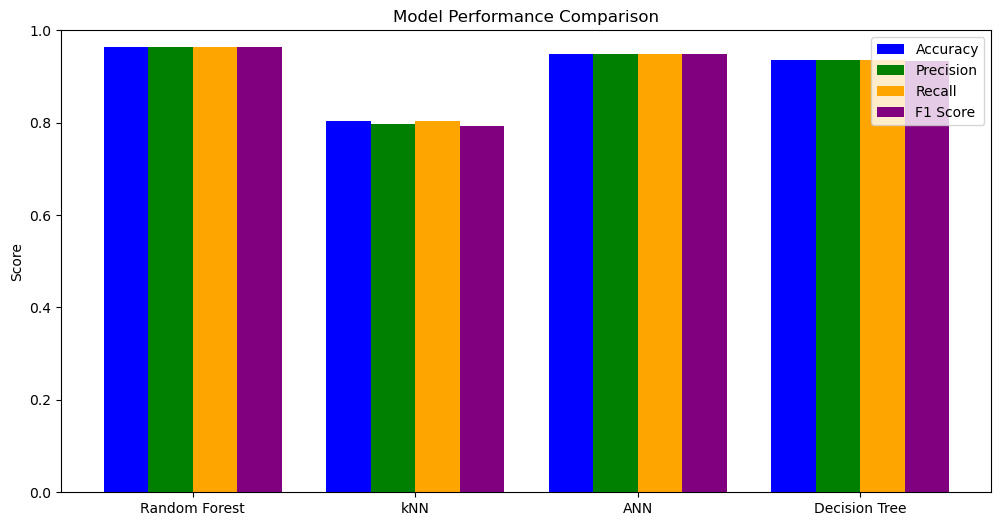
Confusion Matrix Evaluation:

A graph with blue squares and white text

Description automatically generated

To further assess the performance of each classifier, confusion matrices were generated. These matrices provided insights into the true positives, false positives, true negatives, and false negatives for each model. This analysis revealed that while the overall accuracy was high, some misclassifications occurred, particularly between closely related categories like Overweight\_Level\_I and Overweight\_Level\_II. The Random Forest classifier demonstrated the highest accuracy with a score of 96%.

*Compared to other algorithms:*



Based on the evaluation results, Random Forest emerges as the most promising model for this classification task. It consistently demonstrates superior performance across accuracy, precision, recall, and F1-score compared to KNN, ANN, and Decision Tree. This robust performance is attributed to Random Forest's ensemble nature, which enhances its robustness and generalization capabilities. While ANN also achieves high accuracy, it might require more careful hyperparameter tuning and can be computationally expensive. KNN, on the other hand, shows lower performance, potentially due to its sensitivity to the choice of k-value and the curse of dimensionality. Decision Tree, despite its simplicity, is prone to overfitting, leading to lower accuracy and generalization compared to Random Forest. Overall, Random Forest's combination of high performance, robustness, and reasonable interpretability makes it a strong candidate for this classification problem. Each algorithm scores shown below.

* k-Nearest Neighbors (kNN) Evaluation:

Accuracy: 0.8028

Precision: 0.7966

Recall: 0.8028

F1-Score: 0.7918

* Artificial Neural Network (ANN) Evaluation:

Accuracy: 0.9492

Precision: 0.9497

Recall: 0.9492

F1-Score: 0.9494

* Decision Tree Classifier Evaluation:

Accuracy: 0.9350

Precision: 0.9361

Recall: 0.9350

F1-Score: 0.9345

* Random Forest Classifier Evaluation:

Accuracy: 0.9634

Precision: 0.9642

Recall: 0.9634

F1-Score: 0.9636

1. **Conclusion**

We aimed to observe the distribution of obesity levels based on genetics, dietary habits, and daily activities in our analysis.

The analysis utilized three machine learning techniques Association Rule Mining, Clustering, and Classification to uncover insights into obesity-related factors.

Association Rule Mining:

Key rules highlight the impact of healthy eating and regular physical activity on maintaining normal obesity levels.

Support, confidence, and lift metrics ensured the discovery of strong and reliable patterns.

Clustering:

Feature selection improved clustering results significantly, with Silhouette Scores increasing for K-Means (0.610) and Hierarchical clustering (0.721).

The results revealed meaningful lifestyle patterns, with potential regional differences among Colombia, Peru, and Mexico.

Classification:

The Random Forest algorithm identified critical predictors of obesity, such as weight, height, age, and dietary habits.

It demonstrated strong predictive accuracy, emphasizing the importance of key lifestyle factors.

PREPROCCESSING CHALLENGE

Our dataset includes a mix of categorical and numerical data types, necessitating careful preprocessing steps to ensure compatibility with machine learning algorithms. Categorical features, such as gender or CAEC, were encoded with label encoding to convert them into numerical representations suitable for analysis. For addressing class imbalances in the target variable, which can impact model performance, the SMOTE-NC (Synthetic Minority Oversampling Technique for Nominal and Continuous data) method was applied. This technique is particularly useful for datasets with mixed data types, as it generates synthetic samples for minority classes while preserving the relationship between categorical and numerical features.

ASSOCIATION RULE MINING CHALLENGE

The dataset contained many association rules derived from the complex relationships between features. However, due to the large amount of data and the potential for redundancy or irrelevant patterns, not all association rules were equally useful for accurate predictions to make the model more accurate and easier to understand, we chose the most important and useful rules. This selection process prioritized rules with strong support, confidence, and lift values, ensuring that the chosen associations contributed effectively to the estimation of obesity levels while reducing noise and computational complexity.

FEATURE SELLECTING CHALLENGE

The dataset contained 16 features, some of them were noisy or less relevant for the clustering process. To ensure meaningful and accurate clustering, we did feature selection techniques to identify the most significant features that contributed to the understanding of obesity patterns. This process helped eliminate noise and redundancy, improving the clustering's interpretability and performance. Ultimately, four key features were selected from the original 16, focusing on those with the highest relevance and impact on identifying distinct clusters of obesity levels. This simpler approach made the calculations easier and helped show the connections between eating habits, physical health, and obesity more clearly.

**References**

Guide to Intelligent Data Science (2021)

<https://doi.org/10.1007/978-3-030-45574-3>

BMI calculation

<https://www.who.int/data/gho/data/themes/topics/topic-details/GHO/body-mass-index>

**Dataset for estimation of obesity levels based on eating habits and physical condition in individuals from Colombia, Peru and Mexico**

<https://www.sciencedirect.com/science/article/pii/S2352340919306985?via%3Dihub>

**A healthy lifestyle - WHO recommendations**

<https://www.who.int/europe/news-room/fact-sheets/item/a-healthy-lifestyle---who-recommendations>

Estimation of Obesity Levels Based On Eating Habits and Physical Condition (**8/26/2019**)

[**https://archive.ics.uci.edu/dataset/544/estimation+of+obesity+levels+based+on+eating+habits+and+physical+condition**](https://archive.ics.uci.edu/dataset/544/estimation+of+obesity+levels+based+on+eating+habits+and+physical+condition)

**Random Forest Classifier**[**https://scikit-learn.org/1.5/modules/generated/sklearn.ensemble.RandomForestClassifier.html**](https://scikit-learn.org/1.5/modules/generated/sklearn.ensemble.RandomForestClassifier.html)

**SMOTE-NC**

[**https://imbalanced-learn.org/stable/references/generated/imblearn.over\_sampling.SMOTENC.html**](https://imbalanced-learn.org/stable/references/generated/imblearn.over_sampling.SMOTENC.html)