Data Mining Report

Julia Lorenz 156066, Wojciech Bogacz 156034, Mateusz Nowicki 156064, Krzysztof Skrobała 156039, Jakub Kubiak 156049

April 30, 2024

1 Introduction

In this report we aim to explore different preprocessing techniques and their influence on accuracy of classification of dataset "Obesity Levels Classification" found on Kaggle website.

2 Description of the dataset

The dataset consists of information about individuals from Mexico, Peru and Colombia. The aim is to predict obesity levels based on factors like demographics (age, gender, height, weight), habits (diet, physical activity, smoking, alcohol consumption) and technology use. The dataset contains 2111 records with 16 features each and includes a mix of categorical and continuous data. It's worth pointing out that 77% of the data was generated synthetically.

2.1 Input features

2.1.1 Description

There are 16 features overall. As stated before they can be categorical as well as continuous ones.

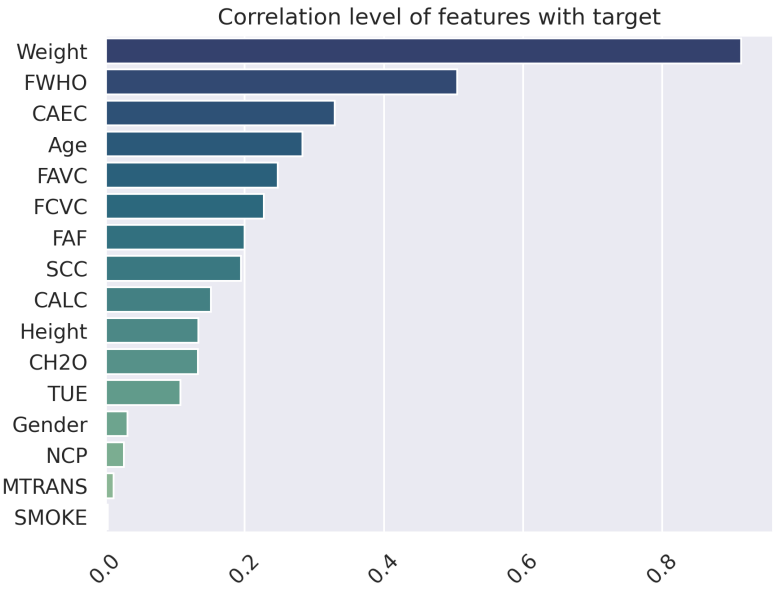
Since substantial amount of data is synthetically generated, some variables can contain unnatural values. For instance feature NCP, which is supposed to contain the number of main meals for the person on daily basis, can contain fractional values like: 1.010319, thus the type is given as Continuous.

2.1.2 Exploratory analysis

We have constructed comprehensive plots to grasp the distributions and potential correlations of our dataset's features. While some variables exhibit non-standard distributions, we have successfully identified several significant relationships, even before performing any preprocessing techniques.

Table 1: Input features explanation

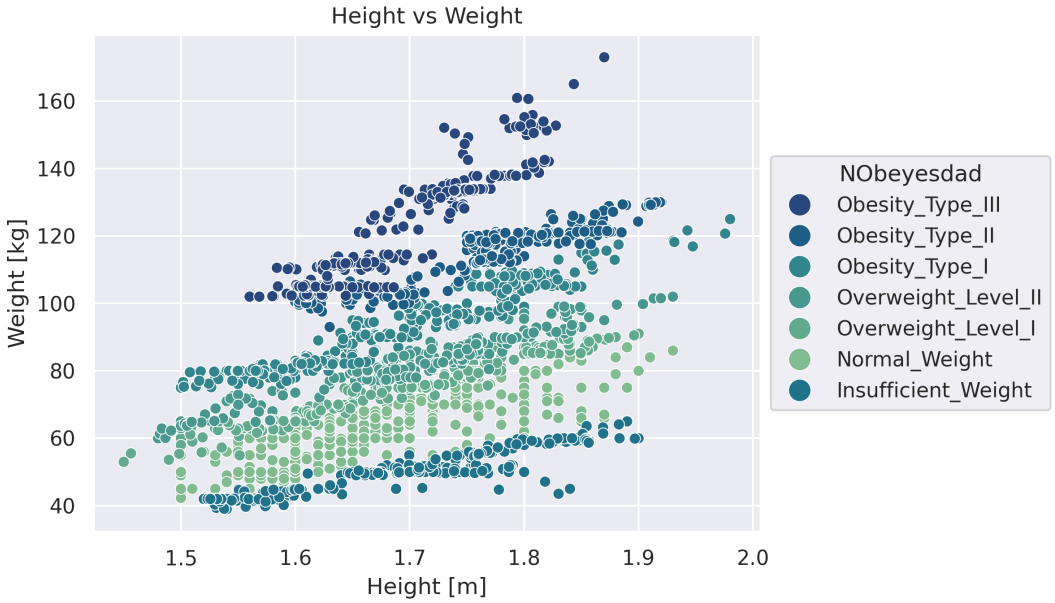
Feature	Type	Description
Gender	Categorical	Gender of the individual
Age	Continuous	Age of the individual
Height	Continuous	Height of the individual (in m)
Weight	Continuous	Weight of the individual (in kg)
FHWO	Binary	Whether a family member suffers from overweight
FAVC	Binary	Whether individual consumes high caloric food
FCVC	Continuous	Frequency of consuming vegetables
NCP	Continuous	Number of main meals daily
CAEC	Categorical	Whether the individual eats between meals
SMOKE	Binary	Smoking status
CH2O	Continuous	Daily water intake (in liters)
SCC	Binary	Whether the individual monitors daily calorie intake
FAF	Continuous	Frequency of physical activity (in days/week)
TUE	Continuous	Time spent on technological devices (in hours/day)
CALC	Categorical	Frequency of alcohol consumption
MTRANS	Categorical	Primary mode of transportation used



Firstly, when examining the correlation between each feature and the target value, it becomes evident that Weight stands out as highly correlated, which comes with no surprise, as it aligns with out intuitive understanding of that the weight is inherently linked to obesity levels. Interestingly, the feature FWHO, denoting the presence of obesity within an individual's family, indicates substantial correlation with the target, reaching about 50%. This underscores the genetic and social predisposition aspect in obesity levels determination.

Additionally, CAEC (indicating snacking habits between main meals during the day) and Age exhibit

notable correlations at 33% and 25% respectively.



Moreover, out exploration extends to the relationship between Height, Weight and the target variable. Visualizing this connection reveals a clear linear separation within the data. This suggests that both Height and Weight play significant role when understanding the target distribution.

3 Preprocessing techniques

- **Removing null values**: Before any analysis, it's essential to check for missing values within the dataset. We have performed the the analysis, fortunately the set was complete (surely the synthetical data generation helped to some extent).
- **Duplicate Identification**: As duplicates can lead to biased results, it is important to get rid of them.
- **Categorical Data Encoding**: Since many algorithms in the libraries we use require numerical input, we have encoded the categorical variables into a numerical format. We have utilized the LabelEncoder for this purpose.
- **Data Normalization (MinMax)**: We have normalized out dataset using MinMax Scaler, this step already vastly improved the accuracy.
- **Correlation Check (Heatmap)**: We have computed the correlations between the variables in our dataset and visually inspected them as a heatmap.
- **Data Standardization for PCA**: Principal Component Analysis (PCA) is a dimensionality reduction technique that identifies the most important features in the dataset. Before applying PCA, it's crucial to standardize the data to have a mean of 0 and a standard deviation of 1 across each feature. Standardization ensures that all features contribute equally to the PCA analysis.

- **Linear Discriminant Analysis (LDA)**: A dimensionality reduction technique that finds linear combinations of features that separates two or more classes.
- **PCA, LDA**: To reduce the dimensionality of data we have tested the behavior of PCA and LDA algorithms. In this study, LDA was chosen as the preferred dimensionality reduction technique due to its effectiveness.

3.1 Pre-processing techniques influence on the accuracy

- Accuracy on encoded data with duplicate values deleted → 67%
- Accuracy after MinMax normalization → 78%
- Accuracy after Principal Component Analysis, having 14 components → 80%
- Accuracy after Feature Selection, choosing attributes 'Weight' and 'Height' → 91%
- Accuracy after Linear Discriminant Analysis, having 2 components → 92%

Classification of the data with redundant values removed resulted in accuracy equal to 63%. Applying MinMax normalization substantially improved that score to 76%. PCA with 14 components further boosted accuracy to 80%. Exploring the relationship between 'Height', 'Weight' and the target variable and selecting only those attributes significantly increased accuracy to 90%, underscoring the importance of relevant feature selection. LDA with 2 components yielded the highest accuracy of 92%, showcasing that LDA effectively reduced dimensionality while preserving class discriminatory information, resulting in superior classification performance.

4 Output features

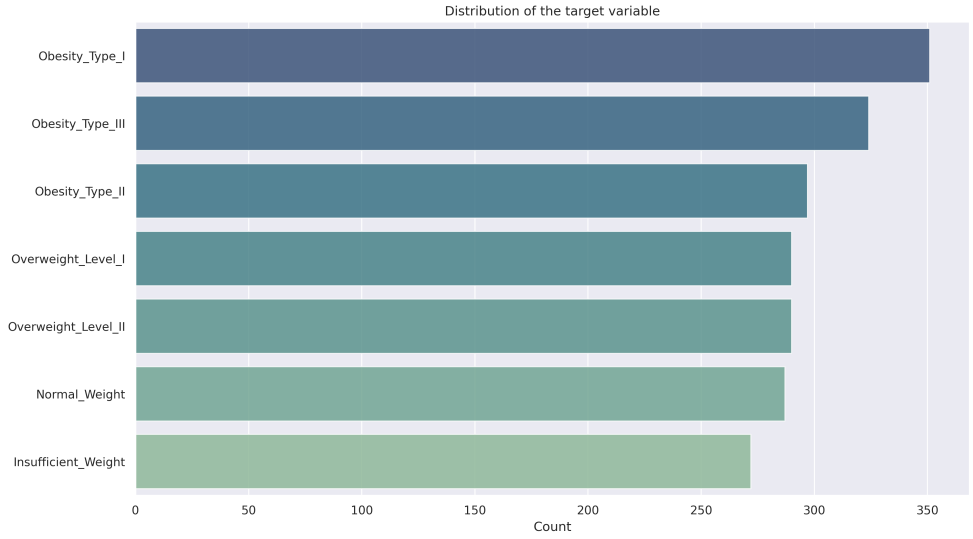
4.1 Description

The output feature (NObeyesdad) in our dataset categorizes individuals into one of seven wight categories based on established criteria. Each weight category reflects different levels of obesity associated with health risks. The categories are:

1. Insufficient Weight
2. Normal Weight
3. Overweight Level I
4. Overweight Level II
5. Obesity Type I
6. Obesity Type II
7. Obesity Type III

4.2 Exploratory analysis

The target variable exhibits a balanced distribution, with each category having similar number of occurrences. It is advantageous, as it ensures that each class is has sufficient amount of instances, leading to more robust and reliable predictions.



5 Conclusions

All of the pre-processing techinques used siginificantly boosted the accuracy of classification of the data. Improving the performance of the classification required a thorough analysis of the input features. Further enhancement of the accuracy would require us to tune the classifier parameters or explore different algorithms. However, in this report we focused on importance of preprocessing techniques, which successfully led us to satisfactory model performance results.