USING MACHINE LEARNING TO PREDICT OBESITY RISK

Iván Martínez Cuevas
Murcia, Spain
imartinezcuevas@gmail.com

ABSTRACT

This work tackles the Kaggle "Multi-Class Prediction of Obesity Risk" challenge, utilizing a data science approach to analyze data derived from a deep learning model trained on the "Obesity or CVD risk" dataset. To predict individuals' obesity risk categories, the study employs an ensemble learning strategy. The approach involves stacking four individual models: Random Forest, XGBoost, CatBoost, and LightGBM. By leveraging the strengths of each model, the stacking ensemble aims to achieve superior predictive performance compared to individual models in classifying individuals into different obesity risk groups.

1 Introduction

Obesity has become a global public health concern, impacting individuals of all ages and backgrounds. This growing problem carries significant consequences, leading to an increased risk of chronic diseases like cardiovascular disease, type 2 diabetes, and certain cancers. Addressing this complex issue requires multifaceted approaches, and Artificial Intelligence (AI) presents a potential avenue for meaningful contribution.

In this study, we leverage the power of AI to contribute to the fight against obesity. We focus on the "Multi-Class Prediction of Obesity Risk" challenge on Kaggle, aiming to develop a robust model capable of accurately classifying individuals into different obesity risk categories. By utilizing data generated from a deep learning model trained on the "Obesity or CVD risk" dataset, we explore the capabilities of AI in identifying individuals at risk, enabling proactive interventions and preventive strategies.

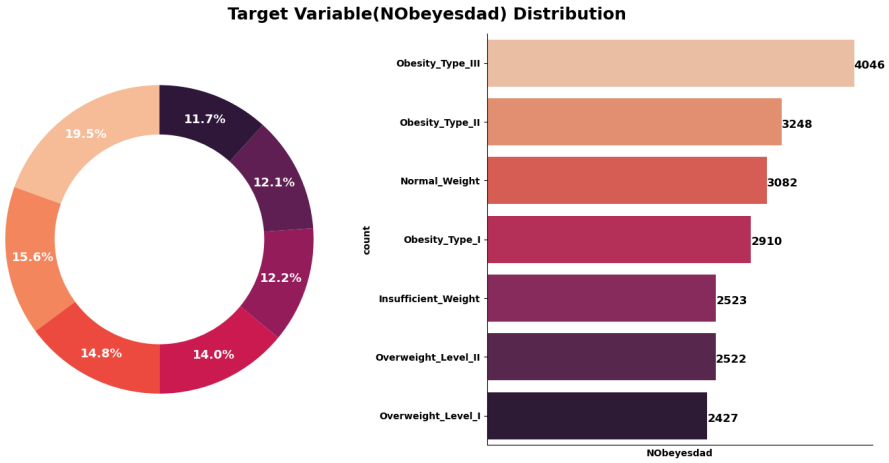
2 DataSet

The dataset for this competition (both train and test) was generated from a deep learning model trained on the Obesity or CVD risk dataset. This dataset consist of 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 condition , data was collected using a web platform with a survey where anonymous users answered each question, then the information was processed obtaining 17 attributes and 2111 records. Feature distributions are close to, but not exactly the same, as the original.

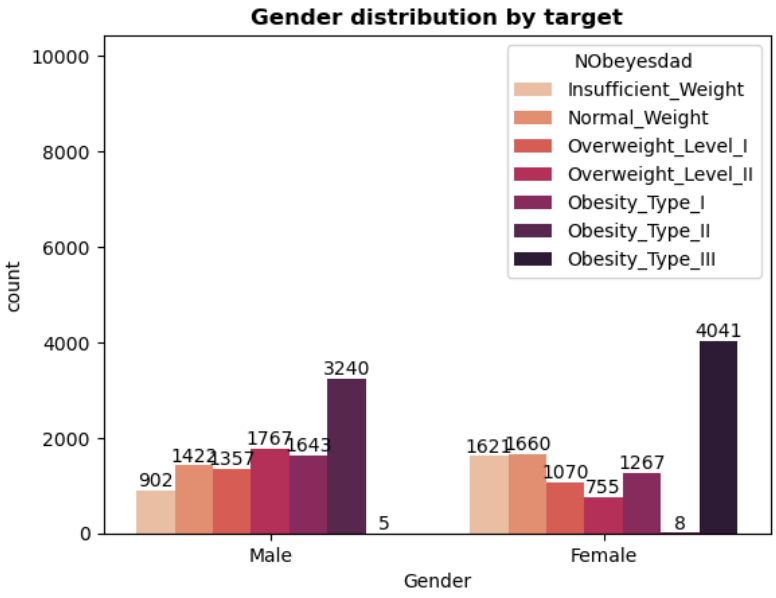
2.1 Categorical Variables

The dataset comprises 9 categorical features. Upon closer examination, it is observed that 5 of these features possess a single value, while the remaining 4 exhibit multiple values. To facilitate analysis, the single and multi-value categorical features are transformed into numerical format using the one-hot encoding technique. This process involves assigning an integer representation to each class, resulting in the conversion of each variable into a set of 0/1 variables corresponding to distinct values.

Within the categorical variables, the target variable is identified, and it is noteworthy that the target variable encompasses 6 different values, displaying a well-balanced distribution, as illustrated in the figure.

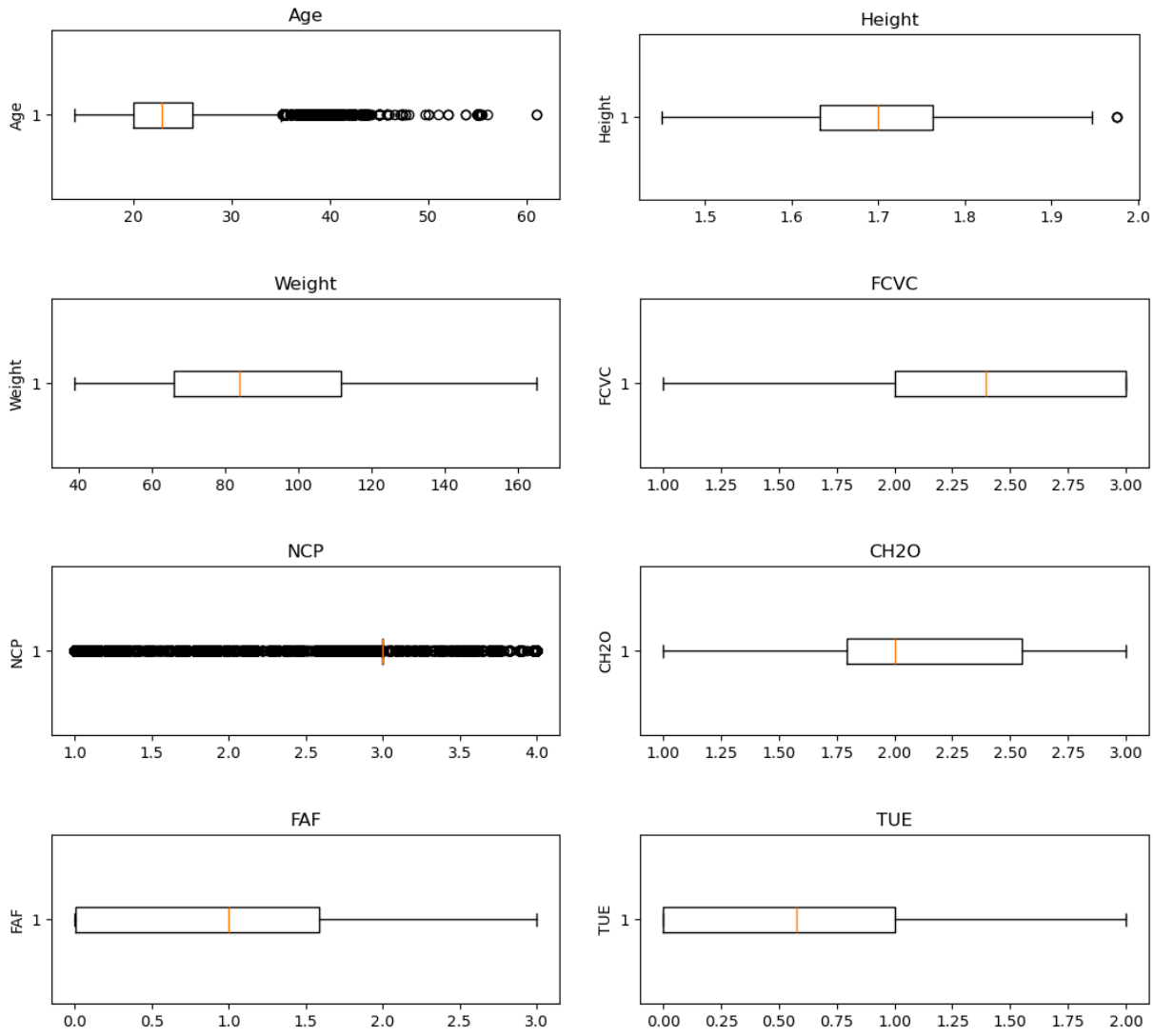


Additionally, the gender variable is observed to be well-balanced. However, upon examining the count of each gender with respect to the target variable, a notable absence of samples is identified. Specifically, there are no instances of men with *"Obesity type III"* and women with *"Obesity type II"* as depicted in the figure



2.2 Numerical Variables

The dataset encompasses 8 numerical variables, and upon preliminary examination, most of them exhibit an absence of outliers. However, a notable exception is observed in the "Age" variable, where the majority of outliers are concentrated. Anomalies are also identified in the "NCP (Number of main meals)" variable, warranting further investigation.



Upon closer analysis of the "Age" variable, the median is found to be approximately 23, with the interquartile range spanning from around 20 to 26 for the middle 50% of the data. The distribution of age displays a right skew, indicating a higher prevalence of younger individuals compared to older individuals. Notably, outliers are predominantly situated in the older age range, starting from 35, suggesting a scarcity of older individuals in the dataset. Additionally, the presence of decimal values in the "Age" variable is noted, raising the suggestion for rounding off, particularly if Age categories are to be created.

Turning attention to the "NCP (Number of main meals)" distribution, a similar issue with decimal values is observed. The occurrence of half meals a day is deemed unusual, resulting in spikes in the frequencies of 1, 3, and 4 meals. Further investigation is essential to understand and address this peculiarity in the dataset.

2.3 Feature Engineering

Two additional features, BMI (Body Mass Index) and BMR (Basal Metabolic Rate), were introduced in an effort to enhance model performance. BMI is calculated by dividing an individual's weight by the square of their height. On the other hand, BMR is computed based on the Harris–Benedict equation, with different formulations for men and women:

- Male: (88.4 + 13.4 x weight in kilograms) + (4.8 x height in centimeters) – (5.68 x age)
- Female: (447.6 + 9.25 x weight in kilograms) + (3.10 x height in centimeters) – (4.33 x age)

The result is an estimated amount of energy in kcal required to maintain the body's basic metabolic activity (without additional activity, so sufficient only for the functioning of the vital organs).

3 Machine Learning Techniques

The dataset, presented in tabular form, has undergone a transformation where all categorical features were converted into numerical types, facilitating the application of conventional machine learning algorithms. In our comparative analysis, we selected Random Forest as the base algorithm, followed by Extreme Gradient Boosting (XGBoost), a non-linear algorithm known for its efficiency with numerical features, minimal requirement for feature engineering, and simplified hyper-parameter tuning.

Lastly, LightGBM, characterized as a fast, highly distributed, and high-performance gradient boosting algorithm, was also included in the analysis. Its notable features include faster training speed and higher efficiency due to its histogram-based algorithm, grouping continuous feature values into bins during training. This results in expedited training and lower memory usage, as it replaces continuous values with discrete containers.

4 Conclusions and future work

In conclusion, our exploration of the dataset involved a comprehensive preprocessing phase, including the conversion of categorical features into numerical types, paving the way for the application of various machine learning algorithms. The selection of algorithms was deliberate, aiming to gauge their effectiveness in handling the dataset's characteristics.

Following a meticulous hyperparameter tuning process using Optuna for each algorithm, a stacking approach was employed to leverage the collective strength of all the tuned models. This comprehensive ensemble strategy led to a remarkable achievement, with an accuracy rate of 90.787% in assessing obesity risk.

The success in hyperparameter tuning and subsequent stacking underscores the effectiveness of the chosen machine learning algorithms and the careful optimization of their configurations. The high accuracy achieved in predicting obesity risk demonstrates the potential of the developed model to make precise and reliable assessments, showcasing the practical application of machine learning in healthcare and risk assessment scenarios.

As a direction for future work, exploring more complex algorithms such as neural networks could provide valuable insights and potentially enhance the model's performance. Neural networks, with their ability to capture intricate patterns and relationships in data, may reveal hidden complexities within the dataset that traditional machine learning algorithms might not capture.

Additionally, considering the use of the original dataset for training and evaluating the model's performance on the private challenge set could offer a robust assessment. This approach enables a more comprehensive evaluation of the model's generalization capability, as the private challenge set serves as an independent and unseen test dataset.

Acknowledgments

References