Obesity Level Prediction Analysis

Problem Statement:

Obesity has emerged as a significant public health concern globally, with its prevalence increasing rapidly over the past few decades. Effective management and prevention of obesity require accurate identification of individuals at risk. If we can uncover the foundational factors driving obesity, it becomes feasible to implement targeted interventions early in life, equipping individuals with the knowledge and resources needed to manage their weight and overall health effectively.

I will be employing decision tree and k-nearest neighbors (KNN) models to predict obesity levels. These models will help us understand which factors are most influential in determining obesity and how they relate to each other. This knowledge will guide us in developing effective strategies to address obesity and meet the unique needs of individuals and communities affected by it.

Objective:

The primary objective of this study is two fold:

- To Identify the major lifestyle, and physiological factors contributing to obesity
- To compare the performance of two classification models in predicting obesity levels based on various factors.

Data:

The dataset was downloaded from UC Irvine Machine Learning Repository.
This dataset contains data for the estimation of obesity levels in people from the countries of Mexico, Peru and Colombia, with ages between 14 and 61 and diverse eating habits and physical conditions. The data was collected using a web platform with a survey where anonymous users answered each question, then the information was processed obtaining 17 attributes. The data contains numerical data and continuous data, so it can be used for analysis based on algorithms of classification.

Overview of the data:

This dataset contains 2111 rows and 17 variables(attributes).
Attributes/variables being tracked in this dataset -

- Gender - Categorical variable - has these values - Male and Female
- Age - This is a numerical variable
- Height - This is a numerical variable
- Weight - This is a numerical variable

- family_history_with_overweight - This is a binary variable which represents if the person’s family has history with overweight
- FAVC - This is a binary variable which answers the question - Do you eat high caloric food frequently.
- FCVC - This is a numerical variable which answers the question - Do you usually eat vegetables in your meals.
- NCP - This is a numerical variable which answers the question - How many meals do you have daily
- CAEC - this is a categorical variable which answers the question - Do you eat any food between meals , and will have these following values - “Sometimes, Frequently, Always, no”
- Smoke - This is a Binary variable, which answers the question - Do you smoke
- CH2O - This is a numerical variable which answers the question - How much water do you drink daily (Probably in liters)
- SCC - This is a binary variable which answers the question - Do you monitor the calories you eat daily
- FAF - This is a numerical variable which answers the question - How often do you do physical activity. (In hours)
- TUE - This is a numerical variable which answers the question - How much is your screen time (in hours)
- CALC - this is a categorical variable which answers the question - How often do you drink alcohol , and will have these following values - “Sometimes, Frequently, Always, no”
- MTRANS - this is a categorical variable which answers the question - Which transportation do you usually use, and will have these following value - “Public_Transportation, Automobile, Walking, Motorbike, Bike”
- NObeyesdad - this is the target variable - and shows the obesity level. Obesity level can be classified into 7 types, but I am going to reduces the levels to 4 types probably.

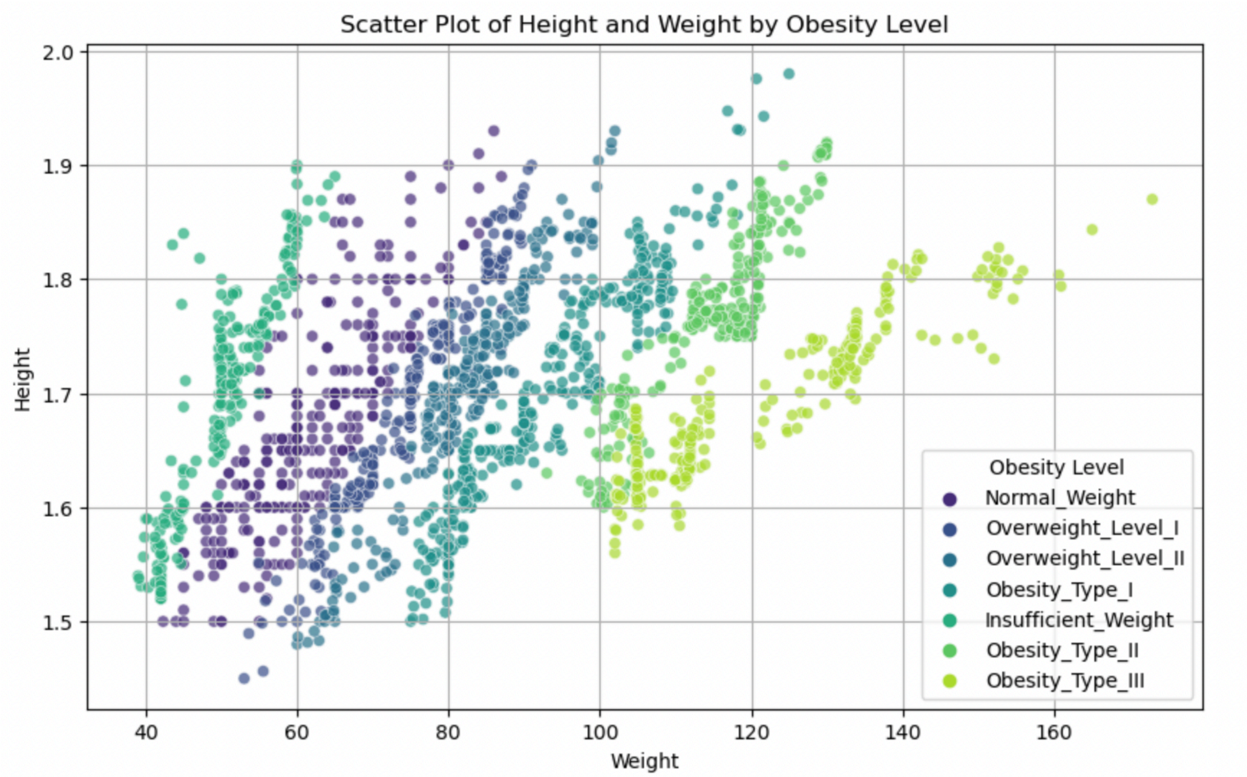
The target variable here is - ‘***NObeyesdad’.*** The possible classes of the target variable and proportions of items in each class are shown below.:

NObeyesdad	
Obesity_Type_I	351
Obesity_Type_III	324
Obesity_Type_II	297
Overweight_Level_II	290
Normal_Weight	279
Overweight_Level_I	276
Insufficient_Weight	265

I conducted this preprocessing step to enhance the clarity and informativeness of the dataset attributes. Initially, I found the attribute names to be lacking in clarity and potentially confusing. Therefore, I opted to rename the attributes to better reflect their respective features and improve understanding during the analysis. This renaming process aimed to streamline the dataset and facilitate more intuitive interpretation of the data. Below are the attributes with their new attribute names:

NObeyesdad : Obese_Type
FAVC : is_high_caloric_food_consumed
FCVC : vegetables_frequency
NCP : Number_of_meals_daily
CAEC : food_between_meals_frequency
CH2O : water_quantity
SCC : are_calories_monitored
FAF : physical_activity_in_hours
TUE : screentime_in_hours
CALC : alcohol_frequency
MTRANS : Transportation_used

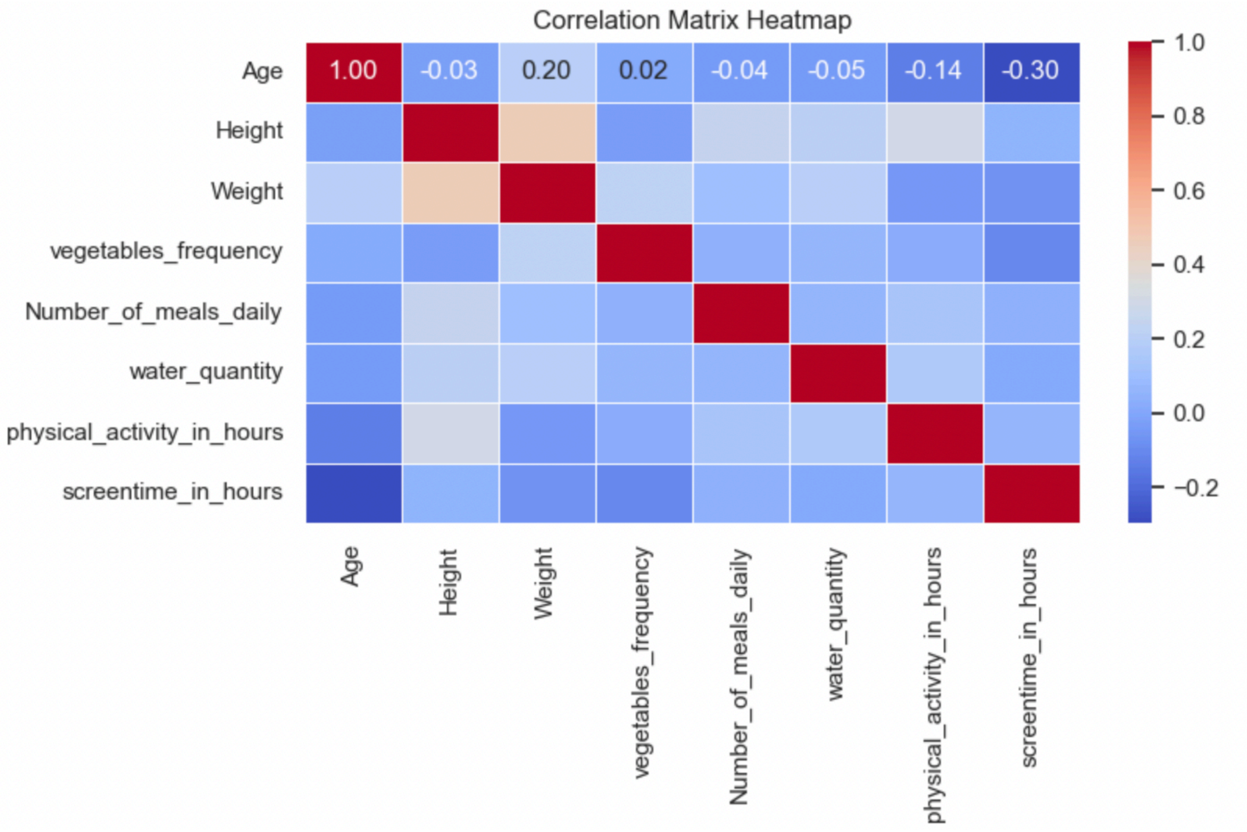
As I began to deeply look in the data and by using plots to understand the data and its correlations. I found the below plot interesting:



This plot shows that while there is no significant correlation between the "Weight" and "Height" attributes, both variables exhibit a substantial impact on obesity levels. The substantial impact of weight and height on obesity levels poses a challenge for predictive modeling. Including these attributes in the modeling process may lead to an overly simplistic approach, as their influence on obesity levels is already well-established. Moreover, incorporating weight and height as predictors may limit the model's ability to capture other significant factors contributing to obesity, potentially hindering its predictive accuracy and generalization to unseen data.

To address the complexity posed by the high impact of weight and height on obesity levels, we propose excluding these attributes during the data preprocessing stage. By removing weight and height from consideration as predictors, we aim to streamline the modeling process and focus on identifying other relevant factors that contribute to obesity outcomes.

As part of exploratory data analysis, I examined the correlation between numeric attributes using a correlation matrix heat-map. This visualization allowed me to assess the degree of linear relationship between all pairs of numeric attributes in the dataset. Below is the matrix:

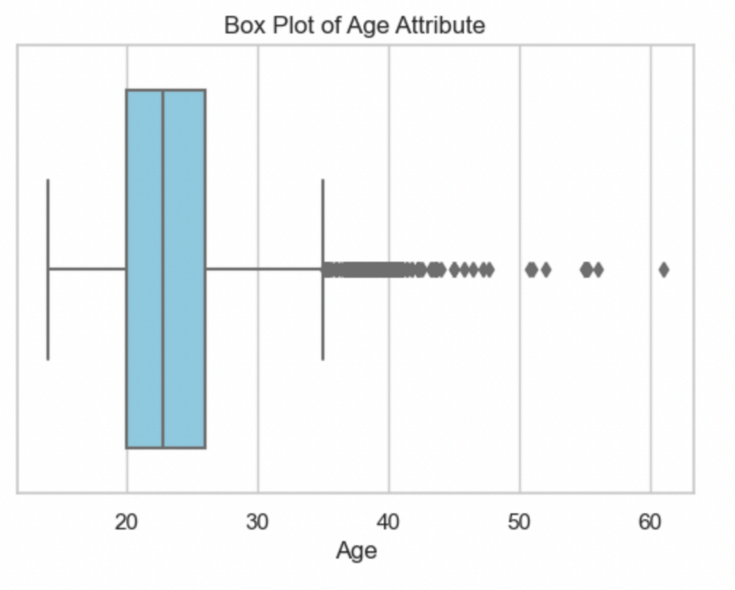


Given absence of major correlations between numeric attributes, I have retained all numeric variables in the modeling process to capture the full spectrum of predictors contributing to obesity levels.

Data Preparation:

No missing values were found in the dataset, eliminating the need for preprocessing steps to handle them. The following data preparation measures were undertaken:

- Attributes are renamed to enhance clarity and understanding.
- The attributes "Weight" and "Height" were excluded from the analysis due to their direct correlation with obesity levels, potentially simplifying the prediction task excessively.
- After removing "Weight" and "Height" attributes, duplicates were identified and subsequently removed. The dataset now comprises 2082 unique rows.
- The "age" attribute exhibits outliers, as illustrated in the accompanying plot. These outliers were not addressed, as age significantly influences obesity levels. Handling outliers could potentially result in the loss of crucial information.



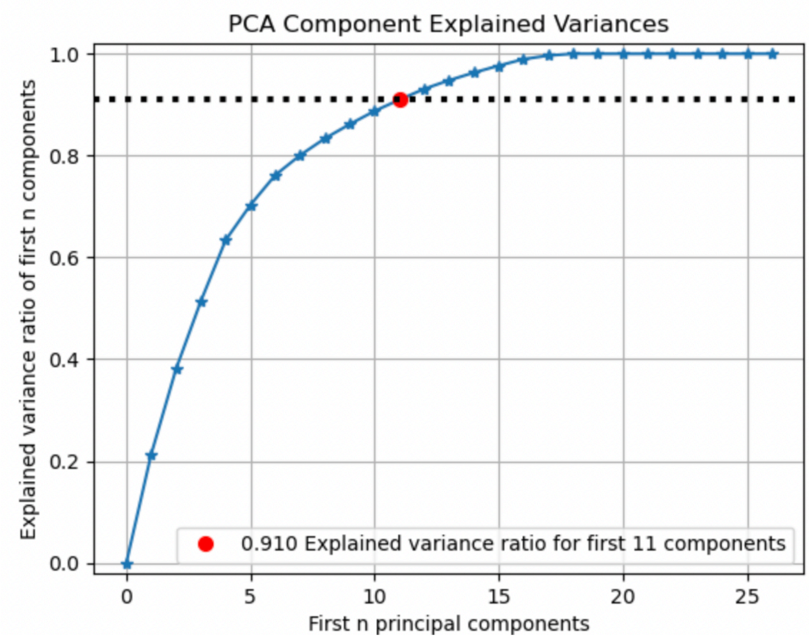
- I have reduced the possible values of the target variable from 7 to 4 by combining similar categories as follows:

    1. **Insufficient Weight**: This category remains unchanged.
    2. **Normal Weight**: This category remains unchanged.
    3. **Overweight**: Combining "Overweight Level I" and "Overweight Level II" into a single category.
    4. **Obese**: Combining "Obesity Type I," "Obesity Type II," and "Obesity Type III" into a single category.

This reduction simplifies the target variable, making it easier for a model to identify patterns in the data and reducing the complexity of the classification task.

- It appears that the value "Always" in the attribute "alcohol_frequency" is likely incorrect, as it is inconsistent with the other values and may be a data entry error. To address this inconsistency, we will change the value "Always" to "Frequently" to maintain consistency within the attribute. This will ensure that the "alcohol_frequency" attribute is consistent and accurate, which is important for maintaining data integrity and reliability in subsequent analyses

- In the 'Transportation_used' attribute, there are a lot of categories. This might reduce the performance of model, so I am going to reduce the number of categories in 'Transportation_used'. Creating three categories for the "Transportation" attribute—Public, Automobile(this will include motorbike), and Walk/Bike—can be a reasonable approach. This simplification retains some level of granularity while still reducing the number of categories. With three categories, the model remains relatively easy to interpret, yet it still captures some nuances in transportation choices that could impact obesity levels.
- Implemented one-hot encoding to transform categorical variables into numerical format, ensuring compatibility with the k-nearest neighbors (KNN) and decision tree algorithms. After transforming I have 30 attributes
- Scaled the data using MinMaxScaler, ensuring that all input attributes (excluding the output attributes) are within a uniform range. This preprocessing step standardizes the features, preventing certain attributes from dominating others due to differences in scale, and facilitating optimal model performance.
- By converting categorical variables into dummy numeric variables, the dataset expanded significantly, introducing numerous new columns and thus introducing the curse of dimensionality. To mitigate this issue, I have used Principal Component Analysis (PCA) to reduce the dimensionality of the dataset while retaining a substantial portion of its information.After doing PCA - I am considering first 11 columns as they capture 90% of the variability in the data as shown below



- I used a train/test split of my data for building the classification model. I allocated 25% of the data for testing purposes and utilized the remaining 75% for training the model. Additionally, I employed cross-validation using grid search specifically for decision trees. For this, I utilized 5 folds to ensure robust model evaluation and parameter tuning.

Modelling:

For this project, I have chosen 2 Classification Models, as shown below:

1. **Decision Tree:** The reasons behind choosing Decision tree are given below
   - Decision trees excel at capturing intricate relationships between features, allowing them to represent complex decision boundaries efficiently.
   - Decision trees can handle noisy data and redundant or irrelevant attributes. They focus on selecting the most informative features at each split, enhancing model generalization and reducing overfitting.
   - Decision trees inherently rank features based on their importance in classification tasks. This feature importance assessment aids in identifying key predictors for obesity levels and refining model architecture.
2. **K-Nearest Neighbors:** The reasons behind choosing KNN are given below
   - KNN is a simple and versatile algorithm that can be applied to classification tasks without many parameters to tune, making it easy to implement and use.
   - It offers decent accuracy, especially when dealing with datasets where the decision boundaries are not linear or easily separable.

Although I initially selected two models, I conducted extensive experimentation to ensure a comprehensive analysis of the dataset. This included:

1. Utilizing decision trees with Principal Component Analysis (PCA), where I retained 11 variables representing 90% of the dataset's variability
2. Employing decision trees with the entire dataset, bypassing PCA.
3. Implementing KNN
4. Conducting grid search with decision trees to determine optimal parameters and performing 5-fold cross-validation.
5. Extending the grid search approach to decision trees with PCA, optimizing parameters, and conducting 5-fold cross-validation.

These diverse approaches allowed for a thorough exploration of model performance and parameter tuning, ensuring robustness in our analysis.

By systematically adjusting the hyperparameters through techniques like grid search, I aimed to identify the optimal configuration that maximizes the model's performance on the given dataset. This process helps mitigate overfitting, enhance generalization capabilities, and ultimately improve the reliability of the model's predictions.

Evaluation:

The metric chosen for evaluation in this report is F1-score as it is a robust metric that balances precision and recall, offering insight into the model's ability to correctly classify instances across different classes. By considering both false positives and false negatives, the F1-score provides a holistic assessment of the model's performance. I have used Micro average F1 score to compare the models as it calculates the F1-score across all classes by summing the individual true positives, false positives, and false negatives before computing the harmonic mean. This metric is particularly useful here, as it accounts for variations in class frequencies and ensures each class contributes equally to the overall score.

Performance of Models: After analyzing four classification reports, it is evident that two models utilizing GridSearch with 5-fold cross-validation did not meet the expected accuracy levels. This suggests that hyper-parameter tuning did not significantly improve model performance. Now, let's contrast the Decision Tree model with the KNN Model.

The Decision Tree models achieved varying levels of performance across different classes. While they demonstrated high precision and recall for certain classes (e.g., Obese and Insufficient Weight), they struggled with others, particularly those with smaller support (e.g., Normal Weight). This indicates potential class imbalance issues or challenges in distinguishing between similar classes. The micro-average F1-score for this model stands at 0.76, which falls short in comparison to the KNN model's performance.

The **KNN Model** exhibited robust performance in precision and recall across most classes, boasting a micro-average F1-score of 0.81. However, it demonstrated relatively lower performance in the "Normal Weight" class.

While analyzing the relative variable importance, it is observed that certain features played significant roles in predicting obesity levels. Notably, variables such as age, family history, and number of meals daily emerged as influential factors in the Decision Tree model. Interestingly, physical activity, while expected to be crucial, exhibited lower importance in the predictive model.

Conclusion:

The model's performance in meeting the original goals of the project can be considered satisfactory, with a room for improvement. While the models demonstrated moderate success in predicting obesity levels based on lifestyle and physiological factors, they fell short of achieving the desired level of accuracy for real-world deployment, particularly in identifying individuals with Normal Weight.

To trust the model in real-life medical scenarios, further validation and fine-tuning are essential. Additional data collection efforts, including gathering more diverse and representative samples, may help address class imbalances and improve model generalization.

Through this project, several key lessons were gleaned:

- Data Preprocessing is Crucial: Thorough data preprocessing, including handling missing values, feature selection, and outlier detection, significantly impacts model performance and interpretability.
- Importance of Evaluation Metrics: Choosing appropriate evaluation metrics is critical for assessing model performance accurately. The F1-score, particularly the micro-average variant, offers a comprehensive evaluation of model precision and recall across all classes, enabling informed decision-making.

In conclusion, while the models presented herein offer valuable insights into obesity prediction, further refinement and validation are necessary before their deployment in real-world scenarios. By leveraging advanced techniques, incorporating domain knowledge, and prioritizing robust evaluation metrics, future iterations of this research hold promise for addressing the complex challenge of obesity prevention and management effectively.