Machine Learning

Dissecting Obesity

Finding the causes with predictive modelling

Group05

20/12/2024

Christian Deluca 20241264

Dino Meng 20241265

Marek Rewolinski 20241452

Michal Marcin Wójcik 20241255

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa

ABSTRACT

Obesity is one of the major humanity's challenges of XXI century. It is highly affecting humans right now, as it is estimated that more than 4 million people die globally each year due to obesity/overweight. [1] It is easier to prevent the illness than cure it, thus it is crucial to find the factors that lead to obesity. In this report we conducted a comprehensive analysis of 1611 observations from collected survey data of people aged 16-56 and living in LatAm. Following the preprocessing stages of removing redundant features, outliers, imputing missing data points and creating new features we adopted 6 various predictive models (Random Forest, Gradient Boosting Classifier, Decision Tree, AdaBoost, Multi-Layer Perceptron (MLP) and Bagging Model) to classify observations into 7 distinct weighting categories. Among tested models through stratified 10-fold cross validation, Gradient Boosting Classifier guaranteed the best and stable results of macro f1 score equal to 95% after selecting important features and tuning the hyperparameters. Our results discovered that BMI class, weight, age and gender help best predict the obesity. However, this does not mean these four features determine obesity alone: an additional twelve features will be used to determine someone's level of obesity. The model tends to misclassify 5% of observations, mostly falling under the Normal Weight and Overweight level I category, due to wider range of weight, height that suits into this group. Overall, this study aims to provide decision makers with information how they can fight with obesity through understanding how it differs among certain groups and what factors contribute to it.

KEYWORDS

Machine Learning; Obesity; Multi-class classification; Gradient Boosting Classifier; Obesity prediction; Supervised learning

INTRODUCTION

Obesity is a health condition where someone's quantity of excessive fat deposits has negative impacts on their health. According to the **World Health Organization (W.H.O.)**, the global obesity landscape has reached a critical milestone: "**more than 1 billion people** were living with obesity" as of **2022**, representing approximately **16% of the global population**. Furthermore, the organization notes that "**43% of adults aged 18 years and over were overweight in 2022**"

The trajectory of this health challenge is particularly alarming. Over recent decades, obesity rates have escalated dramatically, with the W.H.O. reporting that the number of **obese adults has more than doubled since 1990**, while the prevalence among **children and adolescents (5-19 y.o.) has quadrupled**. [2].

As obesity is a multifactorial phenomenon [2], in this study we will attempt to discern the major factors - in particular, **lifestyle and genetic ones** - in impacting individuals' obesity levels. To do this, we will use predictive modelling techniques to accurately classify obesity levels and interpret such models with the results to provide actionable insights that could contribute to more targeted obesity prevention strategies.

BACKGROUND

In this project we used some sophisticated methods that were not covered in class. With this section, we will briefly introduce each of the method used.

The following methods are implemented in the Scikit-Learn module.

LocalOutlierFactor is an unsupervised model which can automatically detect univariate outliers by measuring the "local deviation of the density" of each value, which is given by its k-nearest neighbors [3]. In layman's terms, we are using k-nearest neighbors to detect outliers.

IterativeImputer is an experimental feature from Scikit-Learn, which allows us to impute missing values "by modeling each feature with missing values as a function of other features in a round-robin fashion" [4]; in other words, we are going to impute the missing values iteratively. Moreover, this imputation method is flexible as it allows any estimator (or model) to be used for missing values imputation, making it preferable over KNN-imputing.

GradientBoostingClassifier is a supervised ensemble model, of type sequential. It develops a series of decision trees by calculating the amount of "error" with the real data and assigns a weight to each tree in a similar fashion to gradient descent.

RandomSearchCV is a hyperparameter tuning algorithm, which searches for the best hyperparameters of a model by sampling hyperparameter values randomly. This might be preferable in our case, as the models we will be using have continuous values - making GridSearchCV less suitable, as its brute-forcing nature with continuous hyperparameter spaces can lead to computationally expensive iterations.

HalvingRandomSearchCV is another experimental feature from Scikit-Learn, which attempts to find an optimal set of hyperparameters for a model, using an iterative selection process inspired by genetic algorithms. More precisely, it takes a pre-defined number of candidates, evaluates each of them through cross-validation, and selects the best candidates while removing a portion based on the total number of current candidates [5] - which in our case is half – and iterate over again, changing the parameters slightly. This sophisticated algorithm can help us to converge towards a good set of hyperparameters, rather than just brute forcing it.

Moreover, we used an external tool for data preprocessing purposes – namely feature engineering.

PyGrowUp is a Python library for calculating z-scores for child growth metrics [6]; We used it to calculate age-adjusted BMI scores, which are then segmented into relevant classes. This allows us to calculate one's BMI with as much information as possible, ensuring that our feature engineering process will be as complete.

DATA EXPLORATION

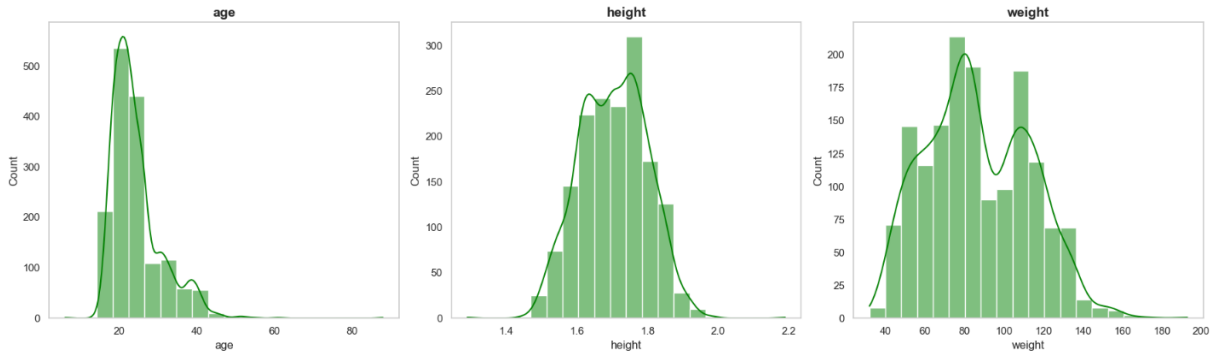
The objective of this section is to summarize the results and observations that can be drawn from the exploratory analysis of the dataset.

UNIVARIATE ANALYSIS

Univariate analysis is carried out to study the main characteristics of each variable. The variables were analyzed with histograms in the case of numeric ones, while with countplots in the case of categorical ones.

Numerical Variables

The most represented age group is between 20 and 30 years old, with an age distribution that has a positive skew. Heights are distributed approximately normally, with an average value around 1.7 meters. Weights present a multimodal distribution with two main peaks, one around 80 kg and the other around 110 kg, showing a slight asymmetry to the right.



Categorical Variables

Eating habits

The count plots show that 87% of participants regularly consume high-calorie foods, and 93% do not monitor their caloric intake. However, consuming vegetables during meals appears to be a positive habit, with 56% always including them and only 5% never doing so. Additionally, 81% reported sometimes snacking between meals and 22% failed to reach the recommended intake of more than 2 liters of water per day.

Parental overweight status

The data indicates that 81% of respondents have at least one overweight parent.

Smoking and alcohol consumption

The data reveals that 97% of respondents do not smoke. Although smoking may seem less relevant for modeling due to limited variation, its potential link to obesity should not be ignored.
Furthermore, 66% of respondents sometimes drink alcohol, while 29% claim to never consume it.

Demographics

Gender distribution among respondents is nearly even, with 51% male and 47% female, and the remainder recorded as null values.

Activities and Transportation

Transportation habits reveal that 73% of respondents primarily use public transport, 21% travel by car, and 3% mainly walk. Device usage shows that 44% of respondents use devices for up to 2 hours per day, while 43% use them for up to 5 hours. Physical activity patterns indicate that 37% exercise 1–2 times per week, while only 6% train more than 5 times per week, representing the group population well.

Bivariate Analysis

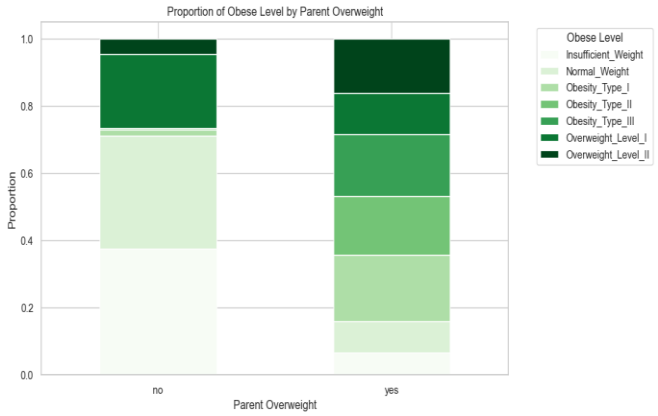
Correlation Matrix for Numerical Variables

The correlation matrix provides a summary of the linear relationships between the numeric variables in the data set. The main correlation is between height and weight and shows a moderate positive correlation with a coefficient of 0.46. This result is in line with the expectation that taller individuals tend to weigh more.

Height vs. Weight by Obese Level

This scatter plot of weight versus height confirms some of the expected pattern, for example the high weights relate to a severe level of obesity, and the smaller weights correspond to normal or insufficient categories. However, different levels of obesity are often covered for people of the same height, indicating that height alone is not a definitive predictor. Furthermore, three outliers stand out.

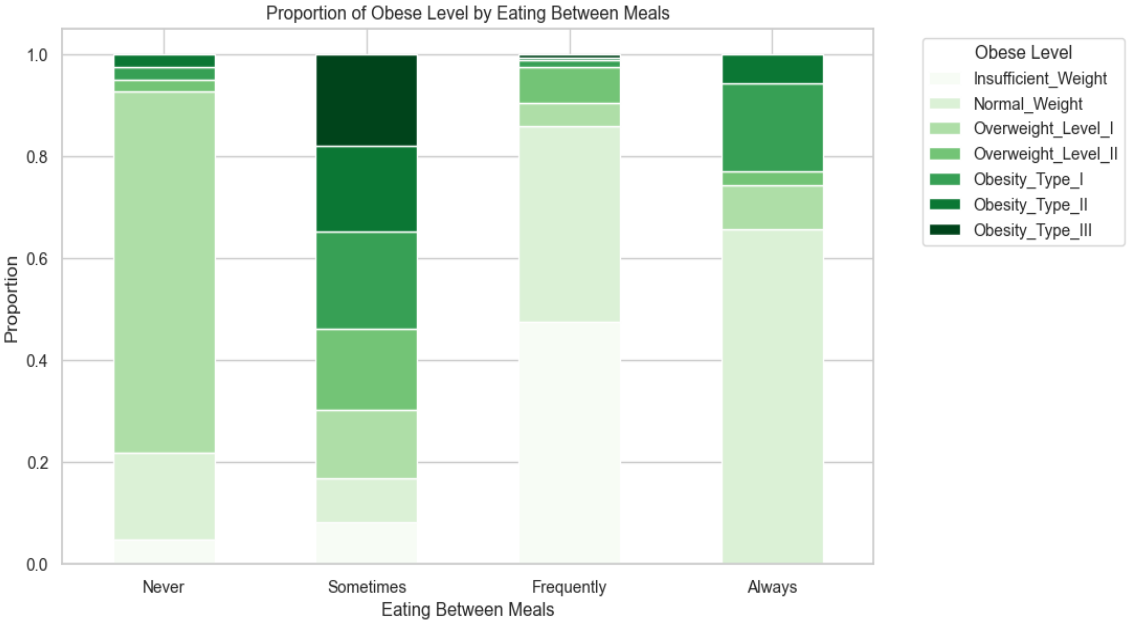
Eating Between Meals vs. Obese Level



The analysis of distribution shows a clear correlation between eating frequency and obesity levels, where most of the individuals are under the "sometimes" category for eating between meals. Obesity_Type_I, Obesity_Type_II and Obesity_Type_III are more present in this category, while insufficient and normal weight are also present, but their counts are lower compared to obese categories, which suggest that "sometimes" eating between meals may be common across all groups but has a greater impact on those with higher obesity levels.

Parent Overweight vs. Obese Level

The graph reveals a strong correlation between the overweight status of parents and obesity levels, as it is observed that people with overweight parents are especially in the highest obesity categories. Instead, people without overweight parents have a more balanced distribution between the levels of insufficient and normal. This trend shows the significant role of hereditary factors or shared lifestyle in influencing obesity.


Proportion of Obese Level by Eating Between Meals
Proportion
1.0
0.8
0.6
0.4
0.2
0.0
Never
Sometimes
Frequently
Always
Eating Between Meals
Obese Level
Insufficient_Weight
Normal_Weight
Overweight_Level_I
Overweight_Level_II
Obesity_Type_I
Obesity_Type_II
Obesity_Type_III

METHODOLOGY

DATA PREPROCESSING

Data preprocessing is a critical step in preparing the dataset for analysis and model training, ensuring that the data is clean, consistent, and ready for accurate predictions. Applied preprocessing process for the purpose of this project involves several key steps:

1. **Dropping irrelevant columns:** Remove *marrital_status* and *region.*
2. **Setting index:** Use *id* column as the index for the dataset.
3. **Handling Outliers:** Remove records with:
    a. *Age* < 16 or > 56.
    b. *Weight* > 167 kg.
4. **Encoding Categorical variables:** Use a predefined hashmap for ordinal encoding.
5. **Handling missing values:**
    a. Apply **KNN Imputer** (numerical) and **Iterative Imputer** (categorical).
    b. Impute with constant value.
6. **Feature Engineering:**
    a. Derive *BMI classifications* based on *weight, height,* and *age.*
    b. Add *life* as a composite lifestyle score.
7. **Scaling numerical features:** Normalize data with StandardScaler.
8. **Spliting data:** Separate features (X) and target (y).
9. **Transforming target variable:** Encode *obese_level* using *hash_obesity*.

DROPPING IRRELEVANT COLUMNS

To simplify the dataset and focus on relevant features, we removed columns ***marrital_status*** and ***region*** from both the training and testing datasets. These variables were deemed non-informative for the prediction task, because they were **uniform** and in the case of **marrital_status** contained only **NULL values.**

HANDLING OUTLIERS

Outlier removal is crucial to eliminate data points that could distort statistical analysis and undermine machine learning model accuracy, thereby preserving dataset integrity and predictive reliability. Utilizing survey information and statistical methods, we identified outliers across three key variables: age, weight, and height.

1. **Age** – observations **outside the range of 16–56** was flagged, due to them being out of scope of the survey conducted.
2. **Weight** – based on the pre-calculated **Inter Quantile Ranges** for numerical variables, weight exceeding **167 kg** was considered suspicious given the dataset context. These records were dropped to ensure data quality and reliability.
3. **Height** – according to the pre-calculated Inter Quantile Ranges observations of **height below 141cm** were excluded from the analysis

We experimented with automated outlier detection techniques (e.g., LocalOutlierFactor), but decided not to use it, as it risked removing too many rows due to non-normal variable distributions.

DATA ENCODING

Categorical variables were manually encoded to numerical variables to retain ordinal relationships where applicable. A custom mapping (hashmap) was created to convert categories into numerical representations (see appendix A). **Manual encoding** was chosen over Scikit-Learn's automatic methods to preserve the ordinal nature of categorical variables, integrate domain-specific knowledge, maintain granular control, and enhance model interpretability through a more precise numerical transformation.

1. Frequency-based and transportation method columns (*alcohol_freq, devices_perday, eat_between_meals, physical_activity_perweek, transportation, veggies_freq, water_daily*) were mapped to increasing integer values reflecting ordinal nature
2. Binary options and gender (yes/no, Male/Female) were also encoded similarly (*monitor_calories, gender, smoke, caloric_freq, parent_overweight*)

HANDLING MISSING VALUES

Our training dataset contains missing values in 16 columns. To address this issue, we applied various imputation methods based on the statistical characteristics of the features and domain knowledge. These methods include the K-Nearest Neighbors (KNN) Imputer, the Iterative Imputer, and filling missing values with a constant value.

It was observed that column *physical_activity_perweek* contained **563 missing values (35%)** and assumed that these columns are not missing at random and act as a distinct category “No physical activity”. Thus, it was imputed with value of 0.

For numerical features (weight, height, age and gender) **K-Nearest Neighbors (KNN) Imputer has** been applied to fill the missing values by leveraging relationships between other variables. To provide accurate results the features first were scaled to calculate proper distances and rescaled back after the process.

Let us observe that as we treated a gender as a number from 0 to 1, we have that by KNN-imputing this variable, we are receiving a continuous value from 0 to 1. To fix this issue, we have rounded it to receive a categorical value, which is either 0 or 1.

Missing values in categorical features have been imputed with Iterative Imputer, where an estimator (KNeighborsClassifier) has been passed over the dataset in iterative process.

For a graphical overview of the missing values imputation, see the table at Appendix C.

FEATURE ENGINEERING

The collected survey data presents significant value for predicting obesity. However, the model performance can be improved by creating new features out of the data. For this model there were two new variables created that are important for model performance: Body Mass Index and Life Index.

Body Mass Index (BMI) is a key metric used to classify individuals into categories such as underweight, normal weight, overweight, and Obesity levels I, II and III. This feature was implemented using age-specific thresholds. For children and adolescents, a standardized metric is used due to the necessity of correcting the BMI by age and gender of the individuals. The thresholds for classifying are predetermined by the WHO and presented in special tables available at their website. [7] This age-

adjusted approach ensures a more accurate assessment for younger individuals. For adult's specific ranges [8] calculated with below formula were applied:

$$\text{BMI} = \frac{\text{Weight in kilograms (kg)}}{\text{Height in meters (m)}^2}$$

See Appendix B for more detailed information about the BMI classes.

To provide a holistic representation of an individual's lifestyle, a composite feature, **Life**, was developed. This feature aggregates multiple variables related to lifestyle habits, offering a single, interpretable measure for analyzing the impact of overall lifestyle on obesity.

The **Life** feature combines the following variables added together: *alcohol_freq, caloric freq, devices_perday, eat_between_meals, monitor_calories, physcial_activity_perweek, smoke, transportation, veggies_freq and water_daily.*

TARGET VARIABLE TRANSFORMATION

The target variable *obese_level* was encoded into integers using a custom mapping *(hash_obesity*) to align with the machine learning model requirements.

FEATURE SELECTION

Feature selection was conducted to identify the most relevant variables for the model, improving its efficiency and performance. Multiple methods were applied, each targeting different aspects of feature relevance

1. **Recursive Feature Elimination (RFE)**

Performed using a Random Forest model to iteratively select features with the greatest impact on predictions.

Output: following features have been selected as important and relevant for the model: *age, alcohol_freq, caloric_freq, devices_perday, eat_between_meals, gender, height, meals_perday, parent_overweight, physical_activity_perweek, transportation, veggies_freq, water_daily, weight, bmi_class, life*

2. **Random forest feature importance**

The Random Forest algorithm was used to assess the importance of each feature in predicting the target, guiding the selection of the most influential variables by measuring how much each feature improves the model's accuracy.

Output: *BMI*, *weight*, *height*, *age*, and *gender* emerged as the **most important features**, ranked in that order. In contrast, *monitor_calories* and *smoke* were the least significant, each scoring **below 1%.**

3. **Chi-Square Test**

Applied to categorical variables, including *alcohol_freq, caloric_freq, devices_perday, eat_between_meals, gender, monitor_calories, parent_overweight, physical_activity_ perweek, smoke, transportation, veggies_freq, and water_daily,* to evaluate their relationship with the target variable. In short, the test evaluates whether the distributions of these variables are independent or if they exhibit significant associations with obesity levels.

Output: each of the tested features was considered important for models under **5% alpha value threshold.**

4. **Correlation Analysis**

Used to evaluate linear relationships between numerical variables, ensuring that highly correlated variables were not redundantly included.

Output: none of the tested features has been removed as none 2 of them have **significant Spearman or Pearson correlation.**

5. **Variance Threshold**

Applied to numerical variables (age, height, and weight) to remove those with low variability, as they add little to model performance.

Output: none of the tested features has been removed as all of them have significant variance.

Considering the above methods results it was decided to drop 3 columns: **monitor_calories, siblings and smoke,** as they were rejected in RFE process and held **less than 2% importance** in Random Forest feature importance test.

For a graphical overview of our feature selection process, see Appendix D.

MODEL ASSESMENT AND SELECTION

1. **Stratified K-Fold Cross-Validation**

We opted for **Stratified K-Fold Cross-Validation** with **10 folds** for model assessment and evaluation. This approach was chosen because it ensures that the class distribution in each fold matches the overall distribution of the target variable. This is especially important for imbalanced datasets, as it prevents models from being trained or evaluated on subsets that misrepresent the actual class proportions.

Using 10 folds provides a good balance between bias and variance in model evaluation:

- A smaller number of folds (e.g., 5) might not fully capture the variability in the data.
- A larger number of folds (e.g., 10) reduces the risk of underestimating performance while keeping computational cost manageable.

By splitting the data into 10 stratified folds, we ensure that all available data is utilized effectively for training and validation without compromising the integrity of the evaluation process.

2. **Macro F1 Average**

We chose the Macro F1 Score as the evaluation metric because it is particularly suited for multi-class classification problems, especially when the dataset exhibits class imbalance. The Macro F1 Score calculates the F1 score independently for each class and then takes the average, assigning equal weight to each class regardless of its size. This ensures that:

The model's performance on minority classes is not overshadowed by dominant classes.

- Each class contributes equally to the overall evaluation metric.

Given that class imbalance can skew accuracy and other simple metrics, the **Macro F1 Score** provides a robust evaluation, helping us identify models that generalize well across all classes.

MODEL OPTIMIZATION

Grid Search CV and Random Search

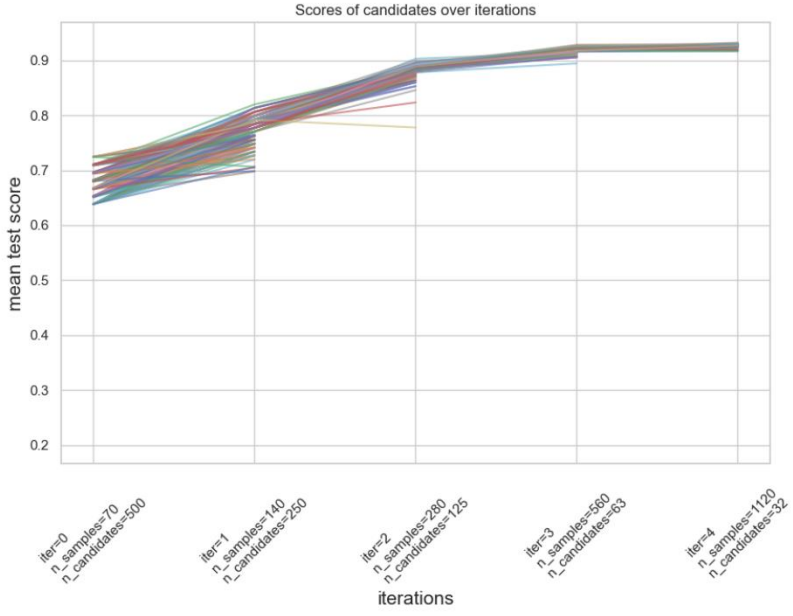
To optimize model hyperparameters, we employed both Grid Search Cross-Validation (Grid Search CV) and Random Search.

1. **Random Search CV**
   Random Search samples hyperparameter combinations randomly instead of exhaustively testing all possibilities. This method is computationally less expensive and can often achieve good results faster than Grid Search, especially when the hyperparameter space is large. Random Search was used as a preliminary step to narrow down the search space before applying Grid Search CV for fine-tuning.
2. **Halving Grid Search CV**

   In addition to traditional Grid Search, we utilized Halving Grid Search, a resource-efficient optimization technique. This approach works as follows:

   - It starts by evaluating a large number of hyperparameter combinations on a small fraction of the dataset.
   - After each iteration, poorly performing combinations are pruned, and only the most promising combinations are retained for further evaluation on larger fractions of the dataset.
   - This process continues iteratively until the optimal combination is found.

   The following image illustrates an example of the Halving Grid Search algorithm.



For both random-based optimization algorithms, we defined the following parameters domain:

Parameter	Domain
Amount of estimators	50,51, 52, ... ,500

Maximum Depth	3, 4, 5, 6, ... , 30
Learning Rate	U(ln(0.001), ln(1))*
Minimum Samples for Split	2, 3, ... , 25
Minimum Samples for Leaf	1, 2, 3, 4, 5
Subsample	U(ln(0.5), ln(1))*

*U(a, b) denotes the uniform distribution between a, b. Observe that U(ln(x), ln(y)) is the log-uniform distribution between x, y.

RESULTS

Outliers Removal

A total of **7 rows** have been identified as outliers and removed from further analysis through statistical analysis.

Model Performance

The following baseline models were assessed using the 10-Fold Stratified Cross-Validation and evaluated based on their Macro F1 Scores on both training and validation data:

Model	Train Score	Validation Score
Gradient Boosting	1.0	0.9323
Bagging	1.0	0.9310
Random Forest	1.0	0.9191
Decision Tree	1.0	0.9133
Multi Layer Perceptron	1.0	0.6753
AdaBoost	1.0	0.3103

As Gradient Boosting had the best validation score, we optimized it and evaluated it. Moreover, we also uploaded their predictions to Kaggle Competition to obtain the test score.

Model	Train Score	Validation Score	Kaggle Score
Optimized Gradient Boosting (with Halving Randomized Search)	1.0	0.9488	0.9548
Optimized Gradient Boosting (with Randomized Search)	1.0	0.9471	0.9599

Another aspect worth analysing is the score on each target class. In this case, we used the Halving Randomized Search-Optimized Gradient Boosting to extract this information.

	Normal_Weight	Overweight_Level_I	Overweight_Level_II	Obesity_Type_I	Insufficient_Weight	Obesity_Type_II	Obesity_Type_III
min	0.857143	0.857143	0.897959	0.943396	0.888889	0.956522	0.958333
max	0.954545	1	0.978723	1	1	1	1
mean	0.897333	0.924004	0.933672	0.969852	0.953842	0.982697	0.993706

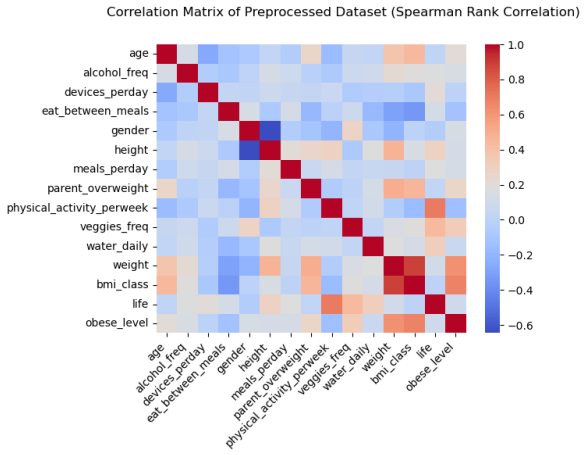
Feature	Importance
bmi_class	26.36%
weight	23.65%
height	7.72%
age	7.43%
gender	6.52%
meals_perday	3.24%
veggies_freq	3.18%
life	3.00%
parent_overweight	2.71%
eat_between_meals	2.61%
alcohol_freq	2.41%
devices_perday	2.04%
physical_activity_perweek	1.96%
transportation	1.96%
water_daily	1.77%
caloric_freq	1.55%
siblings	1.21%
monitor_calories	0.52%
smoke	0.16%

Feature Importances

Another useful point that deserves attention is the numerical values of feature importance, extracted during the feature selection phase. More precisely, they are extracted from the baseline Boosted Gradient classifier.

Post-Cleaning Data Exploration

Our results are not exclusively tied to our trained predictive models. We believe that emphasizing key aspects of exploratory data analysis can be equally important in drawing our conclusions; in particular, we will analyze the preprocessed dataset. First, we calculated the correlation matrix using Spearman's definition as it works both with continuous variables and ordinal variables.

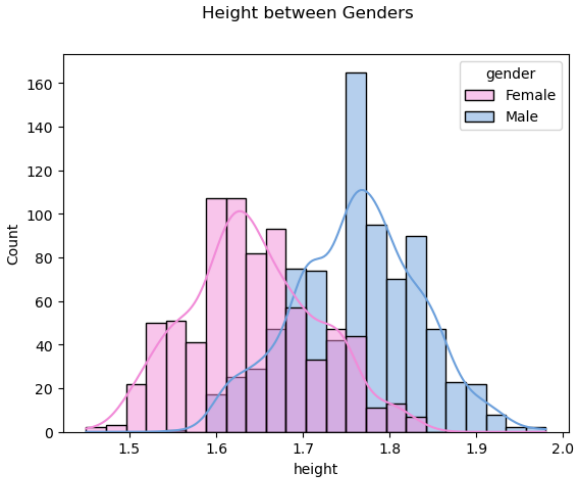


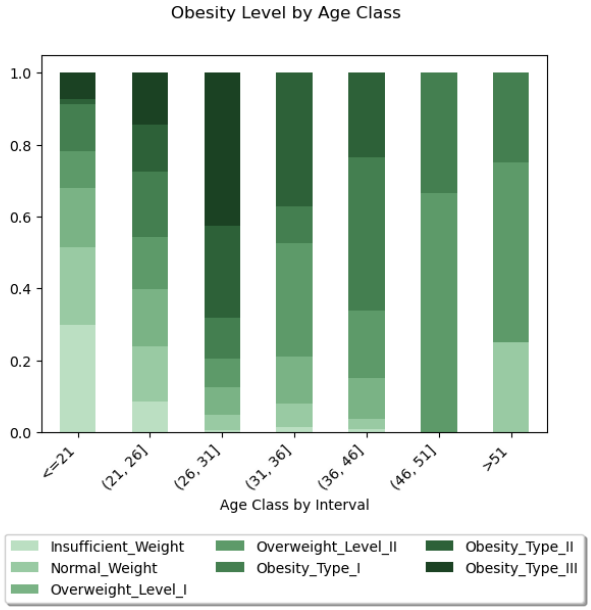
The strongest positive correlations are observed between *weight*, *BMI class*, and *obesity level*, which is expected given their direct relationships. *Physical activity* shows notable negative correlations with obesity, suggesting its protective role against *weight* gain. Lifestyle factors such as *water consumption*, *eating between meals*, and *vegetable frequency* demonstrate moderate correlations with obesity measures, while *device usage* shows relatively weak correlations across most variables. *Parental overweight* status appears to have some influence on obesity measures, indicating a potential hereditary or environmental component. *Age* shows moderate correlations with several factors, suggesting that *weight*-related patterns may vary across different life stages.

By plotting heights separately by gender and age classes for each obesity level, we can gain valuable insights into our problem.

Our selected model indicates that age and gender are significant factors in assessing obesity. To explore these relationships further, the distribution of height by gender and the prevalence of obesity across age groups were analyzed.

The analysis reveals distinct differences in height distribution between males and females. Among males, heights range from **1.4 to 2 meters**, highlighting a broader variation compared to females. These variations underscore the importance of accounting for gender-specific factors when evaluating obesity.

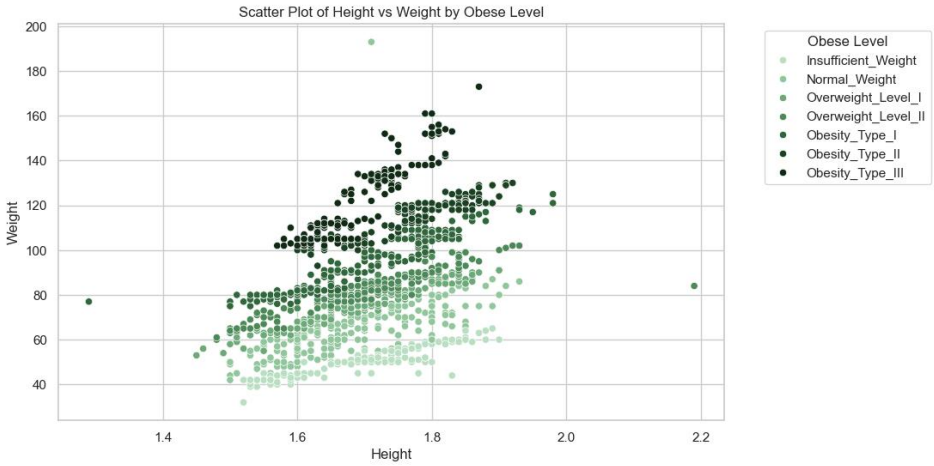




The distribution of obesity among age groups demonstrates that individuals **aged 26 to 31** are the most affected, with over **40% classified as having Obesity Type III**. Notably, cases of Obesity Type III extend beyond this age range, indicating a rising and worsening trend in obesity prevalence over time.

Moreover, individuals born before the advent of the internet era exhibit better overall health, suggesting potential lifestyle or environmental influences on obesity trends. These findings highlight

the need for targeted interventions addressing specific age and gender demographics to combat the growing obesity epidemic.



A clear positive correlation between height and weight is evident, with the data points forming an upward-trending pattern. The distribution shows that individuals classified with higher obesity levels (darker green points) tend to cluster in the upper portion of the plot, typically showing higher weights relative to their heights. Normal and insufficient weight categories (lighter green points) predominantly occupy the lower portion of the plot, demonstrating proportionally lower weights for their heights. It is also important to note that slope gets steeper as obesity levels go higher, which suggests that ratio of weight to height is rising in more obese categories.

DISCUSSION

Model Performance

The Results section highlights that the best performing baseline models are Gradient Boosting and Bagging, both achieving an F1 validation score exceeding 0.93. Furthermore, all models attained a perfect training F1 score of 1.0, underscoring their ability to completely fit the training data. This suggests a risk of overfitting, which was carefully considered when selecting the final model.

We selected Gradient Boosting for hyperparameter optimization because it slightly outperformed Bagging in validation score. With both methods for hyperparameter tuning, the validation score improved significantly, rising from 0.9323 to approximately 0.95 (see Results for more detailed scores). As we considered both optimized models to be "good enough", we uploaded both of their predictions for the test datasets to Kaggle for further analysis.

The model optimized with Random Search showed the best test score at 0.9599; however, we still decided to look at the other optimized model (Halving Random Search). In the case of the Halving Random Search, its results were more stable, with both validation and test scores converging toward 0.95. In contrast, the model optimized with Random Search had a validation score of 0.9471 and a test score of 0.9599, showing signs of instability.

Moreover, since that the Kaggle score accounted for only 30% of the predictions, we deemed it a less reliable metric than the validation score. Having considered all of this, we selected the Gradient Boosting Classifier optimized with Halving Random Search as our best model.

Model's Challenge in Differentiating Normal Weight vs. Overweight Level I

Our best predictive model struggles to accurately differentiate between individuals classified as "Normal Weight" and "Overweight Level I", both achieving scores of 0.8571 (see Results section).
As we delved into this issue, we have found the key issue: BMI. While BMI is effective in identifying individuals who are significantly overweight or obese, it often fails to account for critical factors such as body composition – mainly body fat percentage. For instance, a study [9] shows that BMI is not sufficient to distinguish between physically active individuals (e.g., soldiers in the study) and those who are slightly overweight. Given that our model heavily relies on BMI classes (see Results section), this limitation is likely due to the observed misclassifications.
Therefore, we can simply conclude that this limitation is due to a lack of available data (e.g. muscle mass composition), rather than due to the model itself.
Among other factors that were identified by other studies to have impact on obesity and are not included in our analysis are sleep insufficiency, taken drugs and medicaments, socioeconomic status (link between deprivation and obesity) and stress levels [10].

CONCLUSION

Obesity is a major problem that is caused both by genetic and environmental factors. Thus, it can be prevented, which will enhance the longevity, health and productivity of humans but also relieve the public health system.

It was reviewed, that Machine Learning techniques are commonly used in predicting obesity and helping institutions. Many researchers applied various Machine Learning techniques to solve this problem including Naive Bayes, Decision Trees, SVMs, or XgBoost.

Throughout this project, we aimed to answer the question “Can we predict what causes obesity?”. Thanks to utilization of Machine Learning techniques, we were able to provide a stable Gradient Boosting Classifier Model, that classifies correctly 95%. We have identified 16 features that have impact on obesity, including weight, height, age and BMI that make the biggest difference. Our results suggest also that classification of people in Overweight I and Normal Weight categories is the most challenging and thus exploring different factors is needed.

On the modelling side, picking the best model was based on the model’s performance and stability to produce the same results on different datasets. We have achieved a 95% F1 score both on validation and test sets, which holds a significant premise that the model is ready to be used in broader scope.

We operated only on 1600 observations just from 1 region and imbalanced age groups. Additionally, some of the features were highly imbalanced, which may have caused their removal in feature selection. Though, it should not be assumed that they do not have any impact on obesity if we were studying better representation of world population. Our work can be a foundation to further discover and verify obesity reasons among people living in distinct regions and in distinct conditions. Specifically focusing on diverse groups and verifying whether the socioeconomic conditions and place of living affects people’s obesity level.

REFERENCES

[1] United Nation. Warning 4 Million People Die Anually from Obesity, 2021.

[2] World Health Organization. Obesity and overweight, 2024, https://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight

[3] Scikit-Learn Documentation, https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.LocalOutlierFactor.html

[4] Scikit-Learn Documentation, https://scikit-learn.org/stable/modules/generated/sklearn.impute.IterativeImputer.html

[5] Scikit-Learn Documentation, https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.HalvingRandomSearchCV.html

[6] Official PyGrowUp's README.md file, https://pypi.org/project/pygrowup/

[7] World Health Organization. BMI-for-age (5-19 years) https://www.who.int/tools/growth-reference-data-for-5to19-years/indicators/bmi-for-age

[8] Centers for Disease Control and Prevention. Adult BMI Categories, 2024 https://www.cdc.gov/bmi/adult-calculator/bmi-categories.html

[9] Tyson Grier, Michelle Canham-Chervak, Marilyn Sharp, Bruce H. Jones, Does body mass index misclassify physically active young men, Preventive Medicine Reports, Volume 2, 2015, Pages 483-487, ISSN 2211-3355, https://www.sciencedirect.com/science/article/pii/S221133551500073X?via%3Dihub

[10] B. Masood, M. Moorthy. Causes of obesity: a review, 2023

APPENDICES

APPENDIX A. Conversion table for encoding

Column	Old text value	Converted Value
alcohol_freq	Never	0
	Sometimes	1
	Frequently	2
	Always	3
devices_perday	up to 2	1
	up to 5	2
	more than 5	3
eat_between_meals	Never	0
	Sometimes	1
	Frequently	2
	Always	3
physical_activity_perweek	None	0
	1 to 2	1
	3 to 4	2
	5 or more	3
transportation	Walk	0
	Bicycle	1
	Public	2
	Motorbike	3
	Car	4
veggies_freq	Never	0
	Sometimes	1
	Frequently	2
	Always	3
water_daily	less than 1	1
	1 to 2	2
	more than 2	3

APPENDIX B. BMI categories thresholds

Age Range	BMI Range	BMI Class	Ordinal Value
Children (5-19 years) *Measured in standard deviations	<-2	Underweight	0
	[-2, 1]	Normal Weight	1
	(1, 2]	Overweight	2
	>2	Obesity	3
Adults (20-64 years)	<18.5	Underweight	0

	[18.5, 25)	Normal Weight	1
	[25, 30)	Overweight	2
	[30, 35)	Obesity I	3
	[35, 40)	Obesity II	4
	>=40	Obesity III	5

Appendix C: Missing Value Imputation

Feature	Data type	Number of missing values, share (%)	Imputation technique
Age	Continous	65, 4.0%	KNN Imputer
Alcohol_freq	Ordinal	36, 2.2%	Iterative Imputer
Caloric_freq	Binary	20, 1.2%	Iterative Imputer
Devices_perday	Ordinal	21, 1.3%	Iterative Imputer
Eat_between_meals	Ordinal	59, 3.7%	Iterative Imputer
Gender	Binary	20, 1.2%	Iterative Imputer
Height	Continuous	13, 0.8%	KNN Imputer
Meals_perday	Nominal	9, 0.6%	KNN Imputer
Monitor_calories	Binary	39, 2.4%	Iterative Imputer
Parent_overweight	Binary	20, 1.2%	Iterative Imputer
Physical_activity_perweek	Ordinal	563, 34.9%	Constant Value
Siblings	Nominal	12, 0.7%	KNN Imputer
Smoke	Binary	12, 0.7%	Iterative Imputer
Transportation	Nominal	40, 2.5%	Iterative Imputer
Veggies_freq	Ordinal	26, 1.6%	Iterative Imputer
Water_daily	Ordinal	34, 2.1%	Iterative Imputer
Weight	Continuous	54, 3.4%	KNN Imputer

APPENDIX D. Feature selection methods outputs

Feature	RFE	Feature importance	Variance	Chi-square	Correlation
Age	Keep	Keep	Keep	N/A	Keep
Alcohol_freq	Keep	Keep	N/A	Keep	N/A
Caloric_freq	Keep	Unsure	N/A	Keep	N/A
Devices_perday	Keep	Unsure	N/A	Keep	N/A
Eat_between_meal	Keep	Keep	N/A	Keep	N/A
Gender	Keep	Keep	N/A	Keep	N/A
Height	Keep	Keep	Keep	N/A	Keep
Meals_perday	Keep	Keep	N/A	Keep	N/A
Monitor_calories	Drop	Unsure	N/A	Keep	N/A
Parent_overweight	Keep	Keep	N/A	Keep	N/A

Physical_activity_perweek	Keep	Keep	N/A	Keep	N/A
Siblings	Drop	Unsure	N/A	Keep	N/A
Smoke	Drop	Unsure	N/A	Keep	N/A
Transportation	Keep	Unsure	N/A	Keep	N/A
Veggies_freq	Keep	Keep	N/A	Keep	N/A
Water_daily	Keep	Unsure	N/A	Keep	N/A
Weight	Keep	Keep	Keep	N/A	Keep
BMI	Keep	Keep	N/A	N/A	N/A
Life	Keep	Keep	N/A	N/A	N/A

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa