Estimation of Obesity Levels Based On Eating Habits and Physical Condition

Sara Díez, María Colomer & Gussem Yahia-Cheikh

December 2025

1 Introduction and Goals

This project focuses on the development and evaluation of several supervised learning models in order to address an obesity type classification problem from multiple methodological perspectives. Specifically, we explore a diverse set of algorithms, including Decision Trees, Random Forest, Support Vector Machines (SVM), a Logistic Regression classifier and a Naive Bayes classifier. Each of these models is grounded on different theoretical principles, leading to distinct behaviors in terms of interpretability and sensitivity to data structure; analyzing and comparing them allows to highlight the strengths and limitations of each approach when applied to the same dataset.

Given that no single model outperforms all others across all scenarios, the final stage of the project adopts an ensemble learning strategy. By combining the predictions of the individual models, and comparing both hard and soft voting approaches, the ensemble aims to produce more robust and stable predictions, leveraging probabilistic outputs rather than hard class labels.

With this, our aim is to demonstrate how the aggregation of multiple complementary models can lead to improved predictive stability and overall predictive performance compared to relying on a single algorithm.

2 Initial Steps

First and foremost we imported the dataset using standard data manipulation libraries, specifically with pandas DataFrame, which provides a structured tabular representation that facilitates inspection, preprocessing and subsequent modeling stages.

After, an initial exploration is carried out to verify the data have been correctly loaded. The dataset contains about 2111 observations, each corresponding to an individual, and a set of 17 features describing lifestyle habits and health-related variables. These include both numerical attributes and categorical attributes. The presence of both confirms the need for careful preprocessing in later stages.

The target variable `Obesity_level` represents the obesity level of each individual, is clearly identified at this stage. It is encoded as a categorical variable with multiple classes, reflecting different obesity categories. This confirms that the main task addressed is a **multi-class classification problem**, rather than a regression or binary classification.

3 Pre-Processing Steps

As discussed in the previous section, the dataset contains both numerical and categorical features. To ensure an appropiate preprocessing strategy, variables are grouped according their data type and semantic meaning, allowing transformations consistet with the nature of each feature and the requirements of the models used in this project.

Numerical Variables represent quantitative information expressed in numerical form. This group includes both continuous measurements and discretized numerical values that reflect ordered quantities.

In this dataset, the numerical variables are `age`, `height`, `weight`, and `FCVC` (frequency of consumption of vegetables) which represents an ordered numerical scale rather than a truly continuous measurement. Nevertheless, it is treated as a numerical feature due to its inherent ordering and quantitative interpretation.

Binary Variables encode features with two possible states, typically representing the presence or absence of a behavior or condition. These variables are naturally suited to numerical encoding as 0 and 1, these variables who didn't had the (0,1) value, they were encoded via mapping to these values.

The binary variables in the dataset include `gender`, `family_history_with_overweight`, frequent consumption of high-caloric food (`FAVC`), smoking habits (`SMOKE`), and calories consumption monitoring (`SCC`).

Ordinal Variables have a natural order of the categories. Encoding them using ordinal encoding ensures that the increasing or decreasing levels of frequency are reflected in the numerical respresentation.

The ordinal variables included in the dataset are: consumption of food between meals (`CAEC`), consumption of alcohol (`CALC`), physical activity frequency (`FAF`) and time using technology devices (`TUE`)

Nominal Categorical Variable represent unordered categories, where no essential ranking exists between possible values. In the context of this dataset the variable `MTRANS` (mode of transportation) falls into this category, as transportation methods such as walking, public transport or automobile usage are qualitatively different but cannot be meaningfully ordered.

Using label encoding for nominal variables would introduce artificial numerical relationships between categories. So we opted by using the One-Hot Encoding (OHE); it addresses this issue by transforming a nominal categorical variable into a set of binary indicator variables, one for each category.
Overall this preprocessing strategy ensures that all features are represented in a consistent, meaningful and compatible numerical format.

4 Exploratory Data Analysis (EDA)

In this section we developed the Exploratory Data Analysis (EDA). It was a crucial step to carry out before developing our machine learning models as it aimed to allow a deep understanding of our dataset.

The main focus was to evaluate the quality of our CSV, find out potential issues in order to resolve them plus enable the identification of patterns.

Our goal was to gain knowledge on the dataset structure, identify quality issues such as outliers and skewed distributions, find relationships between the multiple features and target variable as well as gaining information for our preprocessing and future modeling decisions.

4.1 Target Variable Analysis

We decided to develop a target variable plot to visualize the distribution of our target variable.

The plot in Figure 1 illustrated key observations suchlike the distribution not being balanced due to some categories containing more observations and including how 'Obesity_level' variable had a categorical and multiclass character.

This lead us to keep in mind the issue of bias toward majority classes when developing our model along with the influence it has on the evaluation metrics, including accuracy (which later on we will see how it can be misleading).

4.2 Boxplot Analysis

Moving forward, boxplots were generated for our numeric features. We opted to carry out two different approaches: the first plotting each feature by obesity level and the other in the form of a global comparison.

The grouped boxplots allowed us to compare each numeric feature across the obesity categories while the global boxplot (Figure 1) enabled an inspection on the overall spread, skewness and outliers.

Findings included features that showed different median values (indicating relevance when it comes to prediction). Similarly, the appearance of extreme values and the way in variability which tends to increase with higher obesity categories for certain variables.

This indicated that outlier handling was necessary and which features really mattered regarding classification.

4.3 Correlation Matrix Analysis

Following, a correlation matrix was created to assess the relationship between our numeric features. Our goal was to identify highly correlated predictors to note them as sources of multicollinearity. As a result, we could manage properly a feature selection.

The outcome highlighted features that showed both strong and positive correlations, pointing out redundancy. We also noted how certain features correlated moderately with the obesity-related metrics.

4.4 Outliers

As the last step of the EDA, based on the observations made on the outcomes of the target variable plot and the boxplots, we decided it was best if we handled our outliers.

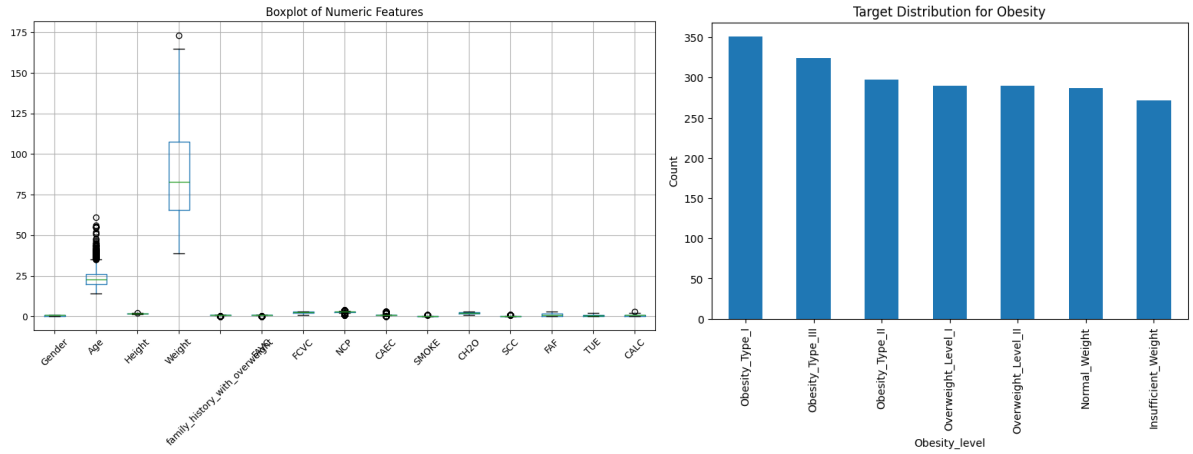

Figure 1: Boxplot of numeric features and obesity's target distribution

4.4.1 Outlier Detection

To detect them we decided to use the IQR method as we were handling human measured variables which tend not to be distributed normally.

We were able to identify that there were numerical feature values far outside the interquartile range. This could mean that the dataset contained measurement errors, rare cases and inconsistencies.

After executing the outlier detection, it was noticed that 10 out of the 15 features contained outliers. Among these, `NCP`(579), `family_history_with_overweight` (385), and `CAEC` (346) contained the highest number of extreme observations.

4.4.2 Outlier Handling

Given that our dataset consisted of 2,118 instances, outlier removal was deemed inappropriate and was not adopted as a handling strategy. That is why we went with IQR Capping.

Consequently, within this method, outliers were removed or capped depending on plausibility while maintaining meaningful variation. In this manner, we could reduce noise and keep intact informative patterns, align the data for model development, and increase data stability.

Right after, we were able to split our data into test and train sets as well as normalize it via Z-score scaling.

5 Modeling

During feature selection, we decided to exclude the `Weight` variable, the reasoning behind being that weight is strongly correlated with the target variable, `Obesity_level`. Including the variable would made the prediction essentially trivial. After treating the dataset, and splitting it into train and test sets (75% and 25%, respectively), we were able to start with the modeling process, where various Machine Learning models were applied to the dataset.

The selected models provide different strengths and flaws, and combining them allows for a comprehensive comparison.

5.1 Tree-Based Models

5.1.1 Decision Tree

First, we started by including a Decision Tree as a fairly simple yet intuitive baseline model, with the aim of assessing the gain in performance through hyperparametrization in the next section with *GridSearchCV*, and through ensembling methods such as Random Forest in later sections.

The initial Decision Tree was trained using the entropy criterion parameter without any more explicit restrictions on depth or node size. This resulted in an accuracy of approximately 0.74 on the test set. Although the performance the model yields is reasonable, it is probably due to overfitting, given the unrestricted growth of the tree, and having a fairly low generalization capability.

5.1.2 Decision Tree with Hyperparameter Tuning

As Decision Trees tend to overfit, tuning its parameters is essential to prevent biased results and obtain the most generalizable model. *GridSearchCV* was used with 5-fold cross-validation to automatically be able to optimize tree parameters, such as maximum tree depth, minimum quantity of samples to split a node, and the maximum leafs a node can have.

The optimal configuration results in a tree with a depth of 17 and 254 leaf nodes, achieving a test accuracy of approximately 0.76, an increase of nearly 2% in performance compared to the baseline model.

Feature importance analysis

5.2 Statistical and probabilistic Models

5.2.1 Logistic Regression Pipeline

Logistic Regression was selected as a baseline classifier because of its compatibility with scaled numeric features, it fits with multiclass classification and how it is easily interpretable. It was implemented as a machine learning pipeline that integrated preprocessing with logistic regression preventing data leakage.

We were able to observe that performance varied among the multiple obesity classes, there was misclassification but still the model performed great on majority classes. Despite obtaining a rather low accuracy, when treating health-related data it is expected. There were indicators of class imbalance and overlapping between neighboring classes, which leads to the model correctly capturing linear relationships but struggling with complex boundaries.

We also decided to compute multiclass ROC AUC in the form of One-vs-Rest (OVR). The high value obtained suggested that the model ranked the correct class higher than others. Meaning that it managed to be highly informative and separate classes very well.

Regardless of having obtained a strong and interpretable baseline, Logistic Regression had clear limitations with our project and desired objectives.

5.2.2 Naive Bayes

Naive Bayes is a conditional probabilistic classifier, which assigns probabilities for each of the possible outcomes. This model assumes conditional independence amongst the features; in real-world data, this assumption does not hold in most cases. Particularly in our dataset, several features show strong correlation, making the assumption largely unrealistic.

However, this model and the results it yields help us understand better the nature of the data we are working with, highlighting how prior assumptions, i.e. conditional independence, affect classification outcomes.

Unsurprisingly, amongst all the models, Naive Bayes yields the lowest accuracy, of approximately 0.4.

5.3 Support Vector Machine (SVM)

Suport Vector Machines (SVM) are well suited for classification tasks due to their strong theoretical foundations and their ability to handle high-dimensional feature spaces. The objective of the SVM is to find the optimal decision boundary that separates the classes while maximizing the margin, defined as the distance between the separating hyperplane and the closest data points form each class, known as support vectors.

Prior to training, feature scaling is applied using `StandardScaler()`, as SVMs rely on distance-based optimization. This ensures that all input features have zero mean and unit variance, improving numerical stability and preventing any single feature drom dominating the decision function.

The model is trained using a soft-margin formulation, where the parameter `C` controls the trade-off between margin maximization and classification error. A non-linear kernel is employed to capture complex relationships between features that cannot be separated by a linear boundary.Under this configuration, the SVM achieves an overall accuracy of approximately 73% on the test set, demonstrating a solid generalization performance.

5.4 Ensemble Models

5.4.1 Random Forest

Random forest is an ensemble learning algorithm that combines the output of multiple decision trees to reach a single result. Each Decision Tree is trained on random subsets of the training data and features, introducing diversity amongst the models.

In this project, the Random Forest classifier was trained using 200 decision trees, allowing for a strong predictive performance. The parameter `n_jobs = -1` was used to allow the algorithm to utilize all available processing cores, and the parameter `Bootstrap=True` allows for each Decision Tree to be trained on a random subset of the data (sampled with replacement). This last parameter is paramount, as it introduces diversity, reducing variance and overfitting, thus improving the generalization capability of the model.

Random forest and the method of aggregating multiple trees achieves robust performance while maintaining readability and interpretability. Especially, it allows for the measurement of feature importance, what enables the identification of the most influential factors affecting obesity levels and the ones that could be disregarded without much loss of information.

5.4.2 Voting Classifier

Lastly, we implemented a voting classifier ensemble model, as it allows for combining several complementary models. The voting classifier algorithm works best when the base models are diverse, as each of them will bring different errors and biases; therefore, we included:

- Logistic Regression as a **linear model**
- Naive Bayes as a **probabilistic model**
- Support Vector Machine as a **margin-based model**
- Random Forest as a tree-based **ensemble model,**

all of them already seen prior in the project.

Model	Accuracy
Logistic Regression	0.509
Naive Bayes (Gaussian)	0.400
Support Vector Machine	0.733
Random Forest	0.877
Voting Classifier (Hard)	0.722
Voting Classifier (Soft)	0.727

Table 1: Accuracy comparison of the evaluated classification models

Table 1 presents a comparison of accuracies amongst the evaluated models. While Random Forest achieves the highest accuracy, approximately 0.88, Voting Classifier models also yield notable results. Moreover, hard voting is *slightly* outperformed by soft voting, demonstrating the advantage of combining probabilistic outputs from diverse models.

It is worth noting that these accuracy values are slightly different from those reported in our prior subsections, as the models used in the ensemble were evaluated with their default hyperparameters, with an exception of setting the parameter `probability=True` for the `SVC` class (and `random_state=RANDOM_SEED` in three of the classifiers).

6 Results & Conclusions

This project aimed to explore multiple supervised learning approaches for a multi-class obesity level classification problem, comparing each model's performance and its predictability capabilities. Although we found ourselves with some difficulties and misconceptions. For instance, we first proposed to reduce the number of target variables from 7 to 3, with the assumption that it would both simplify the problem and improve model accuracy. However, later exploration demonstrated that preserving the initial multi-class structure was more fit to this task. Overcoming the complications allowed us for a deeper understanding of both the data treated and the models used in the process.

In conclusion, the results demonstrate that obesity classification tasks are benefited by modeling with ensemble methods, which highly outperformed tree-based and statistical and probabilistic models. Especially, Random Forest did best at capturing the complex behavior and non-linear relationships of all the variables concerned.