Estimating Obesity Levels Based on Eating Habits and Physical Condition

Stat 362 Final Report

Emma Wiebe – 20105831, 17ejw8@queensu.ca
Jack Dunn – 20169797, 18JWRD@queensu.ca
Robin Joshi – 20052972, 16rj17@queensu.ca
Yun Kyaw – 20177325, 19ytzk@queensu.ca
Zirong Liu – 20031058, 16zl14@queensu.ca
Zoe Jin – 20013771, 15sj53@queensu.ca

- Emma Wiebe: Data description and Data wrangling
- Jack Dunn: Introduction and Visualizations
- Robin Joshi: Classification – decision tree, KNN, neural network
- Yun Kyaw: Discussion and Conclusion
- Zirong Liu: Random Forest Classification, Hypothesis and Hypotheses Testing
- Zoe Jin: Multinomial Logistic Regression
- All group members have discussed every part of the project together.

1.0 INTRODUCTION

The dataset we are working with will allow us to analyze how different factors contribute to obesity. Our goal is to pinpoint which of the attributes are the biggest contributors to weight gain, and what habits may result in weight loss/ a normal body weight.

Previous studies have found that smoking can potentially lead to obesity [1] and that frequent drinking also results in increased risk of weight gain, and higher chances of obesity [2]. We hope to gain some insight on the topic through some statistical analysis with hypothesis testing, training a regression model, and classification by decision trees.

We hope to find meaningful results that could be applied in the real world. Great things could come out of analysis of data on obesity/overweight people. If there was one attribute found to make a huge difference, certain habits frequently leading to obesity as an example, doctors or whoever is dealing with these problems could begin to look for these habits in their/their patients' lives and correct them before it is too late.

2.0 DATASET DESCRIPTION

The dataset we are using is derived from research regarding obesity levels in Mexico, Peru, and Colombia [3]. This data contains a total of 2111 records 23% of which was collected from individuals taking a web survey while the other 77% was synthetically generated using the tool Weka and a filter SMOTE [3]. The data contains 17 different attributes including obesity level which is our response variable. The remaining 16 attributes can be grouped into three categories: general information, eating habits, and physical condition.

The general attributes that they collected data for include, age of participant, gender, height, weight, and if any of their family members are/were overweight. Age is a numerical value representing how old the participant is, and ranges from 14-61 in this dataset. Gender is a categorical variable with either “Male” or “Female” as the response. Height is a numerical value representing the height of the participant in meters (m). Weight is another numerical variable with values representing the participants’ weight in kilograms (kg). Finally, has a family member ever suffered from obesity is a categorial variable with response options being “Yes” or “No”.

Attributes regarding eating habits include: Frequency of consumption of high caloric food (FAVC), frequency of consumption of vegetables (FCVC), consumption of food between meals (CAEC), number of main meals (NCP), consumption of water daily (CH2O), consumption of alcohol (CALC), and smoking habits. Do you consume high caloric food frequently has categorical responses being “yes” or “no”. Do you usually eat vegetables with your meals has numerical responses with possible answers being “never” = 1, “sometimes” = 2, or “always” = 3. Do you eat food between meals, is a categorical variable with responses being “no”, “sometimes”, “frequently”, or “always”. Next, number of main meals consumed in the day is

represented numerically with responses ranging from 1-4. How much water do you consume daily has numerical responses “less than a liter” = 1, “between 1 and 2L” = 2, or “more than 2L” = 3. How often do you drink alcohol has responses “I don’t drink”, “sometimes”, “frequently”, or “always”. Finally, do you smoke has responses “yes” or “no”. These categories are all classified and measured in accordance with WHO guidelines of healthy diet standards [4].

Attributes regarding the participants physical condition include: Frequency of physical activity (FAF), calorie consumption monitoring (SCC), time using technology devices (TUE), and transportation type used (MTRANS). How often do you participate in physical activity has response categories “I do not” = 0, “1 or 2 days” = 1, “3 or 4 days” = 2, or “4 or 5 days” = 3. Do you monitor calorie consumption has responses “yes” or “no”. How much time do you spend on technology (which includes cellphones, television, videogames, computers, and others) has possible numerical responses of “0-2 hours” = 0, “3-5 hours” = 1, and “more than 5 hours” = 2. Finally, types of transportation include “automobile”, “motorbike”, “bike”, “public transport”, and “walking”.

All the various attributes are being looked at for possible correlation with obesity levels. The researchers who collected this data were attempting to identify a factor or a combination of factors that could accurately predict the obesity level of the participant. Obesity is calculated using the standard BMI calculation:

$$BMI = \frac{weight\,(kg)}{height^2\,(m^2)}$$

Then split into 7 different categories. If the BMI is under 18.5 the participant is considered “underweight”. If it ranges between 18.5 - 24.9 the participant is “normal weight”. If the BMI falls between 25 – 27.4 the participant is “overweight I”. If the BMI falls between 27.5 - 29.9 the participant is “overweight II”. If the BMI is between 30 – 34.9 the participant is “obesity I”. If the BMI is between 35- 39.9 the participant is “obesity II”. Finally, if the BMI is 40 or higher then the participant is “obesity III”.

3.0 METHODS

3.1 Multinomial Logistic Regression

We assume the variable NObesity (Obesity Level) as the response variable, and the other attributes (except weight and height) as covariates. Under this assumption, we look for a suitable approach to describe the relationship between predictors and the response. Before modeling the data, we process our data and convert some of categorical data that in numeric format into factor, and then we partition our data and use around 70% of the data to form our training dataset and the rest to form our testing dataset so that we can use the training dataset to fit a model and evaluate the model performance. Logistic regression is a technique used when the dependent variable is categorical (or nominal). A multinomial logistic regression that is a type of logistic

regression and can be used when the dependent variable is more than two levels [5], in which the log odds of the outcomes are modeled as a linear combination of the predictor variables as form:

$$log \frac{\pi_{ij}}{\pi_{ij}} = \alpha_j + x_i{}' \beta_j \quad (1)$$

where α_i is a constant and β_j is a vector of regression coefficients, for j=1, 2, …, J−1, and X is the model matrix does not include a column of ones [6]. Since NObesity is a categorical variable with 7 levels, we will use the multinominal logistic regression in nnet library in R.

The full fitted model has 14 explanatory variables, but it may be possible that not all of them are useful to model and predict the response. To seek a simpler multinomial log-linear model, our approach is to operate the stepwise regression where the final part of the output tells us which explanatory variables we need to keep. We then run the likelihood ratio test using ANOVA to compare the fit of the model containing all explanatory variables with the final model obtained by the stepwise regression, and check whether we really need the bigger model. We and then use the 2-tailed z test to see if all variables are statistically significant at 95 percent CI. Overall, we will perform model selection procedure as follow:

1. Start with the full model contains all 16 explanatory variables as our fitted model 1.
2. Use step function to find a subset of explanatory variables that are useful to model and predict the response.
3. Select the final part of the output from step function as our fitted model 2. If the model returned same as the full model. STOP. Otherwise, we move to the next step.
4. Set a null hypothesis H_0 (e.g. Model 2 is statistically equivalent to Model 1).
5. Operate ANOVA and find the probability of obtaining test.
6. If null hypothesis H_0 reject, then the new model is different in fitting the data than the model with more terms.
7. If null hypothesis H_0 cannot reject, use the new model instead.

After find the “best” model, we will use it and our testing dataset to feed prediction and then evaluate model performance. In this case, the response is multi-class and not continuous, we can use table and/or CrossTable function introduced in class to obtain accuracy and confusion matrix, where the confusion matrix tells whether the model is good at identifying each of the class. However, unlike binary output, there are no positive or negative classes here, so we have to find TP, TN, FP and FN for each individual class. We will also use confusionMatrix function to find sensitivity and specificity for each class and use 10-fold cross-validation to estimate the future performance. The potential problem with this approach for this dataset could be difficultly interpreting categorical data, and outliers can temper the accuracy (see Data Visualization).

3.2 Classification

The goal of performing classification tasks on this dataset is to see if we can predict BMI levels based on the eating/living habits of people. The dataset provides seven groups of BMI levels as we discussed in the dataset description section. Ideally our findings will show how eating/living habits can influence BMI levels and can be used by people who want information on this topic. Note that we do not use weight and height as features since the response (BMI) is calculated using these quantities.

The first method we use is k-nearest neighbours. The idea behind this method is to use the k-nearest data points to decide what group the new data point belongs to. For example, if k = 3, and two out of three points belong to group A, then the new data point will be put into group A according to this method. The Euclidean distance is considered between two points. We also normalize the data using the min-max normalization method before implementing a 3-nearest neighbours' function from the class library in R. We used dummy indices for categorical features in the dataset.

The second method of classification we use is the classification tree. This method uses a tree-like structure to classify the data. At each node in the tree one branch is followed based on some splitting rule. We use the C5.0 function from the C50 library in R. This uses the difference between the entropy before the split and the entropy after the split to determine the splitting to create a branch.

The final method we use is the neural network. This method builds a model for the input and output signals mimicking how a human brain works. This method can be used for our classification model building. We use the neuralnet function from the neuralnet library in R. Like the KNN technique we first normalize the data using the min-max normalization.

3.3 Random Forest Classification

In addition to decision tree, we also used the random forest classification method. As a supervised learning algorithm, random forest uses the "bagging" method to ensemble several decision trees to get a more accurate and stable prediction [7]. In the training process, each individual decision tree that the random forest generated gives out a class prediction and the most accurate prediction becomes the model's prediction [8]. Due to its applicable for both classification and regression and the simplicity and general turn out good results, we decided to use this method. The main limitation for random forest is that a lot of trees can make the algorithm too slow, however, it is not applicable to our case. As previous sections, we will use table function and mean function to tell us whether this model is a good fit for our case. We will also use 10-fold cross-validation to estimate the future performance.

3.4 Hypothesis and Hypotheses Testing

Before applying machine learning methods for regression and classification to the dataset, we made a list of topics that we were interested and performed related testing to better understand the dataset. The topics were asked based on our curiosity such as which eating habits and physical conditions would potential factors that affect the weight gain/loss. We used "t.test", "prop.test", and "ks.test" functions in R to answer the questions we listed in this section.

The topics we discussed are listed in the following:

1. Average age for Obesity type I group, Obesity type II group, and Obesity type III group of people?
2. Average height for three groups of people who are categorized as Obesity and people who have the normal weight?
3. Average weight for three groups of people who are categorized as Obesity?
4. Whether the average consumption of water daily is equal to two for three groups of people who are categorized as Obesity?
5. What is the 99% confidence interval for average number of main meals of Obesity group (included three types together)?
6. Are the means of the weights in the smoking group and non-smoking group different?
7. For question 6, is it same for people who are categorized in Obesity?
8. Are the means of the weights for overweight level I people and overweight level II people who have family history with overweight different?
9. If three types of obesity people have equal portion of people who use Automobile as transportation.
10. If the difference between the two proportions is statistically significant for smokers in type I Obesity group and type II Obesity group.

4.0 RESULTS

4.1 Multinomial Logistic Regression

In the results that follow, since no extra variables are dropped from step function, so we can say these variables are useful for predicting the response. The model summary returns a block of coefficients and a block of standard errors. Each block has a row of values corresponding to the model equations. Focusing on parameter estimates (see Table 1 and Table 2), we can see the relationship between predicators and the response by comparing two classes from the response. Here, we treat inefficient weight as the baseline and compare everything else against it. Therefore, we can have estimated 6 model equations as follow:

1. a model equation for normal weight relative to insufficient weight,

2. a model equation for overweight level I relative to insufficient weight,

3. a model equation for overweight level II relative to insufficient weight,

4. a model equation for obesity type I relative to insufficient weight,

5. a model equation for obesity type II relative to insufficient weight, and

6. a model equation for obesity type III relative to insufficient weight.

Since the interpretations of above model equations are more or less the same, we will use the model equation for normal weight relative to insufficient weight as an example. First, we will focus on the coefficients, and interpret the coefficient as follow:

- The coefficient of GenderMale is the multinomial logit estimate comparing females to males for normal weight relative to insufficient weight given the other variables in the model are held constant. The log-odds of being in normal weight level vs insufficient weight level will increase by 0.7 if switching females to males. In other words, males are more likely than females to being in normal weight level relative to insufficient weight level.
- If a one-unit increase in age, the log-odds of being in normal weight level to insufficient weight level would be expected to increase by 0.2 unit while holding all other variables in the model constant.
- Those who have family history with overweight are more likely to being in normal weight level relative to insufficient weight level.
- Those who eat high caloric food frequently are less likely to being in normal weight level relative to insufficient weight level.
- Those who sometimes eat vegetables are more likely than those who never eat vegetables to being in normal weight level relative to insufficient weight level; Those who always eat vegetables are less likely than those who never eat vegetables to being in normal weight level relative to insufficient weight level.
- If a one-unit increase in number of main meals, the log-odds of being in normal weight level to insufficient weight level would be expected to decrease by 0.16 unit while holding all other variables in the model constant.
- Those who always consumption of food between meals are more likely than those who don't consumption of food between meals to being in normal weight level relative to insufficient weight level, while those whose sometimes or frequently consumption of food is less likely to being in normal weight level relative to insufficient weight level.
- Those who smokes are more likely than those who don’t to being in normal weight level relative to insufficient weight level.
- Those who drink between 1 and 2 L or more likely than 2 L water are less than those who drink less than a liter to being in normal weight level relative to insufficient weight level.
- Those who monitor calories consumption are more likely to being in normal weight level relative to insufficient weight level.

- Those who have physical activity for 1 or 2 days, or 4 or 5 days are more likely than those who have no physical activity to being in normal weight level relative to insufficient weight level, while it is interesting to see that those who have physical activity for 2 or 4 days are less likely than those who have no physical activity to being in normal weight level relative to insufficient weight level.
- Those who spend 3 to 5 hours or more than 5 hours on technology are less likely than those who spend 0 to 2 hours to being in normal weight level relative to insufficient weight level.
- Those who sometimes or frequently drink alcohol are more likely than those who never drink to being in normal weight level relative to insufficient weight level, while people always drink alcohol are less likely than those who never drink to being in normal weight level relative to insufficient weight level.
- Those who don't use automobile are more likely than those who use to being in normal weight level relative to insufficient weight level.

Secondly, we will consider exponentiation of the coefficients which are the odds (or relative risk) ratios for the explanatory variables. For example, for males relative to females, the odds ratio for being in normal weight level relative to insufficient weight level would be expected to increase by a factor of 2 given the other variables in the model are held constant. In other words, males are more likely than females to being normal weight relative to being insufficient weight. After checked all exponentiation of the coefficients, we find that the interpretations of exponentiation of the coefficients agree with the finding in the interpretations of coefficients.

Based on the result obtained from our training dataset, we find the following points:

- Males are more likely being in levels from normal to Obesity Type II.
- If people have family history with overweight are more likely in higher obesity levels.
- People who smoke are more likely being in normal and Obesity Type II levels.
- People who sometimes drink alcohol are more likely to being in the Obesity Type III, those who frequently drink alcohol are more likely being in the Overweight Level I, and those who always are more likely being in the normal weight. This is said that a certain amount of alcohol intake may let people in a higher obesity level while more alcohol intake can let people in a lower obesity.

The overall accuracy and kappa are around 0.619 and 0.55 respectively (see Table 4), which indicates we have a moderate agreement. From Table 5, our specificity values are close to 1, which means that there few false positive results in each class. However, the sensitivity values of normal weight and Overweight level I and Level II are small, and this means there are a few false negative results, it is not good at identifying normal weight and overweight classes. The

mean accuracy from 10-fold cross-validation is 0.616 which indicates our future performance is roughly good but can be improved.

Table 1: Parameter Estimates Part I

y.level	term	estimate	p.value	Exp(estimate)
Normal_Weight	(Intercept)	-3.771	0.024	2.302745e-02
	GenderMale	0.715	0.017	2.044893e+00
	Age	0.230	0.000	1.258183e+00
	family_history_with_overweightyes	0.671	0.020	1.955247e+00
	FAVCyes	-0.618	0.077	5.390874e-01
	FCVC2	0.433	0.439	1.542133e+00
	FCVC3	-0.544	0.339	5.804282e-01
	NCP	-0.160	0.346	8.523101e-01
	CAECSometimes	-0.910	0.351	4.026644e-01
	CAECFrequently	-1.824	0.065	1.613608e-01
	CAECAlways	1.819	0.156	6.168548e+00
	SMOKEyes	3.488	0.012	3.272823e+01
	CH2O2	-0.258	0.403	7.723018e-01
	CH2O3	-1.251	0.005	2.861140e-01
	SCCyes	0.309	0.510	1.362048e+00
	FAF1	0.557	0.090	1.744757e+00
	FAF2	-1.012	0.005	3.634501e-01
	FAF3	0.978	0.077	2.659914e+00
	TUE1	0.028	0.925	1.028322e+00
	TUE2	-1.398	0.001	2.469833e-01
	CALCSometimes	-0.130	0.632	8.780783e-01
	CALCFrequently	2.982	0.025	1.973126e+01
	CALCAlways	5.225	0.000	1.859221e+02
	MTRANSBike	7.866	0.000	2.606716e+03
	MTRANSMotorbike	11.721	0.000	1.231403e+05
	MTRANSPublic_Transportation	1.266	0.003	3.547671e+00
	MTRANSWalking	3.108	0.000	2.236876e+01
Overweight_Level_I	(Intercept)	-6.035	0.001	2.392848e-03
	GenderMale	0.304	0.357	1.355563e+00
	Age	0.349	0.000	1.417932e+00
	family_history_with_overweightyes	1.672	0.000	5.322271e+00
	FAVCyes	0.521	0.239	1.684246e+00
	FCVC2	0.159	0.789	1.172704e+00
	FCVC3	-0.949	0.123	3.871450e-01
	NCP	-0.470	0.011	6.251863e-01

	CAECSometimes	-1.419	0.137	2.418830e-01
	CAECFrequently	-5.195	0.000	5.546471e-03
	CAECAlways	-1.860	0.206	1.556809e-01
	SMOKEyes	1.750	0.299	5.754821e+00
	CH2O2	-0.548	0.120	5.783445e-01
	CH2O3	-0.795	0.091	4.516852e-01
	SCCyes	1.545	0.001	4.690204e+00
	FAF1	0.363	0.315	1.438319e+00
	FAF2	-0.945	0.018	3.888646e-01
	FAF3	0.692	0.270	1.998145e+00
	TUE1	-0.550	0.095	5.770805e-01
	TUE2	-1.583	0.000	2.053219e-01
	CALCSometimes	0.737	0.020	2.088983e+00
	CALCFrequently	4.207	0.003	6.715166e+01
	CALCAlways	-0.430	0.000	6.506935e-01
	MTRANSBike	8.475	0.000	4.794189e+03
	MTRANSMotorbike	-5.035	0.000	6.504621e-03
	MTRANSPublic_Transportation	1.420	0.002	4.139136e+00
	MTRANSWalking	1.578	0.089	4.842939e+00
Overweight_Level_II	(Intercept)	-14.833	0.000	3.613338e-07
	GenderMale	1.805	0.000	6.077384e+00
	Age	0.550	0.000	1.732942e+00
	family_history_with_overweightyes	3.649	0.000	3.843466e+01
	FAVCyes	-1.653	0.000	1.915311e-01
	FCVC2	0.103	0.890	1.108433e+00
	FCVC3	-1.151	0.133	3.161816e-01
	NCP	-0.678	0.001	5.074342e-01
	CAECSometimes	1.546	0.293	4.691473e+00
	CAECFrequently	-1.927	0.208	1.456442e-01
	CAECAlways	1.433	0.445	4.189560e+00
	SMOKEyes	2.847	0.084	1.723026e+01
	CH2O2	0.062	0.876	1.064285e+00
	CH2O3	-0.500	0.329	6.068042e-01
	SCCyes	-0.755	0.424	4.698895e-01
	FAF1	0.774	0.045	2.167470e+00
	FAF2	-0.951	0.026	3.864345e-01
	FAF3	0.520	0.443	1.682667e+00
	TUE1	0.737	0.036	2.088958e+00
	TUE2	-1.097	0.021	3.337631e-01
	CALCSometimes	-0.897	0.006	4.077782e-01
	CALCFrequently	3.039	0.033	2.087712e+01
	CALCAlways	-2.309	0.000	9.938229e-02

	MTRANSBike	-9.880	0.000	5.120491e-05
	MTRANSMotorbike	-7.210	0.000	7.389584e-04
	MTRANSPublic_Transportation	2.595	0.000	1.339292e+01
	MTRANSWalking	2.244	0.019	9.428583e+00

Table 2: Parameter Estimates Part II

y.level	term	estimate	p.value	Exp(estimate)
Obesity_Type_I	(Intercept)	-12.553	0.000	3.533612e-06
	GenderMale	0.969	0.005	2.635291e+00
	Age	0.445	0.000	1.559836e+00
	family_history_with_overweightyes	4.399	0.000	8.134395e+01
	FAVCyes	1.456	0.012	4.286811e+00
	FCVC2	-0.408	0.523	6.651480e-01
	FCVC3	-2.017	0.002	1.331130e-01
	NCP	-0.826	0.000	4.378585e-01
	CAECSometimes	1.521	0.298	4.576001e+00
	CAECFrequently	-3.036	0.053	4.803261e-02
	CAECAlways	1.555	0.384	4.734628e+00
	SMOKEyes	3.393	0.040	2.975109e+01
	CH2O2	-0.670	0.078	5.114619e-01
	CH2O3	-0.093	0.847	9.112967e-01
	SCCyes	-28.064	0.000	6.488234e-13
	FAF1	0.344	0.365	1.410405e+00
	FAF2	-0.840	0.042	4.318724e-01
	FAF3	0.341	0.604	1.406803e+00
	TUE1	-0.088	0.797	9.157405e-01
	TUE2	-1.239	0.005	2.896820e-01
	CALCSometimes	-1.003	0.002	3.668847e-01
	CALCFrequently	3.009	0.036	2.027498e+01
	CALCAlways	-1.058	0.000	3.470456e-01
	MTRANSBike	-6.324	0.000	1.793321e-03
	MTRANSMotorbike	11.973	0.000	1.584534e+05
	MTRANSPublic_Transportation	1.962	0.000	7.112439e+00
	MTRANSWalking	0.657	0.618	1.929335e+00
Obesity_Type_II	(Intercept)	-29.353	0.000	1.786476e-13
	GenderMale	10.472	0.039	3.532786e+04
	Age	0.643	0.000	1.902477e+00
	family_history_with_overweightyes	7.377	0.000	1.599331e+03
	FAVCyes	-0.142	0.833	8.674357e-01
	FCVC2	-0.918	0.204	3.991785e-01
	FCVC3	-0.774	0.302	4.611658e-01
	NCP	-0.015	0.954	9.854256e-01

	CAECSometimes	-0.406	0.841	6.665731e-01
	CAECFrequently	-6.401	0.004	1.659555e-03
	CAECAlways	0.274	0.910	1.314765e+00
	SMOKEyes	3.487	0.042	3.269364e+01
	CH2O2	-0.711	0.110	4.911456e-01
	CH2O3	-1.876	0.001	1.531580e-01
	SCCyes	0.335	0.844	1.398575e+00
	FAF1	0.595	0.165	1.812702e+00
	FAF2	-1.067	0.024	3.440255e-01
	FAF3	-17.221	0.000	3.317454e-08
	TUE1	-0.075	0.842	9.278734e-01
	TUE2	-2.368	0.000	9.362768e-02
	CALCSometimes	-0.331	0.401	7.178884e-01
	CALCFrequently	1.689	0.327	5.412019e+00
	CALCAlways	0.595	0.000	1.813512e+00
	MTRANSBike	7.234	0.001	1.385116e+03
	MTRANSMotorbike	-3.828	0.000	2.174674e-02
	MTRANSPublic_Transportation	3.225	0.000	2.514647e+01
	MTRANSWalking	0.727	0.622	2.068263e+00
Obesity_Type_III	(Intercept)	-50.996	0.000	7.125886e-23
	GenderMale	-7.412	0.000	6.039311e-04
	Age	0.404	0.017	1.498027e+00
	family_history_with_overweightyes	9.630	0.000	1.520748e+04
	FAVCyes	5.075	0.000	1.599547e+02
	FCVC2	2.341	0.545	1.039176e+01
	FCVC3	10.749	0.001	4.658600e+04
	NCP	3.282	0.000	2.663013e+01
	CAECSometimes	1.734	0.813	5.662659e+00
	CAECFrequently	-8.239	0.289	2.640237e-04
	CAECAlways	-7.201	0.424	7.461482e-04
	SMOKEyes	2.933	0.537	1.877695e+01
	CH2O2	-0.052	0.959	9.490630e-01
	CH2O3	0.944	0.441	2.570617e+00
	SCCyes	-5.198	0.263	5.528301e-03
	FAF1	-0.682	0.548	5.056789e-01
	FAF2	-0.581	0.670	5.594921e-01
	FAF3	-0.534	0.895	5.864422e-01
	TUE1	-0.208	0.834	8.124162e-01
	TUE2	-2.650	0.459	7.066904e-02
	CALCSometimes	4.477	0.001	8.799552e+01
	CALCFrequently	-1.665	0.909	1.892751e-01
	CALCAlways	0.838	0.000	2.312584e+00
	MTRANSBike	7.911	0.000	2.727845e+03

	MTRANSMotorbike	5.764	0.000	3.186344e+02
	MTRANSPublic_Transportation	6.721	0.004	8.298071e+02
	MTRANSWalking	-0.645	0.901	5.248571e-01

Table 3: Predicted NObeyesdad Table

	Insufficient_Weight	Normal_Weight	Overweight_Level_I	Overweight_Level_II	Obesity_Type_I	Obesity_Type_II	Obesity_Type_III
Insufficient_Weight	70	9	8	1	4	4	0
Normal_Weight	23	31	15	13	7	1	0
Overweight_Level_I	6	9	33	8	22	11	1
Overweight_Level_II	4	5	7	21	17	15	1
Obesity_Type_I	2	3	5	14	62	10	4
Obesity_Type_II	0	0	0	2	9	76	0
Obesity_Type_III	1	0	0	0	0	0	99

Table 4: Overall Statistics output from confusionMatrix function

	Overall_Statistics
Accuracy	0.6192733
Kappa	0.5543998
AccuracyLower	0.5801716
AccuracyUpper	0.6572570
AccuracyNull	0.1579779
AccuracyPValue	0.0000000
McnemarPValue	NaN

Table 5: Statistics by Class output from confusionMatrix function

	Sensitivity	Specificity
Class: Insufficient_Weight	0.7291667	0.9329609
Class: Normal_Weight	0.3444444	0.9521179
Class: Overweight_Level_I	0.3666667	0.9355433
Class: Overweight_Level_II	0.3000000	0.9325044
Class: Obesity_Type_I	0.6200000	0.8893058
Class: Obesity_Type_II	0.8735632	0.9249084
Class: Obesity_Type_III	0.9900000	0.9887430

4.2 Data Visualization

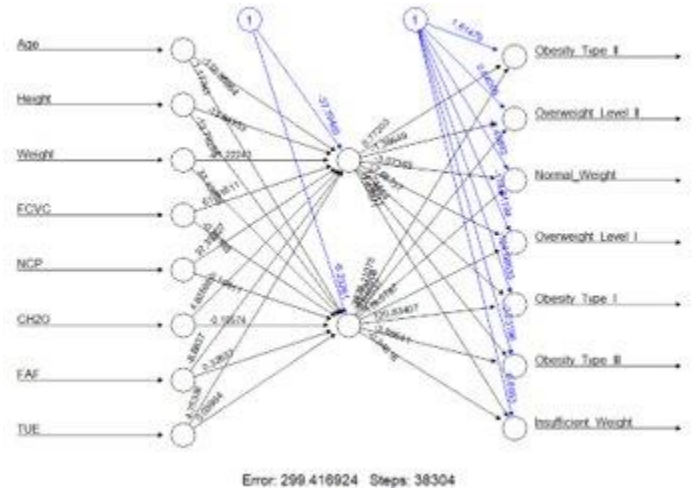

Figure 1. Neural Network trained for classification. One layer with two neurons was used for this figure.

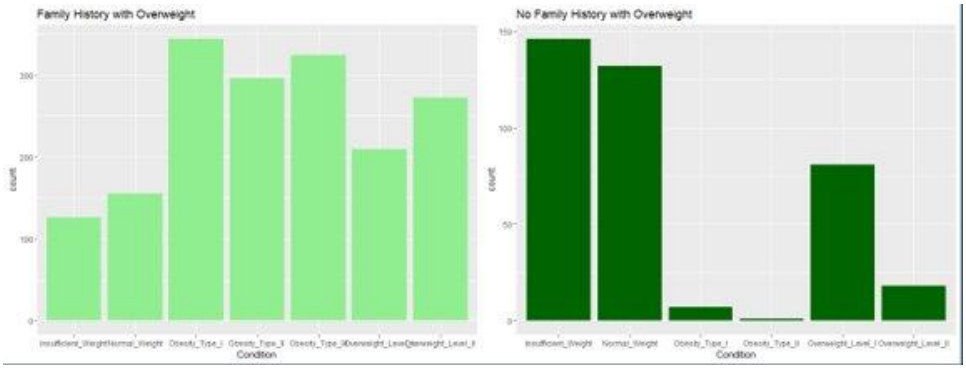

Figure 2. Comparison of those with family history of being overweight vs. Those who don't.

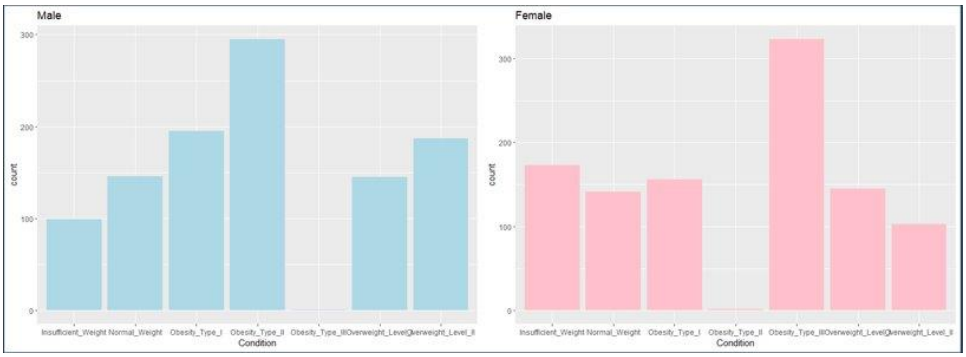

Figure 3. Comparison of cases between Male and Female.

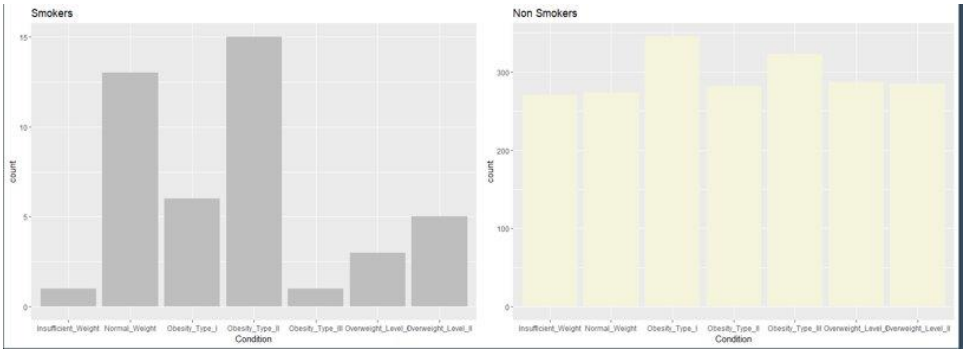

Figure 4. Smokers vs. Non-Smokers

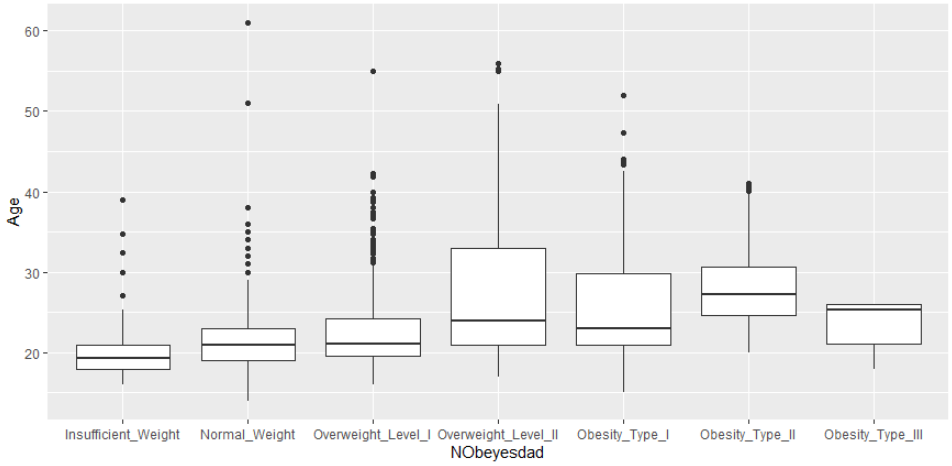

Figure 5. Boxplot of age for obesity level.

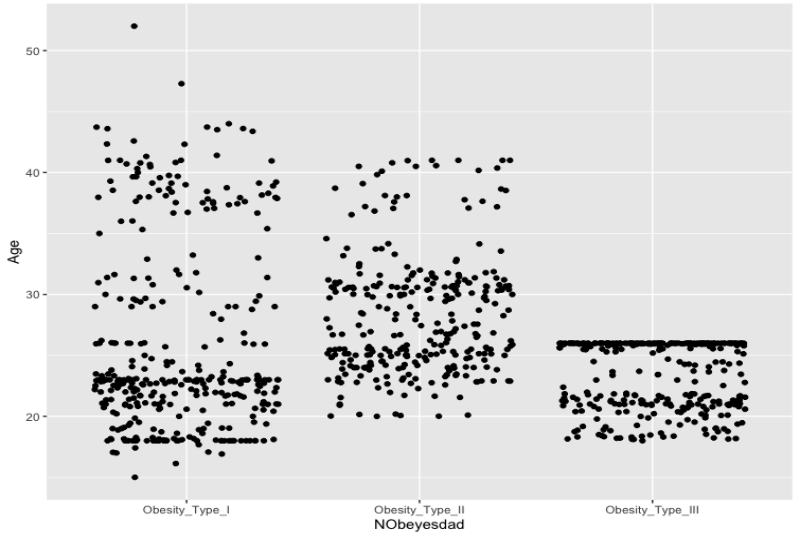

Figure 6. Jitter plot of age of Obesity Type I, Obesity Type II, Obesity type III.

4.3 Classification

Using the 3-nearest neighbours' method we find an accuracy of about 0.77. The confusion matrix is shown in Table 6.

Table 6: Confusion matrix from 3-NN

	Insufficient Weight	Normal Weight	Overweight 1	Overweight 2	Obesity Type 1	Obesity Type 2	Obesity Type 3
Insufficient Weight	80	10	2	3	1	0	0
Normal Weight	7	43	8	6	6	1	1
Overweight 1	5	11	61	1	4	1	0
Overweight 2	1	15	7	51	9	2	0
Obesity Type 1	1	8	3	5	70	2	0
Obesity Type 2	2	2	9	4	6	81	0
Obesity Type 3	0	1	0	0	4	0	99

Using the C5.0 classification tree we find an accuracy of about 0.78. The confusion matrix is shown in Table 7.

Table 7: Confusion matrix from classification tree

	Insufficient Weight	Normal Weight	Overweight 1	Overweight 2	Obesity Type 1	Obesity Type 2	Obesity Type 3
Insufficient Weight	78	11	4	0	1	1	0
Normal Weight	10	56	13	5	7	2	1
Overweight 1	5	4	56	3	5	1	0
Overweight 2	2	11	8	52	11	4	0
Obesity Type 1	1	8	7	6	73	1	0
Obesity Type 2	0	0	2	4	1	78	0
Obesity Type 3	0	0	0	0	2	0	99

The accuracy of our neural network using two hidden layers with four neurons each is 0.65. The confusion matrix is shown in Table 8. This method took 226 seconds to run. It is likely that the accuracy could be improved by increasing the depth of the neural network.

Table 8: Confusion matrix from neural network

	Insufficient Weight	Normal Weight	Overweight 1	Overweight 2	Obesity Type 1	Obesity Type 2	Obesity Type 3
Insufficient Weight	76	23	7	4	2	0	0
Normal Weight	12	45	25	8	7	0	1
Overweight 1	6	7	41	7	22	2	1
Overweight 2	1	7	4	28	15	2	0
Obesity Type 1	1	4	12	5	42	1	0
Obesity Type 2	0	4	1	16	8	82	0
Obesity Type 3	0	0	0	2	4	0	98

Overall, we see that eating/living habits can predict your BMI to a good accuracy (using KNN and classification trees). The accuracy could be improved upon if a deeper neural network was trained on a high-performance computing system.

4.4 Random Forest Classification

By using the mean function, we could see that the accuracy for random forest model is 0.8056872, which is higher than decision tree's accuracy – 0.78. The mean accuracy from 10-fold cross validation is 0.9075928, which indicates the future performance for this model is promising.

Table 9: Confusion matrix from Random Forest

Prediction	Insufficient_Weight	Normal_Weight	Overweight_Level_I
Insufficient_Weight	81	10	1
Normal_Weight	9	55	9
Overweight_Level_I	3	8	68
Overweight_Level_II	2	6	4
Obesity_Type_I	1	9	7
Obesity_Type_II	0	2	1
Obesity_Type_III	0	0	0

Prediction	Overweight_Level_II	Obesity_Type_I	Obesity_Type_II	Obesity_Type_III
Insufficient_Weight	2	0	0	0
Normal_Weight	6	4	1	1
Overweight_Level_I	5	7	0	0
Overweight_Level_II	45	3	0	0
Obesity_Type_I	6	79	1	0
Obesity_Type_II	6	3	85	0
Obesity_Type_III	0	4	0	99

4.5 Hypothesis and Hypotheses Testing

1. The average age for Obesity type I group is 26, Obesity type II group is 28, and Obesity type III is 24. Comparing with the first two types of Obesity groups, Obesity type III seems is the group that have the youngest average age. For this dataset, the average testing sample's age range is from 14 to 61 years old. Also, by statistics provided by the Centers for Disease Control and Prevention, "the prevalence of obesity was 40.0% among adults aged 20 to 39 years, 44.8% among adults aged 40 to 59 years, and 42.8% among adults aged 60 and older [9]." Our dataset doesn't seem to match the general situation. As shown in figure [6] in section 4.2 data visualization, our dataset doesn't have any observation for people who are older than 26 and categorized in Obesity type III.

2. The average height for three groups of people who are categorized as Obesity are 1.69, 1.77, and 1.69. Meanwhile, the average height for normal weight group is 1.68. By comparison, it seems that the average height for the obesity type II people is far higher than other groups, which could be caused by multiple factors, since it doesn't make much sense.

3. The average weight for three groups of people who are categorized as Obesity are 92.87, 115.31, and 120.94.

4. The average consumption of water daily for three groups of Obesity people are 2.11, 1.88, and 2.21. Which are all not far away from 2. Hence, it may indicate that average consumption of water daily did not affect much for defining types of Obesity.

5. The 99% confidence interval of Obesity group (included three types of Obesity groups together) is 2.665547 and 2.768026.

6. The means of weight for the smoking group and non-smoking group are 91.21 and 86.49. The smoking group tends to be heavier than non-smoking group.

7. The means of weight for the smoking Obesity group and non-smoking Obesity group are 114.50 and 108.96. The smoking group still tends to be heavier than non-smoking group.

8. The means for overweight level I group with family history with overweight and without family history with overweight are 74.82933 and 73.09394. The means are 82.60761 and 80.99613 for overweight level II group. Hence, whether family history with overweight or not does not seem like a huge factor that affect the weight of overweight group.

9. Three types of obesity people have the following portion of people who use Automobile as transportation: 0.31339031, 0.31986532, and 0.00308642. The type III obesity group seems have a far smaller proportion than other two types of obesity group. This could cause by many reasons.

10. We wanted to know whether between the two proportions (smokers in type I obesity group and type II obesity group) is statistically significant. By performing ks.test, we know that the p-value < 2.2e-16, which is smaller than 0.05. Therefore, we conclude that the difference of the two distributions is statistically significant.

5.0 DISCUSSION AND CONCLUSION

The selected dataset from Universidad de la Costa includes data that was used for estimating obesity levels in individuals from Mexico, Peru, and Columbia given their eating habits and physical condition. This dataset consists of 17 attributes and 2111 records. This required minimal data wrangling, only involving the reformatting of data for the different statistical analyses.

The first statistical analysis performed on the dataset was a classification through a k-nearest neighbours. This resulted in an accuracy of around 0.77. The second statistical analysis, another classification through a classification tree, was more successful and had a higher accuracy of 0.78. The third statistical analysis performed on the dataset was a random forest

classification and had an even higher accuracy of 0.81. The fourth statistical analysis performed on the dataset was a multinomial logistic regression and this analysis had an overall accuracy of 0.619. And the regression model concluded four things. Firstly, males are more likely to have a BMI in the normal to Obesity Type II range, secondly, people whose families have a history of being overweight are more likely in the higher obesity, thirdly, people who smoke are more likely to be in normal and Obesity Type II levels, and lastly, a certain amount of alcohol intake may let people in a higher obesity level or in a lower obesity.

There were ten hypotheses that were tested. Here will only conclude first three. Firstly, the average age of groups of people with obesity type I, type II, and type III was found to be 26, 28, and 24, respectively. Secondly, the mean weights of smoking and non-smoking groups differed, with smoking groups weighing an average 114.5 kg, and non-smoking groups weighing an average 109kg. Lastly, the groups of people with different types of obesity differed in the proportion of people who selected “automobile” as their main form of transportation, with the proportion of those with type III being relatively lower than those with type I and type II.

One limitation to this dataset is that 77 percent of the data is synthesized, thus it is not an ideal dataset for application. As well there are some limitations in relation to different statistical analyses. In the C50 classification tree, a limitation that affected our dataset is the possible over or underfitting of trees, and the difficulty in interpretation of large trees [10]. A limitation from the k- nearest neighbours is that it does not produce a model [11]. There was also an ordinal logistic regression performed on the dataset, though this was unsuccessful with a very low accuracy as 0.427 and low mean accuracy from 10-fold cross validation as 0.387. This may be due to the numerous assumptions this algorithm involved—firstly, the dependent variable should be in order. Secondly, the independent variables are assumed to be continuous, categorical, or ordinal. Thirdly, there is assumed to be no multi-collinearity, and lastly, there is the assumption that the relationships of the different outcome groups have proportional odds [12]. In the random forest classification, some limitations that affected out analysis includes its requirement of computational power, and the inability to know the significance of each variable [13]. Lastly, for the multinomial logistic regression, there is difficulty in the interpretation of categorical data, and there are some outliers, and it can affect the overall accuracy.

6.0 REFERENCES

[1] Arnaud Chiolero et al. 2008

[2] Gregory Traversy and Jean-Philippe Chaput, 2015

[3] 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.

[4] Healthy diet. (n.d.). Retrieved April 15, 2021, from https://www.who.int/news-room/fact-sheets/detail/healthy-diet

[5] Logistic regression. (2021, March 23). Retrieved from wikipedia.org/wiki/Logistic_regression #:~:text=Logistic regression is a statistical,a form of binary regression).

[6] Rodríguez, G. (n.d.). GR's Website. Retrieved April 11, 2021, from data.princeton.edu/wws509/notes/c6s2

[7] Niklas Donges (June 16, 2019) A COMPLETE GUIDE TO THE RANDOM FOREST ALGORITHM.

Retrieved from https://builtin.com/data-science/random-forest-algorithm

[8] Tony Yiu (Jun 12, 2019) Understanding Random Forest. Retrieved from https://towardsdatascience.com/understanding-random-forest-58381e0602d2

[9] U.S. Department of Health & Human Services (February 11, 2021) Adult Obesity Facts

Retrieved from https://www.cdc.gov/obesity/data/adult.html#:~:text=The%20prevalence%20of%20obesity%20was%2040.0%25%20among%20adults%20aged%2020,adults%20aged%2060%20and%20older.

[10] Ling, B 2021, *Chapter 11 Classification Trees*, lecture notes, R for Data Science STAT 362, Queen's University.

[11] Ling, B 2021, *Chapter 10 k-nearest neighbours*, lecture notes, R for Data Science STAT 362, Queen's University.

[12] Lee, E. (2019, May 29). Ordinal logistic regression on world happiness report. Retrieved from https://medium.com/evangelinelee/ordinal-logistic-regression-on-world-happiness-report-221372709095

[13] N.a (2020, March 11). Random Forest Algorithm- An Overview. Retrieved from mygreatlear ning.com/blog/random-forest-algorithm/