TECHNICAL UNIVERSITY OF DENMARK

INTRODUCTION TO MACHINE LEARNING AND DATA MINING

02450

Project 1 - Data: Feature extraction and visualization

Group 127

Lukas Øystein Normann Stjernen– s222992
Maximilian Klein – s222856
Maciej Salwowski – s223525

Mandatory Section

Group member contributions

Section	Contributors and rates
1	Maximilian: 100%
2	Lukas: 100%
3	Maximilian: 100%
4	Maciej: 100%
5	Lukas: 100%

Exam problems

Problem 1

Table 1: Answer for Question 1

Answer	**B**
Explanation	x_1 - ordinal - the intervals are coded as indexes which cannot be added/subtracted, but the smaller the index the earlier interval it indicates. The information about x_1 is enough to pick the correct answer, while the other ones states that it is not ordinal.

Problem 2

Table 2: Answer for Question 2

Answer	**A**
Explanation	Each p-norm can be calculated with the equation $d_p(x,y) = \left(\sum_{j=1}^{M} \lvert x_j - y_j \rvert^p\right)$. For special case of $p = \infty$, d_p becomes highest distance between all the observations. Therefore, $d_{p=\infty} = 26 - 9 = 7$

Problem 3

Table 3: Answer for Question 3

Answer	**A,B,C,D**
Explanation	The get the explained variance, the singular values must be used. They can be found as the diagonal of **S**. Than the explained variance for each principal component is calculated be dividing each singular values by the sum of the singular values. For this case, the explained variance are (in order from first to fifth component): 0.242, 0.218, 0.200, 0.175, 0.165. By adding these values up, the questions can be answered. Thereby it can been shown that all answer are correct, with the exact values being A: 0.835, B: 0.54, C: 0.46, D: 0.66

Problem 4

Table 4: Answer for Question 4

Answer	**C**
Explanation	The five columns of the **V** matrix are the five principle components found from doing a PCA analysis. The value of a projection onto a principle component n is decided by the coefficients in the column of the V matrix vector v_n. Since the observation description in option C fits the coefficients of its described principle component v_3 this seems like the most correct option.

Contents

1 Description of the data set

The given data set contains information about the physical conditions (age, height and weight), eating habits (e.g. daily vegetable consumption) and a few more health related habits (e.g. smoking) from individuals from Colombia, Peru and Mexico. It consists of 17 features and 2111 observations. Of the observations, 23 % were created using questionnaires and 77 % were generated from the questionnaires data by a machine learning tool called Weka and a filter called SMOTE. The data set was created with the intent to be used to train computational tools for obesity prediction and can be downloaded from the enclosed repository [1].

As described in [2], the data for weight and height from the questionnaire was used to determine the *Body Mass Index* (BMI). The BMI was then used to assign the observations to classes with the following criteria:

- Underweight: Less than 18.5
- Normal: 18.5 to 24.9
- Overweight: 25.0 to 29.9
- Obesity I: 30.0 to 34.9
- Obesity II: 35.0 to 39.9
- Obesity III: Higher than 40

After adding this classification as a feature to the data set a very unbalanced distribution towards a "Normal" BMI classification was discovered. This could pose a problem for machine learning methods, since the very unbalanced distribution could introduce bias. For this reason, Palechor and Manotas decided to add observations using the SMOTE filter and Weka tool in order to balance out the distribution of obesity levels.
With this data set we aim to predict obesity based on the physical conditions, eating habits and health related habits of a person. To approach this task as a regression problem, *weight* is used as a continuous response y. For a classification task, the *BMI class* will serve as the discrete response y. For both tasks the features *weight* and *BMI class* will not part of the data matrix X. For the following, the classification of the BMI classes will be denoted as the primary machine learning task As the data is derived from a questionnaire, some features are not numerical, but in form of text. Depending on the type of attribute the text will be transformed into a numerical value. Binary ordinal values will be transformed into a 0/1 format. Discrete ordinal values with x-different entries will be translated into an integer 0 to x scale. For discrete nominal attributes, One-out-of-K coding will be applied.

2 Attributes explanation

Table 5 describes all the attributes found in the original data set. Some of the attributes that would usually be discrete are continuous in the data set. This is because large part of the data has been synthesized. The most prominent example of this is the attribute for the age of the person. A human would usually describe their age in integer values, this is not the case for the synthesized data. This is however not an issue for the purposes of this report. In contrast to the feature labels provided by the authors, we decided to change the short name of two attributes: one related to family history with overweight, the other describing class.

Table 5: Properties of the attributes in the data set.

Id	Attribute name	Short for	Continuous/ Discrete/ Binary	Attribute type
1	Gender	N/A	Discrete	Nominal
2	Age	N/A	Continuous	Ratio
3	Height	N/A	Continuous	Ratio
4	Weight	N/A	Continuous	Ratio
5	FHO	Family history with overweight	Binary	Nominal
6	FAVC	Frequent Consumption of high caloric food 6	Binary	Nominal
8	FCVC	Frequency of consumption of vegetables	Continuous	Ratio
9	NCP	Number of main meals	Continuous	Ratio
10	CAEC	Consumption of food between meals	Discrete	Ordinal
11	SMOKE	N/A	Binary	Nominal
12	CH2O	Consumption of water daily	Continuous	Ratio
13	SCC	Calories consumption monitoring	Binary	Nominal
14	FAF	Physical activity frequency	Continuous	Ratio
15	TUE	Time using technology devices	Continuous	Ratio
16	CALC	Consumption of alcohol	Discrete	Ordinal
17-20	MTRANS	Transportation used	Discrete	Nominal
N/A	BMI_class	BMI classification	Discrete	Ordinal

Table 6 shows summary statistics for the numerical attributes. Table 7 summarizes the discrete attributes in the data set.

Table 6: Summary statistics for continuous attributes.

Attribute	Mean	Median	Standard Deviation
Age	24.3126	22.7779	6.346
Height	1.7017	1.7005	0.0933
Weight	86.5861	83	26.1912
FCVC	2.419	2.3855	0.5339
NCP	2.6856	3	0.778
CH2O	2.008	2	0.613
FAF	1.0103	1	0.8506
TUE	0.6579	0.6254	0.6089

Table 7: Discrete attribute details

Attribute	Most Frequent Value	Percentage
Gender	Male	50.59%
FHO	yes	81.76%
FAVC	yes	88.39%
CAEC	Sometimes	83.61%
SMOKE	no	97.92%
SCC	no	95.45%
CALC	Sometimes	66.37%
MTRANS	Public_Transportation	74.85%
BMI_class	Obesity_Type_I	16.63%

3 Data visualization

When analysing a new data set, visualisations can help understand the structure of the data. Especially of interest are the questions of how the attributes are distributed and potential correlations between the attributes. With this data set there are two challenges to overcome for visualisation. First, the attributes are very diverse regarding their types. Secondly, the high amount of attributes results in a great amount of potential correlations between variables. A compact way to check for correlations between variables and also visualise the distributions of the variables is a pair plot. It consists of scatter plots for each possible combination of attributes. Thereby, on the diagonal the distributions are displayed. This is shown in Figure 1 for five exemplary variables. The variables or all ether ratio or ordinal.

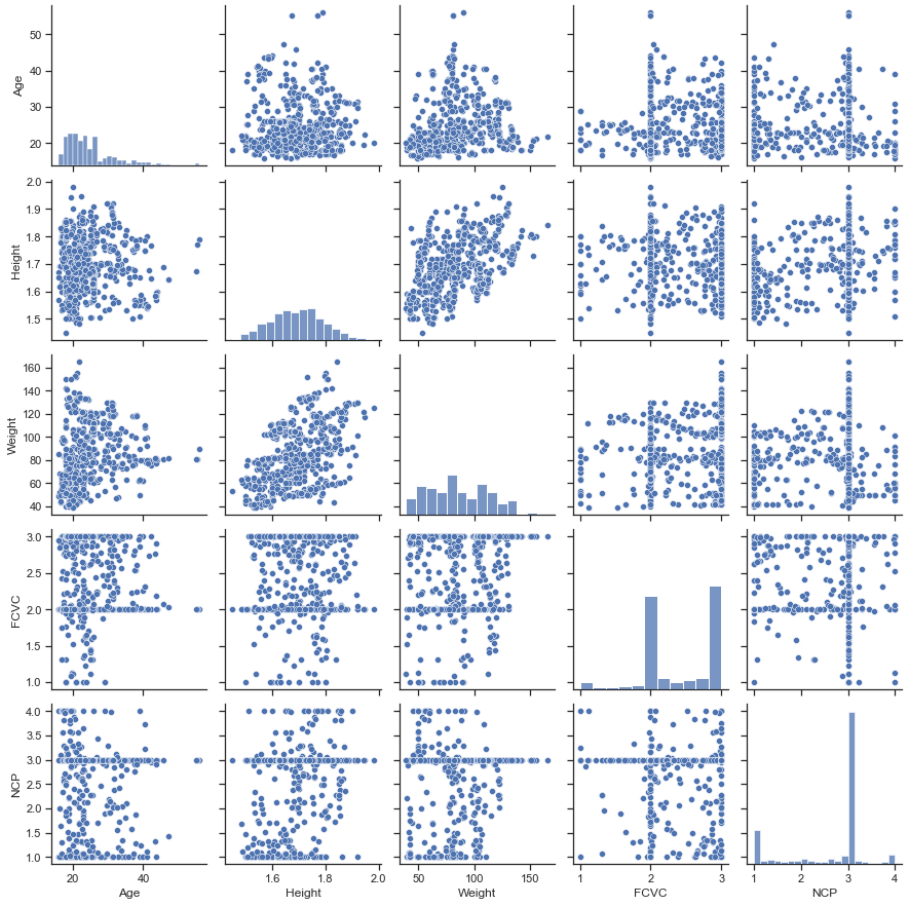

Figure 1: Pair plot of age, height, weight, frequency of vegetable consumption (FCVC) and number of main meals (NCP). To improve the clarity of the plot, the figure shows just 30 % of the observations, which were randomly chosen.

In the pair plot two different types of attributes can be distinguished. The attributes *age*, *height* and *weight* all seem to be normally distributed (*height*) or nearly normally distributed (*age* and *weight*). The variable *weight* could be slightly bimodal distributed. For the two other variables the observations agglomerate at specific modals. This is due to these

variables originally being discrete values, for which observations in between the discrete values were added using the Weka tool. Such behaviour can also be clearly seen in the scatter plot, were data points overlap at specific values. This kind of distribution also applies to the remaining variables not shown in Figure 1. One positive correlation can be seen in the scatter plots between the height and weight. The taller a person is, the more the person tends to weigh. Nonetheless, scatter plots are a sub-optimal way of getting insight into this data set. One type of attribute which can be easily visualized are binary attributes. Therefore, in Figure 2 the influence of three binary attributes on the response variable *weight* is visualized. To achieve this, the distribution of *weight* was split between the possible values of the binary attribute. The result are two different distributions for each binary attribute. The three investigated variables are FAVC (frequent consumption of high caloric food), FHO (Family history of overweight) and SCC (counting caloric intake).

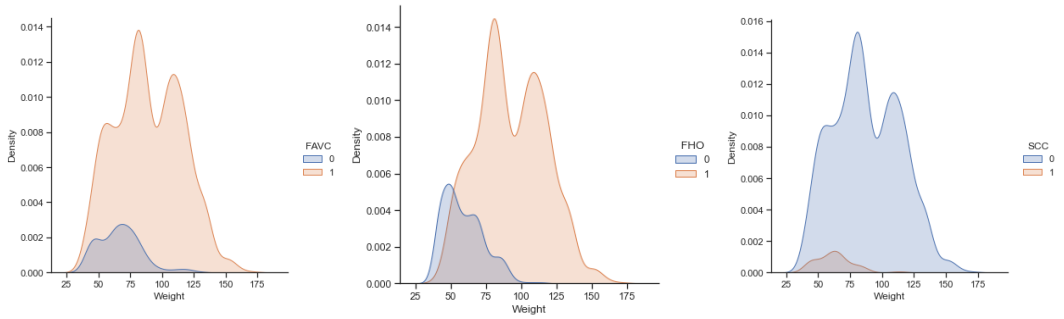

Figure 2: Distributions of weight for three different binary variables (0 = no, 1 = yes).

To answer the question if eating high caloric, having overweight persons in the family or counting the caloric intake is a predictor for overweight, the positions of the *weight* distribution must be observed. For all three variables it is visually clear that the distribution curves for each value of the variable are centered around different values. The height of the distribution curves do not matter in this case, since it just refers to which value is more common. All in all, it can be referred from Figure 2 that frequently eating high caloric food, having overweight persons in the family and not counting the caloric intake is linked to having an increased weight.

For more complex correlations these simple visualisations tend to struggle with revealing all the information given. It is therefore also to early to make assumptions about the success of the primary machine learning task. Therefore, principal component analysis will be applied in the following section.

4 Data analysis

4.1 Preprocessing

As mentioned in the first chapter the real data (the records that had been gathered by the online survey) accounts for less than a quarter of the whole set. The reason behind adding the artificially generated answers was an immense imbalance towards the *normal weight* class [2]. The problem of unequal distribution is one of the major ones that Machine Learning techniques are facing. Most of the times, if the data set is not corrected, the results of classification and/or regression are hardly reliable. In the state-of-the-art articles numerous solutions for that issue are suggested [3]. In this case, in order to resolve the imbalance problem, the *oversampling method* was used. A detailed description of how the additional data was generated can be found in [2]. It is important to say, that during this process the author decided to get rid of the atypical data which included records with missing values and the outliers. The absence of outliers is also shown later. Eventually the authors achieved the expected goal, as shown in Figure 3.

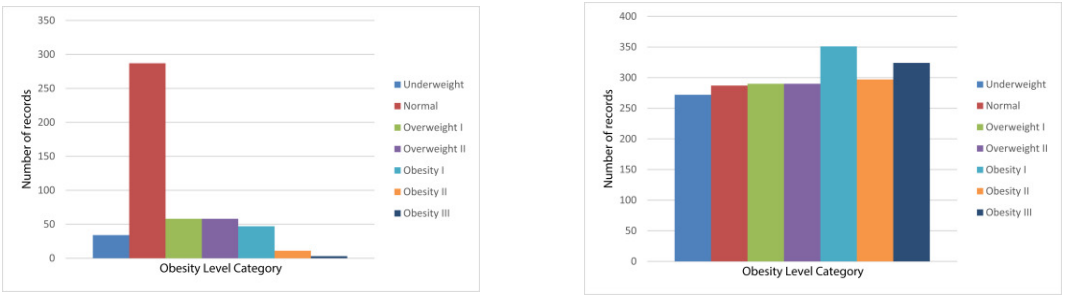

Figure 3: Data distribution between classes before (left) and after (right) oversampling. (source: `www.sciencedirect.com/science/article/pii/S2352340919306985/`)

Apart from above mentioned preprocessing techniques we decided on applying the normalization step. Having that big number of features which differs in nominal value significantly it was crucial to transform the data in this way in order to obtain reliable results.

4.2 Feature correlation

In order to find correlations we have conducted analysis between each pair of the features. The data set consists of 21 attributes after One-out-Of-K coding and because of that it is impossible to illustrate every possible combination. In the whole set of attributes there are pairs which are associated. Most of the proven correlations were obvious, i. e. it is known that taller people tend to weigh more than shorter ones, so that positive correlation between height and weight is clearly visible, and people who are born in the families with obesity history are more likely to weigh more (Figure 4).

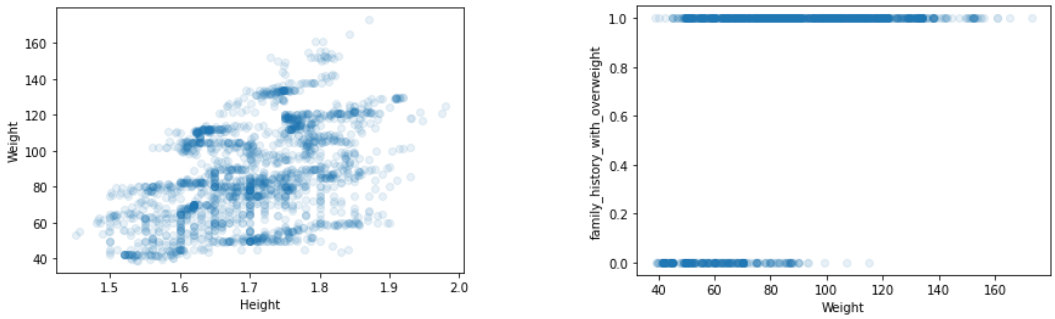

Figure 4: Features of weight and height (left), weight and family history with overweight (right) plotted against each other.

On the other hand there are some features that have shown correlations that were unexpected but could not be seen at first glance. Figure 5 for example reveals that there is a positive correlation between height and daily water consumption.

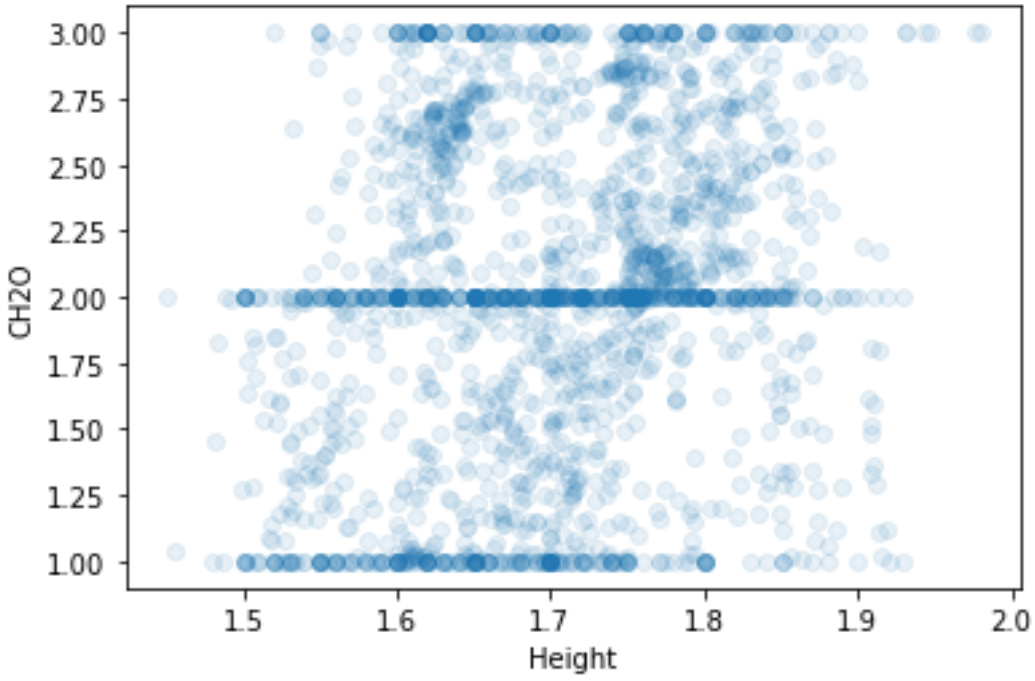

Figure 5: Features of height and weight plotted against each other.

4.3 PCA

In the Figure 6 we can clearly see that the variance of the data is utterly split across the principal components. This suggests that with each attribute different information is encoded, which means that there are hardly any possibilities for dimensions reduction. It is worth mentioning that although first three principal components account for above 30%, further

group of 15 attributes differ by only up to 1%. The figure also shows that we need 15 principal components in order to explain the variation with the threshold of 90%.

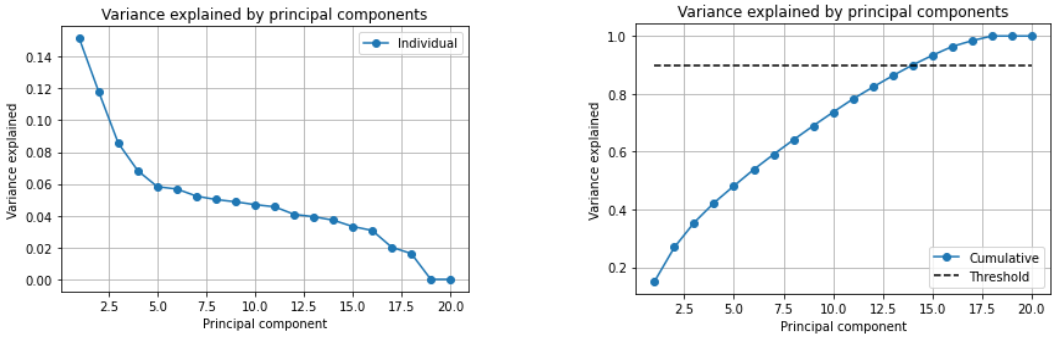

Figure 6: Individual (left) and cumulative (right) variance explanation by principal components.

Due to the fact that the data consists of a vast number of components that influence the output, it is challenging to choose which ones should be looked at thoroughly. Apart from that we would like to pay greater attention to the first three principal components (PCs), which comprise nearly one third of the variance. It is clearly visible in Figure 7 that this group explains features associated with the personal information (attributes with indexes ranging from 1 to 4) and transportation used (indexes 5 to 20) by the surveyed. On the other hand, the features related to eating habits and physical condition (indexes 6 to 15) are explained by the rest of components, what can be seen in the plot containing PC7, PC8 and PC9.

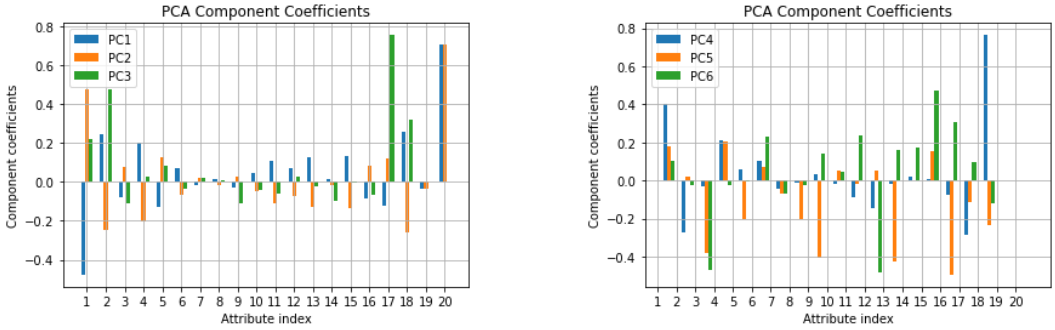

Figure 7: Level of component coefficient for the features in terms of first (left) and latter (right) principal components (indexes explanations can be found in Table 5).

While we were conducting the analysis on each possible pair of PCs, we realised it is hard to distinguish the 6-scale classes model from the data plotted against the principal components. Additionally, single PCs do not distinguish the class of data. This behaviour was expected due to the high spread of importance of each component, which was explained in the previous paragraph.

Thus we decided to transform the 6-classification into a 3-classification for the visualization purpose. As each class concerned the range of variation of BMI, this approach does not affect the reliability of the analysis.

Looking at the Figure 8 it is visible that while data points with relatively low BMI levels are spread mostly alongside principal component number 3, the other ones lies down the PC2 direction. Although we cannot precisely distinguish each class we can see the trend that the classes concerning overweight tend to be more concentrated in terms of those principal components.

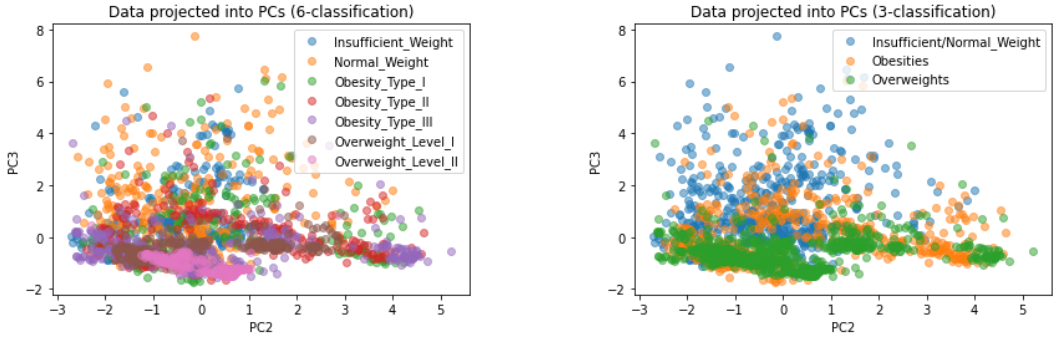

Figure 8: Data plotted on two principal components labelled with 6-scale (left) and 3-scale (right) classification.

5 Discussion

A lot has been learned and quite a few insights have been gained through analyzing and preprocessing this data set. The data set contains a lot of attributes of different types. By using preprocessing methods such as one-out-of-K encoding, we were able to analyze the data set more efficiently. Visualizing the data showed that some binary attributes were indicators for differences in weight. It also showed that simple scatter plots were not sufficient in visualizing correlations between the more complex variables. To combat this, principal component analysis was used to reveal the complexity of information that can be found in the data set. This suggestion undoubtedly has to been taken into consideration when building models for classification and regression. When looking at the principal directions found in the analysis (Figure 7), the contributions of the attributes to each principal component varied vastly. This makes precisely predicting the outcome of the main machine learning aim difficult. However, based on the findings in this report, predicting BMI class using classification within a reasonable margin of error seems feasible.

References

[1] https://archive.ics.uci.edu/ml/machine-learning-databases/00544/.

[2] Alexis de la Hoz Manotas Fabio Mendoza Palechor. Dataset for estimation of obesity levels based on eating habits and physical condition in individuals from colombia, peru and mexico. *Data in brief*, 25, 2019.

[3] Xinjian Guo, Yilong Yin, Cailing Dong, Gongping Yang, and Guangtong Zhou. On the class imbalance problem. In *Fourth international conference on natural computation*, volume 4, pages 192–201. IEEE, 2008.