Statistical Analysis for Pre-treatment of dataset before Classification Analysis

All of the features in the dataset were standardized to avoid bias and to bring all values of the variables in common scale without distorting differences in the range of values. Standardization is important, as later in KNN analysis the values of the features need to be in the same scale or else some features might be falsely weighted too high or too low when calculating distances. Figure 3 below shows a visual example of the distribution of one categorical turned binary feature in histogram before and after standardization.

Binary feature CALC Before Normalization **CALC After Normalization**

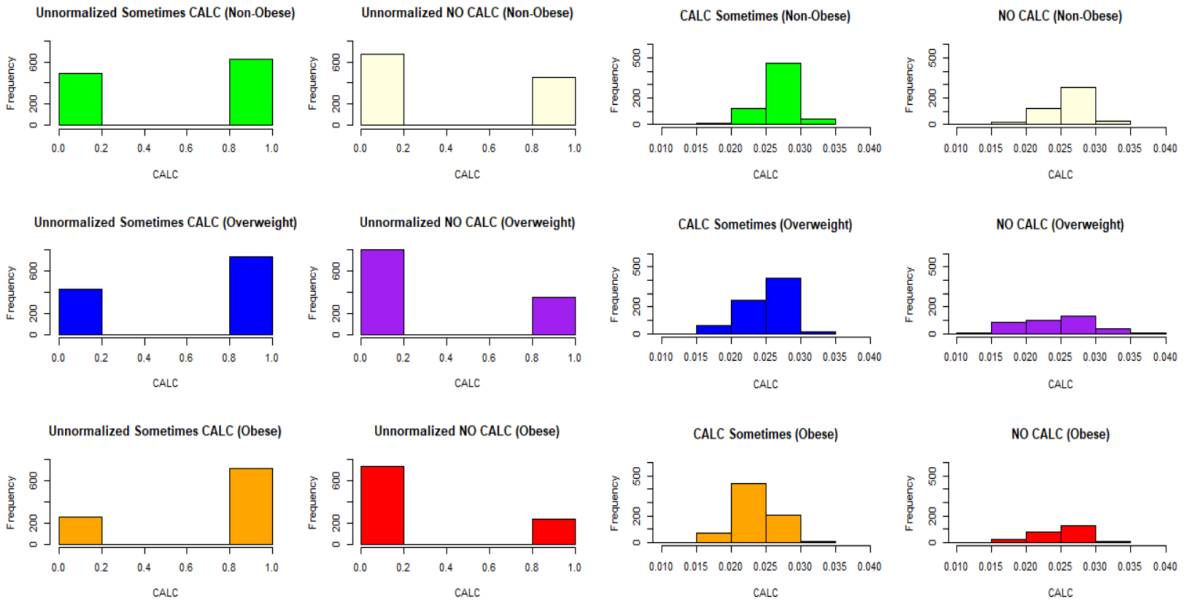

Figure 1. Visual Illustration of Before and After Normalization.

It is interesting to determine if any of the features discriminates well or poorly between two classes. In order to determine discrimation power, the histograms of all 14 features were plotted and compared for each class Non-Obese (CL1), Overweight (CL2) and Obese (CL3). Then, after visually comparing histograms, few features were narrowed down for comparison based on distinct looking distribution among classes between histograms. The selected features were Age, number of main meals (NCP), daily consumption of water (CH2O), Time using technology devices (TUE), binary features Sometimes CALC and No CALC where CALC means consumption of alcohol and binary features Automobile Transportation and Public Transportation. In Figure 4, it can be seen that the histogram of Age follows similar distribution but histogram for Age of Non-Obese seems skewed right whereas the the histogram for Age of Obese seems shifted to the little right. Similar patterns can be observed in histograms of NCP between classes. In Figure 5, the

histograms of CH2O between all three classes look very similar making it a poor discriminating feature. The Obese class in CH2O seems to have smaller highest frequency compared to other two class. Similarly, histograms of TUE feature are skewed right and Obese class and Overweight class have lower frequency than Non-Obese class which suggests it carries better discriminating power. In Figure 6, distributions of highest frequency are in different intervals in the histograms of Sometimes CALC making it a good discriminating feature. In the histograms of NO CALC, Non-obese class seems to have concerntrated frequency around 0.025 and classes of Overweight and Obese seems to have very few observations. In Figure 7, the binary features of Automobile transportation shows few yet sharing different intervals among classes and public transportation have similar patterns among different classes indicating poor discriminating features.

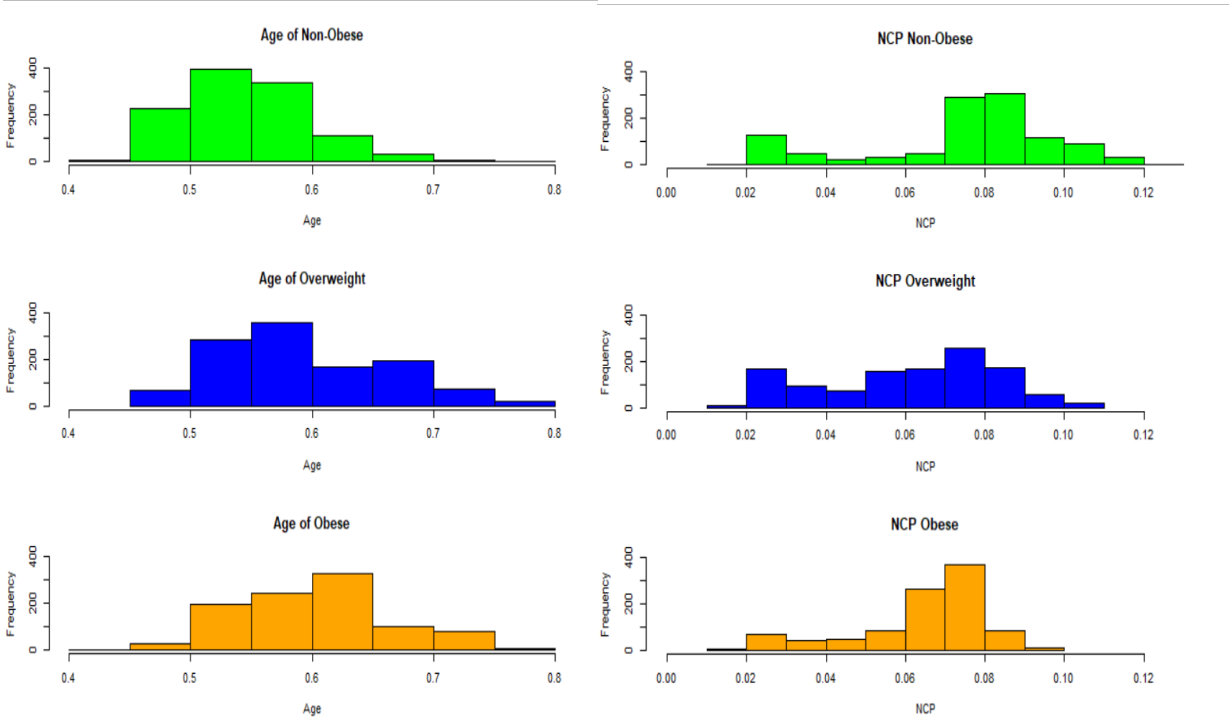

Figure 2. Selected Feature Age and Selected Feature NCP (number of main meals)

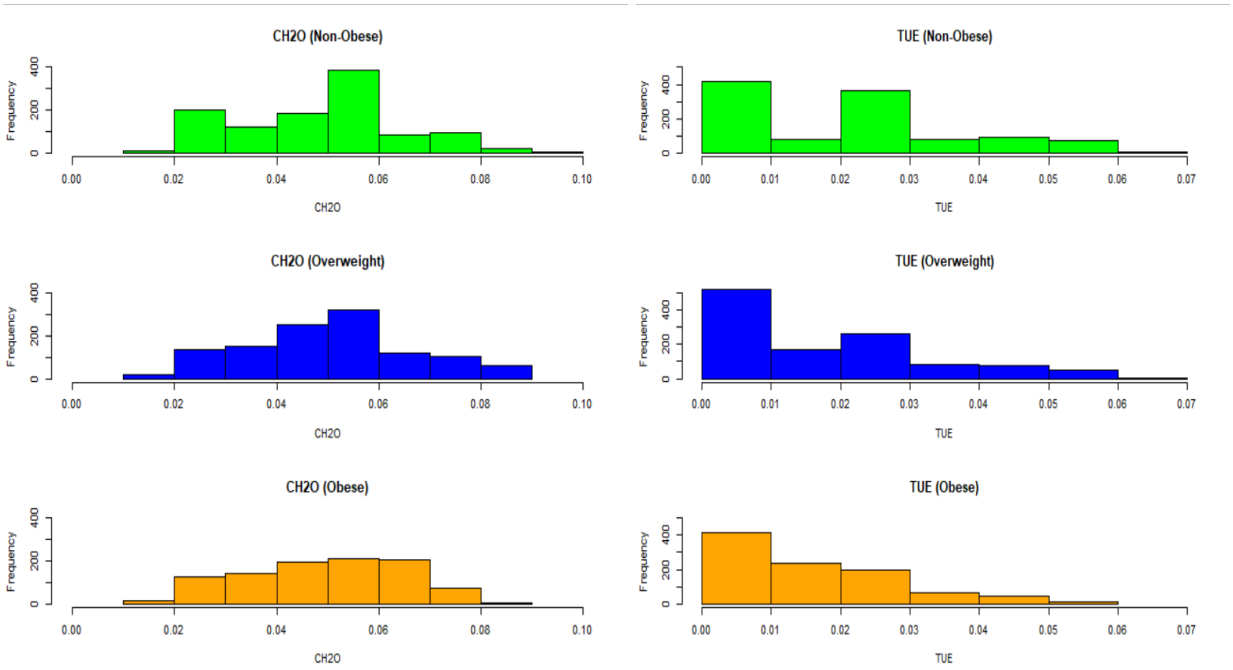

Figure 3. Selected feature CH2O (Consumption of water) and Selected feature TUE (Time using technology devices).

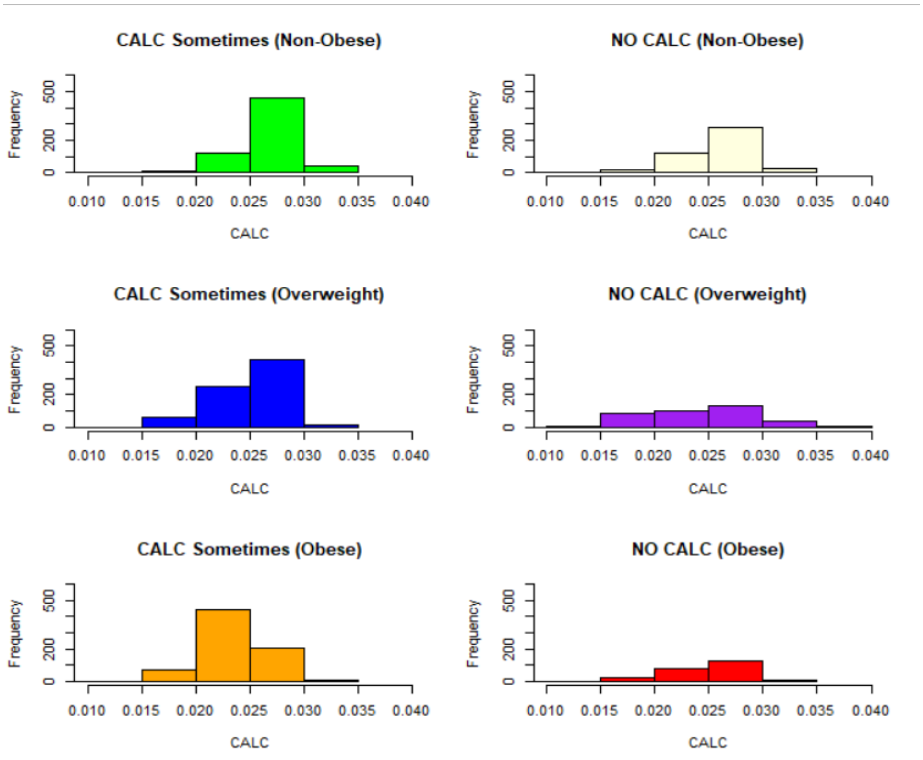

Figure 4. Selected binary features CALC (Consumption of Alcohol) Sometimes and No CALC.

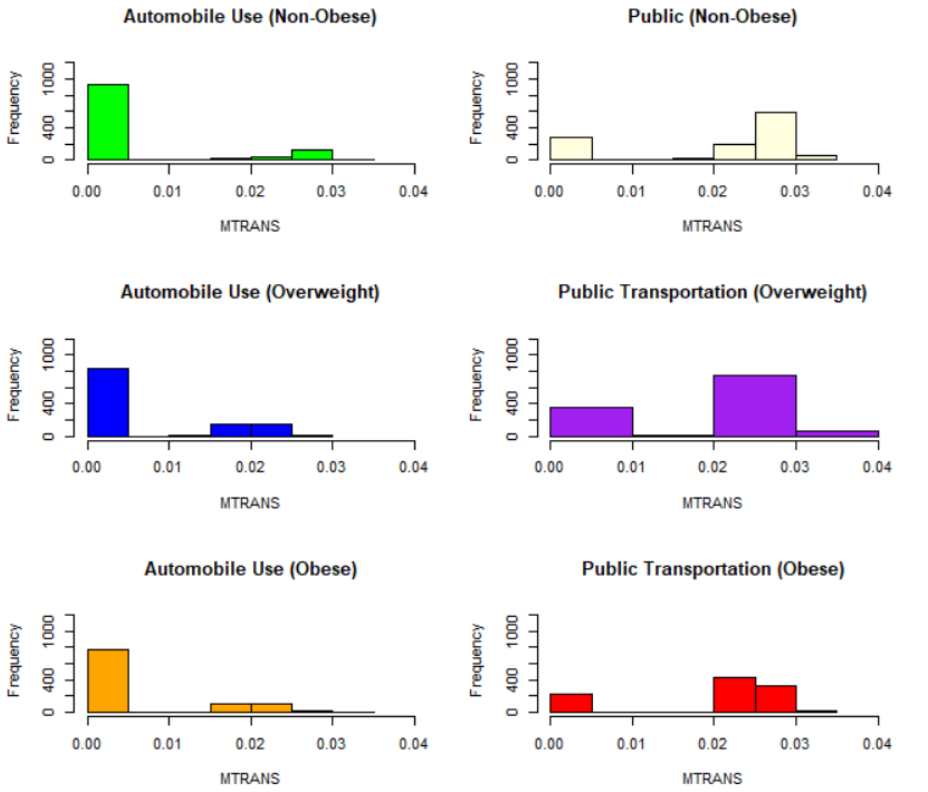

Figure 5: Selected Binary features Automobile use and Public Transportation MTRANS

In order to solidify the understanding of discriminatory features among different classes, the discriminatory power was calculated for each selected feature using t-test. T-test is a statistical test of determining the difference between means of two groups. In our case, there are three groups. Therefore, to compare between two groups, datasets will be split into three groups which includes the first dataset of CL1 and CL2, the second dataset of CL1 and CL3 and the third dataset of CL2 and CL3 groups. As a result, three values of discriminating powers will be obtained to compare between different classes. The t-test was computed using statistical software R. The p-values obtained from each t-test are shown in Table 2. Afterwards, discriminatory power can be obtained by subtracting p-value from 1 (q=1-p). The calculated discriminatory powers of each feature between different groups of classes using this approach is shown in Table 3. As noticeable, most of the p-values are very small and close to 0 resulting in higher q-value indicating higher discriminating power. The reason behind indicating values using scientific notation is that p-values are not exact 0 even though one can interpret it that way. Similarly, for discriminatory power obtained from p-value, the values with scientific notation as shown in table 3 includes values that can be interpreted as 1 even though is not exact 1. The discriminatory power of 1 is the highest discriminatory power.

	Age	NCP	CH2O	TUE	CALC Sometimes	CALC NO	Automobile MTRANS	Public Transp. MTRANS
P-value CL1 & CL2	$2.2x10^{-16}$	$2.2x10^{-16}$	0.07	$3.8x10^{-7}$	0.33	$6.4x10^{-9}$	$3.3x10^{-4}$	$1.2x10^{-4}$
P-value CL1 & CL3	$2.2x10^{-16}$	$2.2x10^{-16}$	0.64	$5.8x10^{-15}$	$3.3x10^{-5}$	$2.2x10^{-16}$	0.7	0.08
P-value CL2 & CL3	0.12	$5.42x10^{-6}$	0.19	0.01	$8.1x10^{-4}$	0.007	$6.7x10^{-4}$	0.02

Table 1. P-values obtained from t-test.

It is evident that Age has the highest discriminatory power between CL1 (Non-Obese) and CL2 (Overweight) as well as between CL1 and CL3 (Obese). Hower, the discriminatory power of Age is lower among CL2 and CL3. The feature NCP (number of main meals) and binary feature (CALC NO) have highest discriminatory power between all groups of classes. Oppositely, the feature CH2O have weak discriminatory power among all classes, especially between the group of Non-Obese and Obese. The binary feature CALC sometimes has weak discriminatory power between CL1 and CL2 but has high discriminatory power among other two groups of classes. Similarly, the binary feature Automobile transportation (MTRANS) has the weakest discriminatory power among CL1 and CL3 but has the highest discriminatory power among the other two classes.

Discriminatory power obtained from p-value of t-test (1-q)	Age	NCP	CH2O	TUE	CALC Sometimes	CALC NO	Automobile MTRANS	Public Transp. MTRANS
Between CL1 and CL2	$9.9x10^{-1}$	$9.9x10^{-1}$	**0.93**	$9.9x10^{-1}$	**0.67**	$9.9x10^{-1}$	$9.9x10^{-1}$	$9.9x10^{-1}$
Between CL1 and CL3	$9.9x10^{-1}$	$9.9x10^{-1}$	**0.36**	$9.9x10^{-1}$	$9.9x10^{-1}$	$9.9x10^{-1}$	**0.3**	**0.92**
Between CL2 and CL3	**0.88**	$9.9x10^{-1}$	**0.81**	**0.99**	$9.9x10^{-1}$	$9.9x10^{-1}$	$9.9x10^{-1}$	**0.98**

Table 2. Discriminatory Power using p-values.

As many of the obtained p-values and discriminatory power from p-values were very close, another approach of calculating discriminatory power was done to verify the obtained results without the help of built-in function in statistical software R. This approach calculates

discriminatory power using formula for difference of means which is similar to the formula of calculating t-statistic in t-test. The means and standard deviations of each selected feature were obtained and below formulas were used to compute discriminating power of each selected feature.

The difference between means of each feature of HIGHmpg and LOWmpg cases were computed using the below formula:

$$s(F)=\sqrt{(sd\text{CLi})^2+(sd\text{CLj})^2}/\text{N}$$

Where i,j are distinct classes and N= number of cases in CLi ≃ number of cases in CLj, resulting from roughly same total number of cases in CLi and CLj.

After then, the discriminatory power for each feature was calculated using the below formula:

$$discr(F)=|mean\text{CLi}(F)-mean\text{CLj}(F)|/s(F)$$

Discriminatory power using difference of means formula	Age	NCP	CH2O	TUE	CALC Sometimes	CALC NO	Automobile MTRANS	Public Transp. MTRANS
Between CL1 & CL2	19.41	14.46	1.82	5.04	0.95	6.21	3.58	3.85
Between CL1 & CL3	21.92	11.45	0.46	7.86	4.11	8.59	0.29	1.71
Between CL2 & CL3	1.52	4.56	1.28	2.67	3.35	2.68	3.40	2.27

Table 3. Discriminatory power using the formula of difference between means.

Table 4 shows the higher the number, the higher the discriminatory power and lower the number, the lower the discriminatory power. Although this is the rough evaluation of discriminatory power, it verifies the results of discriminatory power obtained from p-values because wherever the discriminatory power was high in table 3, the discriminatory power is high for the same feature between groups of classes in table 4. The evaluation of these discriminatory powers helps to avoid placing weaker discriminating features in the same pack of features when doing knn analysis with weighted features in classification analysis.