**The Leading Cause of Obesity**

Sawyer Mason, [smason@bellarmine.edu](mailto:smason@bellarmine.edu)

John Cason, [jcason@bellarmine.edu](mailto:jcason@bellarmine.edu)

**ABSTRACT**

The link between weight levels and various other factors has become a further topic of interest for study. Through the course of this analysis, the various factors are explored as independent variables with one of the six personal weight levels becoming the dependent variable. By the method of One-Hot encoding, the independent categorical variables were altered to binary variables. At the same time, Label encoding was used to transform the dependent variable, Obesity level, to 0 through 6. Finally, a logistic regression model was used, trained, and verified in an effort to create a machine learning model to predict weight level. This experiment was repeated containing different parameters and re-checked for accuracy.

1. **INTRODUCTION**

The Data set we chose to use for this project was Obesity Levels. This dataset was found on Kaggle and includes data for the estimation of obesity levels in individuals from the countries of Mexico, Peru and Colombia. The estimations for this data were based on the subject's eating habits and physical condition. The data contains 17 columns and 2111 entries. This data set is very usable and has zero null entries. The one drawback to this data is it is specific to Latino genetics so when using our data for other areas in the world the results could vary. Along with this, most of the data was generated synthetically using the Weka tool and the SMOTE filter. Only 23% of the data was collected directly from users through a web platform. Through this dataset we will try to find how physical activities relate to obesity and create a logistical regression model to predict if a person is obese based on its variable entries.

1. **BACKGROUND**

Obesity is a disease involving having so much body fat that it leads to other health risks. Obesity is a major problem in the United States today. It is estimated that 69% of the US population is overweight and 36% of that are obese. Even though this data was collected from Mexico, Peru, and Colombia the hope is that this data will still reveal effective results for the US. By finding the leading factors to Obesity America can start to set out programs to become a healthier nation.

1. **EXPLORATORY ANALYSIS**

This section will be similar to your exploratory analysis project. First, provide a summary of the data set similar to your first exploratory analysis: *e.g. this data set contains 398 samples with 7 columns with various data types*. In this summary, provide the data types of your columns (in a table) and then rather than providing tabular statistics and plots for each variable, provide only statistics and plots that seem unusual. For example, if one or two variables have significant missing values or the distribution of the variable is skewed or looks unusual note that. Provide the unusual statistics or plots in this section. Provide any other appropriate plots (e.g. correlation matrix, heatmaps, bar charts, etc.) that you deem necessary.

The dataset being tested contains 17 attributes and 2111 records. This dataset has 8 Float64 variables and 9 Object variables. All data was kept the same while looking through statistics and plots. However, when getting into the logistical Regression extra data cleaning steps took place. By looking at figure 2 the heat map you will see that there is not many good linear correlations in this dataset which seems weird. You would think that weight would fluctuate more with eating habits and physical activity however the heatmap shows otherwise. Object variables FAVC, SCC, SMOKE, Family\_history\_with\_overweight, and CAEC were changed through One-Hot encoding to binary code with 1=yes 0=no. Also, N0beyedsdad was encoded using Label encoding. Thus, Underweight is , normal weight is 1, Obesity level I is 2, Obesity level II is 3, Obesity level III is 4, overweight level I is 5, and overweight level II is 6 . If you look at Figure 1 you will see that there is a close to even split of these dependent variables in the dataset.

**Table 1: Data Types**

|  |  |
| --- | --- |
| *Variable Name* | *Data Type* |
| Age | Float64 |
| Gender | Object |
| Height | Float64 |
| Weight | Float64 |
| CALC | Object |
| FAVC | Object |
| FCVC | Float64 |
| NCP | Float64 |
| SCC | Object |
| SMOKE | Object |
| CH20 | Float64 |
| Family\_history\_with\_overweight | Object |
| FAF | Float64 |
| TUE | Float64 |
| CAEC | Object |
| MTRANS | Object |
| N0beyesdad | Object |

1. **METHODS**
   1. *Data Preparation*

No data normalization was required for this dataset as it was sourced intact from the original authors. No variables were dropped from the data set as all were useful and could be utilized to add definition to the logistic model. The main source of data preparation was through variable encoding. One-Hot encoding was used for the Object variables FAVC, SCC, SMOKE, Family\_history\_with\_overweight, and CAEC in order to make them binary, and thus, usuable in a logistic model. In contrast, Label encoding was used to convert the “N0beyesdad” object into integer values between 0-6 inclusive.

* 1. *Experimental Design*

You will run your model several times with different parameters to see what different results you get. In a table, describe your experimental parameters. Three or four experiments are sufficient. This is where you will describe how you divided your data into train, validate and test data sets. For example:

**Table X: Experiment Parameters**

|  |  |
| --- | --- |
| **Experiment Number** | **Parameters** |
| 1 | All sixteen (16) raw features with test size .15 and random state 20 for train, test and results check |
| 2 | All sixteen (16) raw features with test size .20 and random state 30 for train, test and results check |
| 3 | All sixteen (16) raw features with test size .10 and random state 25 for train, test and results check |

* 1. *Tools Used*

The following tools were used for this analysis: Python v3.5.2 running the Anaconda 4.3.22 environment for Apple Macintosh computer was used for all analysis and implementation. In addition to base Python, the following libraries were also used: Pandas 0.18.1, Numpy 1.11.3, Seaborn 0.7.1, Matplotlib 1.5.3, and SKLearn 0.18.1. We used Anaconda to run Python and Python to write the code. We Pandas to import the dataset and to get dummy variables as well. Numpy was used to create our arrays and Seaborn was used to create the heat map. SKlearn was used to create the training and test sets as well as the label encoder and creating the logistical regression model.

1. **RESULTS**
   1. *Classification Measures/ Accuracy measure*

To compare across the 3 models created in this analysis we used Python to create a classification report and did multiple tests to see which model preformed the best. Model 1, which had a test size of .15 and a random size of 20, had an f1 weighted average score of 71%, a recall weighted average score of 72%, a precision weighted average score of 71%. Model 2, which had a test size of .20 and a random size of 30, had an f1 weighted average score of 69%, a recall weighted average score of 70%, a precision weighted average score of 69%. Model 3, which had a test size of .10 and a random size of 25, had an f1 weighted average score of 71%, a recall weighted average score of 72%, a precision weighted average score of 72%. These scores show that Model 3 is the most accurate model and provides the most accurate predictions. However, when the models were tested using samples Model 1 had a 100% accuracy rate while Models 2 and 3 only had a 33% accuracy rate.

* 1. *Discussion of Results*

We think Model 1 is the best. This due to its 100% accuracy rate when test compared to the other 2 models only being 33% accurate. Model 1 also provided the second-best scores for the classification report to Model 3. Model 1 was only behind 1% in the precision area to model 3 and the f1 score and recall score was the same. With this minimal difference and the better test we think the Model 1 is the best choice. Model 2 is the worst. It only had a 33% test rate accuracy and had the lowest classification report scores for precision, f1, and recall. For these reasons we believe Model 2 is the worst.

* 1. *Problems Encountered*

Overall, the majority of the project went smoothly. However, there were some issues with certain areas. The main difficuly that we initially had was with the encoding of the data. Due to the need for two different types of encoding, the data had to first be separated. Unfortunately, this was only found through trial and error with it eventually finding a solution. Another key issue that presented was the question of how to test whether or not or model predicted the correct result for the dependent variable. Eventually, we decided to take a random sample of the X value, use the independent variables to plus into our prediciton model, then use the iloc function to call that specific entry from the data. This allowed us to check the accuracy of our model against the actual reported data.

* 1. *Limitations of Implementation*

The main limitation to our model was the need to encode our categorical data into mostly binary values. Due to our model being based on a simple logarithmic regression, we were unable to use the data in its raw form. While this linekly did not pose a significant limitation to our model’s accuracy, it is still a limitation that should be noted. Likely, a most advanced statistical regression would be needed to accommodate the categorical variables in the dataset.

* 1. *Improvements/Future Work*

Possible future improvements to our model would mainly be an expanded selection of biological data. While this dataset does provide a decent amount of information, it is limitied by the answers being either self-reported or generated. Thus, a more fleshed out study with more datapoints would be beneficial. This would provide more independent variables to improve the model. Another possible improvement for future work would be using different parameters to further narrow down the best model for this dataset.

1. **CONCLUSION**

This experiment was designed to take various metrics about lifestyle are use them to predict overall weight level. Through the analysis, design, training, and implementation of this regression model, we feel that we have demonstrated the development of a machine learning model that, while not infallible, can predict the dependent variable with a good level of accuracy. Model 1 was decided as the best model due to both its classification report and the manual checks of the results that were preformed. Overall, working within the dataset, this model did provide decent results with the possibility of improvement through dataset additions or a more complex regression model for the basis.

Finish up with a paragraph or two of summarizing your problem, the results and your conclusions (good model, bad model, needs more work, etc.).

 