Obesity Project Report

Tianqi Tang

October 2020

1 Introduction

The topic of this project is obesity. This is an important topic because obesity affects many aspects of our lives. The data set[1] in use has 16 features and 1 target variable, NObeysdad. The target variable represents categorical obesity levels consisting of Insufficien tWeight, Normal Weight, Overweight Level I, Overweight Level II, Obesity Type I, Obesity Type II and Obesity Type III. The problem that I will work on is categorical and is about making predictions about a person's obesity level based on other features. The 16 features are Gender, Age, Height, Weight, family history with overweight, Frequent consumption of high caloric food (FAVC), Frequency of consumption of vegetables (FCVC), Number of main meals (NCP), Consumption of food between meals (CAEC), SMOKE, Consumption of water daily (CH2O), Calories consumption monitoring(SCC), Physical activity frequency (FAF), Time using technology devices (TUE), Consumption of alcohol (CALC), Transportation used (MTRANS) and NObeyesdad. There are 2111 data points with no missing data. A brief summary of data values is presented below:

1. Gender: Male: 1068, Female: 1063
2. Age(years): min: 14, max: 61, mean: 24.31
3. Height(m): min: 1.45, max: 1.98, mean: 1.70
4. Weight(kg): min:39, max: 173, mean: 86.59
5. Family history with overweight: Yes: 1726, No: 385
6. FAVC: Yes: 1866, No: 245
7. FCVC: min: 1.0, max: 3.0, mean: 2.42
8. NCP: min: 1.0, max: 4.0, mean: 2.69
9. CAEC: No: 51, Sometimes: 1765, Frequently: 242, Always: 53
10. SMOKE: Yes: 44, No: 2067

11. CH2O: min: 1.0, max: 3.0, mean: 2.01
12. SCC: Yes: 96, No: 2015
13. FAF: min: 0.0, max: 3.0, mean: 1.01
14. TUE: min: 0.0, max: 2.0, mean: 0.66
15. CALC: No: 639, Sometimes: 1401, Frequently: 70, Always: 1
16. MTRANS: Public Transportation: 1580, Automobile: 457, Walking: 56, Motorbike: 11, Bike: 7
17. NObeyesdad: Insufficient Weight: 272, Normal weight: 287, Overweight Level I: 290, Overweight Level II: 290, Obesity Type I: 351, Obesity Type II: 297, Obesity Type III: 324

Many continuous features do not have units because they initially assumed categorical values and then preprocessed during the data set generation. The previous works using this data set includes Obesity Level Estimation Software[2] exploring 3 different methods for data set modeling. It was found that Decision Trees (J48) achieved the best results based on metrics Precision, recall, TP and FP rate. The authors also built a software using J48 and achieved 97.4% precision.

2 Exploratory Data Analysis

In this data set, every feature is identified as either categorical or continuous based on whether it contains 10 or less distinct values. Several figure are then created for data visualization. Here are some interesting examples. An inter-

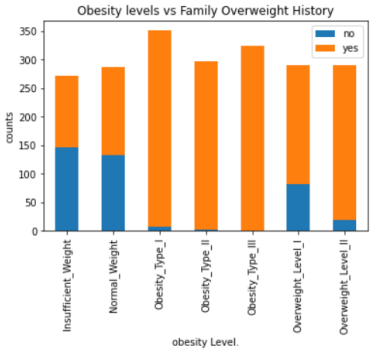

Figure 1: This figure shows how data points corresponding to different obesity levels are distributed. For each level, the bar is divided into two parts with the orange parts indicating people having family overweight history whereas the blue parts show the opposite case.

esting observation about this figure is that people with insufficient weights or normal weights have higher probability of having no family overweight history. Moreover, people with moderate, rather than high obesity levels have the greatest chance of having overweight family members.
Another comparison is between age and daily length of technological devices usage. We notice from this figure that people tend to use less technological

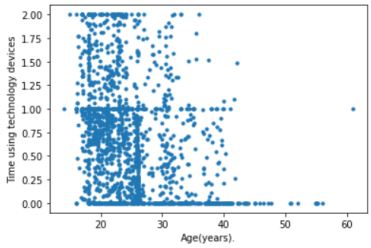

Figure 2: This figure shows how frequently people have technological devices usage based on their age. Many data points assume integer y-values.

devices such as cell phones, TVs, etc., as they grow older.
One more comparison sets between alcohol consumption frequency and water intake.

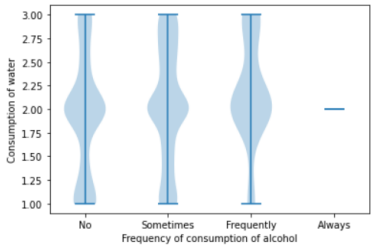

Figure 3: This figure shows how much water people drink based on their alcohol consumption. There's only 1 data point for the person who always drinks alcohol.

It can be seen that there's a severe data imbalance as people who classify as always drinking alcohol are few. Furthermore, for people who never or sometimes drink alcohol, some of them also drink little water whereas it is rare for people who frequently drink alcohol to drink little amount of water.

3 Data Processing

Because this data set is not very large, I will ration 20% of data for testing. For the rest of 80%, I will ration another 20% of them for validating. Therefore, I use $64\%, 16\%$, and 20% of data for training, validating, and testing, respectively. There are several imbalanced features, but MTRANS is quite severe so

I stratified splitting based on this feature. There's only one person who always drink alcohol, so no matter how the data is split, the imbalance of that will occur, so I did not stratified split based on alcohol consumption. I suggest splitting the data in a similar fashion based on these concerns. Since each data point contains information about a unique person who are independent of others, my data is IID with no group structure. It also does not involve a time series. I used 3 preprocessors: OrdinalEncoder, One-HotEncoder and StandardScaler and LabelEncoder on my Data. For categorical non-target features, only CAEC and CALC have ordered structures, so they will be processed by OrdinalEncoder, while the rest of them will be processed by One-HotEncoder. Since StandardScaler is always suitable for continuous features, all of the continuous features will be preprocessed by it. The target variable is categorical and ordered. Therefore, it will be preprocessed by LabelEncoder. Here's a summary of preprocessor usage on features.

1. OrdinalEncoder: CAEC, CALC
2. One-HotEncoder: Gender, Family History with Overweight, SMOKE, SCC, MTRANS
3. StandardScaler: Age, Height, Weight, FAVC, FCVC, NCP, CH2O, FAF, TUE
4. LabelEncoder: NObeyesdad

After preprocessing, there are 25 features and 1 target variable.

References

[1] Palechor, F. M., de la Hoz Manotas, A. (2019)., *Dataset for estimation of obesity levels based on eating habits and physical condition in individuals from Colombia, Peru and Mexico*. Data in Brief, 104344.

[2] De-La-Hoz-Correa, E., Mendoza-Palechor, F. E., De-La-Hoz-Manotas, A., Morales-Ortega, R. C. Beatriz Adriana, S. H. (2019). *Obesity Level Estimation Software based on Decision Trees*, Journal of Computer Science, 15(1), 67-77. https://doi.org/10.3844/jcssp.2019.67.77