Predicting Obesity in Individuals

Kadin McWilliams, kmcwilliams@bellarmine.edu

ABSTRACT
Obesity is a physical disease that affects every country all over the world. There are serious health risks to being obese and knowing if one is obese is important for one's health. To better understand that question and answer it. An analysis was done on obesity from an obesity dataset. For this analysis is used logistic regression to predict obesity in individuals. 4 different models are used to investigate this issue. This analysis also investigates the difference in using 80% training and 20% testing models vs 70% training and 30% testing models. The results found that predicting obesity class 3 in all models was extremely accurate. Model 1 had the best f1-score of 0.74 and predictions for all obesity levels. Model 1 needs more work to be able to be in a medical setting.

I. INTRODUCTION

Obesity has been a prevailing problem in the United States and researchers are still trying to fully understand this phenomenon. This project's main goal is to be able to use logistic regression to accurately guess obesity levels of any individual. Other goals of this project are to learn more about training data, test data, and draw conclusions about obesity using results found. To complete the regression, I will be using an obesity and health lifestyle choice dataset to perform regression on. The dataset is from a Spanish research study on the potential factors of obesity in people.

II. BACKGROUND

Obesity is a physical disease where an individual has a BMI of 30.0 or higher. It is categorized into three classes; class 1 (low-risk) obesity is a BMI from 30 to 34.9, class 2 (moderate-risk) obesity is a BMI from 35 to 39.9, and class 3 (high-risk) obesity is a BMI greater than 40. Obesity can have many negative health impacts including cardiovascular disease, type 2 diabetes, musculoskeletal disorders, sleep apnea, liver disease, high blood pressure, and some forms of cancer (endometrial, breast, and colon). These conditions can lead to a premature death or disability in an individual. Obesity is a treatable and preventable disease; however, genetic factors can lead people to being more susceptible to becoming obese. The dataset used for this investigation was collected from participants living in Mexico, Peru, and Colombia. This data is self-reported data on individuals' physical activity, caloric intake, and health lifestyle. Health lifestyle choices include things like eating vegetables in one's diet, how many meals a day one has, daily water intake, and daily physical activity. In general, health lifestyle choices are how an individual lives their life with physical health in mind. The authors of this data collected height and weight to calculate obesity levels for each of the participants in this study. 23% of the data was directly collected from a web platform. The other 77% of the data was generated synthetically using WEKA and the SMOTE filter. The original goal of the author's study was to try to find the predicting factors for obesity.

III. EXPLORATORY ANALYSIS

This dataset contains 2,111 samples with 17 columns with 9 categorical and 8 numerical variables. The dataset contained no missing values within the dataset. However, some of the values were in decimal form when they should have only been integers. For example, when asked, "How much water do you drink daily?" The participants reported they drink 1.0234 bottles of water. The participants would most likely only use whole numbers to answer the question instead of a specific decimal. I hypothesize that the synthetic data that was created did not have certain values set to integers only, causing the unusual decimal inputs. There are 4 other columns that have the same issue as previously stated. This dataset has also been translated from Spanish to English, so the variable names are not in English. There has been other important information lost in translation such as unit values and the specificity of questions. A complete list of the dataset, with the variable names translated to English, is presented in **Table 1**.

Table 1: Data Types

Variable Name	*Data Type*
V1 Age	float64
V2 Gender	object

V3 Height	float64
V4 Weight	float64
V5 Alcohol	object
V6 HighCaloricIntake	object
V7 VegtableIntake	float64
V8 DailyMeals	float64
V9 CalorieMonitoring	object
V10 Smoke	object
V11 DailyWaterIntake	float64
V12 family_history_with_overweight	object
V13 PhysicalActivity	float64
V14 TechnologyUsage	float64
V15 FoodBetweenMeals	object
V16 MTRANS	object
V17 ObesityLevel	object

When getting a general idea of what the data set contains, there are a few instances of unusual occurrences in the dataset. When looking at the heat map correlations (figure 1) there appears to only be a moderate correlation between height and weight for this dataset. I would have suspected that they would have had a stronger correlation considering normally they are strongly correlated. Other unusual correlations from the heat map (figure 1) showed that height tended to have a stronger correlation than weight did with daily meals. I would think that weight would have a similar or stronger correlation with daily meals, as one must eat to gain weight.

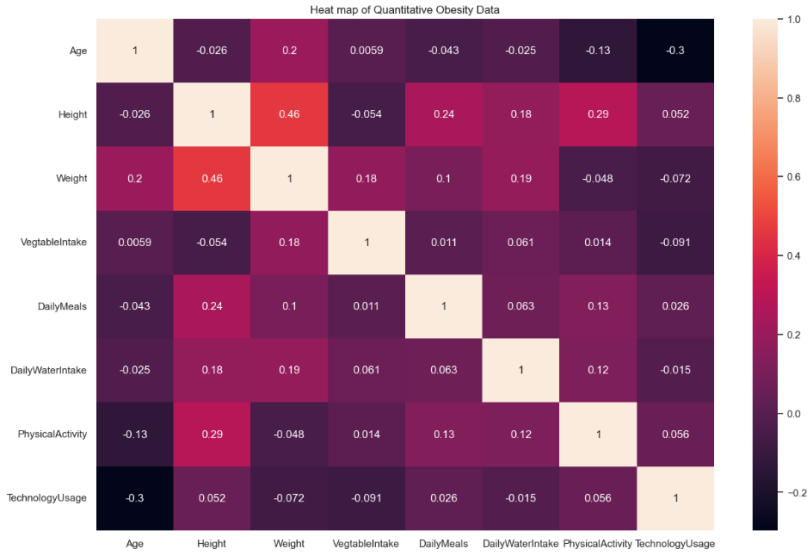

Figure 1: Heat map of Quantitative Obesity Data

The average participant in this study is 5 feet and 7 inches tall and weighs 190 pounds. This means that the average BMI of a participant is 29.8, which is 0.2 away from being considered obese. I think this statistic helps understand the lower correlation between height and weight. The average person in this study tends to weigh more than a typical individual, which would lower the correlation between the two. The pie plot (figure 2) shows that 73.52% of participants in the study are obese or overweight and 26.48% of participants are normal weight or underweight. The pie plot (figure 2) also shows that this data is heavily skewed toward obese and overweight obesity levels.

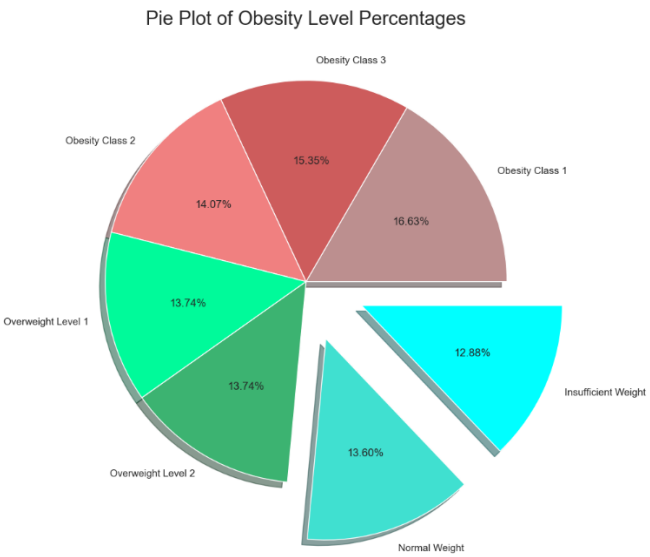

Figure 2: Pie Plot of Obesity Levels Percentages

The linear distribution plot (figure 3) investigates the relationship between height and weight with obesity level. The plot looks at the individual weights and heights for each obesity level and displays the line of best fit. In the graph the lines of best fit from insufficient weight to obesity class 2 all seem to have similar slopes. However, obesity class 3 has a less steep slope meaning the correlation between height and weight is lower than the rest of the obesity levels. This shows that people with obesity class 3 have a significantly larger weight to height ratio. Meaning that people with obesity level 3 are going to have significantly more weight on them than any other obesity level. This graph also gives us insight on what to expect from the height and weight of someone from each obesity level.

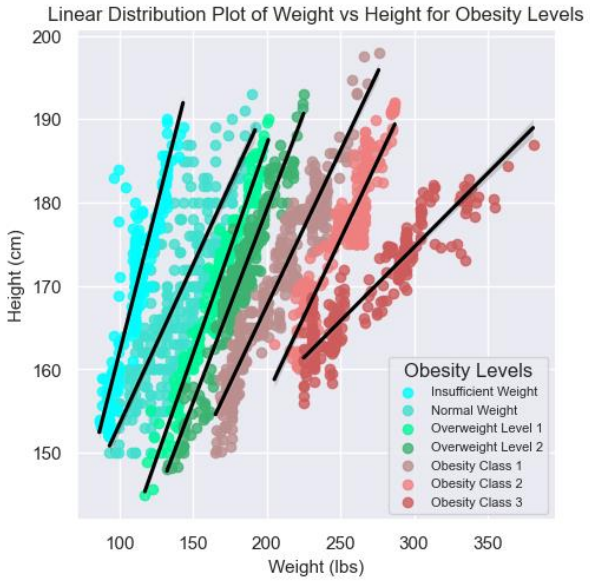

Figure 3: Linear Distribution Plot of Weight vs Height for Obesity Levels

The last unusual occurrence in this data set is the frequency of time spent on technology. The question asked for a general estimate of how much time a participant spends on all forms of screen technology. The question did not specify in what units or duration of time it is over. The other questions are all asked in the daily format and the histogram (figure 4) below shows a minimum value of 0 and maximum value of 2. I made an educated guess that this question asked the number of hours a day that people spend on devices, but this dataset has many 0 entries.

Considering that the survey was taken online, having approximately 33% of the entries be zero is illogical. Everyone who participated in this survey must have had at least 10 minutes of internet usage to take the survey. I think that the translation of this column makes getting meaningful information from it nearly impossible. I think another potential reason for the 0 entries could be how the synthetic data was produced.

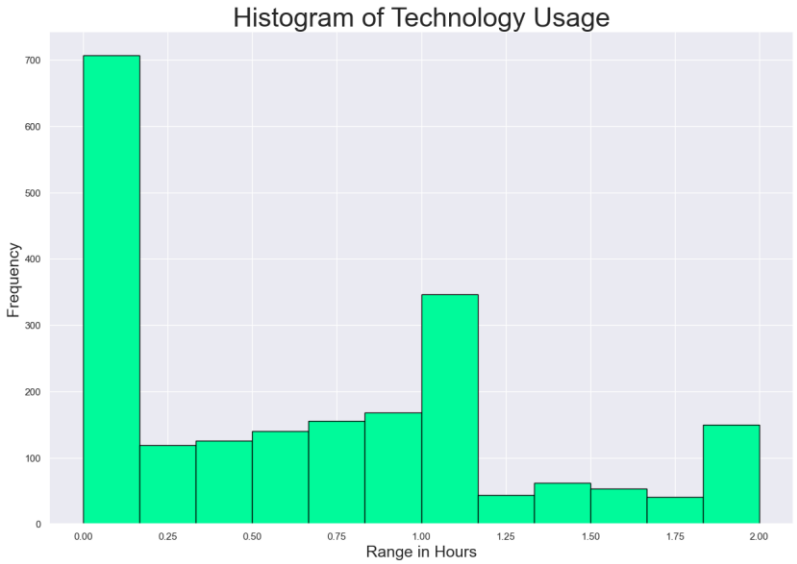

Figure 4: Histogram of Technology Usage

Overall, this dataset at a surface level comes across clean and put together. However, after careful analysis of the dataset there are a few unusual occurrences within it. Some of the values are inputted as a form that makes no logical sense. The columns are hard to interpret due to the translation and leave some columns up to guessing for what they are providing. The dataset shows an interesting trend of obesity classes. Specifically, how the weight of individuals can significantly affect the correlation between height and weight. Other than the usual occurrences listed the dataset is useful and provides good meaningful information.

IV. METHODS

A. Data Preparation

Preparing the dataset for logistical regression required splitting the dataset into independent variables and one dependent variable. The independent variable decided for this analysis was obesity level. The independent variables for this analysis are age, gender, height, weight, alcohol consumption, if one has a high caloric intake, daily vegetable intake, daily meals, if one monitors their calories, if the individual smokes, daily water intake, having a family history with overweight family members, daily physical activity, technology usage per day, if one eats food in between meals, and daily transportation method. I added and dropped a variable I created called imperial height. I created the column to make data interpretation easier for me during the data exploration process. I removed it from the independent variables as the data type was categorical and the dataset already had a quantitative height variable. I used one hot shot encoding for all the independent categorical variables. I then normalized all the independent variables using scaling to range normalization. This put all the independent variables between the values of 0 and 1. After I normalized the data, I created two experimental independent variable groups. The first group contains all the independent variables, and the second group contains all the independent variables except height and weight. For the dependent variable I used ordinal encoding to encode each of the types of obesity level.

B. Experimental Design

For my experiments I wanted to test 4 different ideas. I wanted to see the difference between using 80% training data and 20% testing data versus 70% training data and 30% testing data. I then wanted to apply both different ways of training data to my two independent variable groups. The first independent group, that will be referred to as raw data, is testing to see how much of an impact height and weight have on predictions of obesity. Raw data is also being used as a control experiment to compare to the second group. The second group, that will be referred to as

health lifestyle choices, is testing to see if the model can predict obesity based on health lifestyle choices only. **Table 2** below shows all the experiments and their parameters.

Table 2: Experiment Parameters

Experiment Number	Parameters
1	All independent variables with 80/20 split for train and test.
2	All independent variables with 70/30 split for train and test.
3	All independent variables excluding height and weight with 80/20 split for train and test.
4	All independent variables excluding height and weight with 70/30 split for train and test.

C. Tools Used

The following tools were used for this analysis: Python v3.11.4 running Anaconda v2.5.0 for a Windows 10 computer, was used for all analysis and implementation. In addition to base Python, the following Python libraries were used: Pandas v1.5.3, Matplotlib v3.7.1, Seaborn v0.12.2, Numpy v1.24.3, SKLearn v1.3.0, Statsmodels v0.14.0. Anaconda was used to download and use Python. Python was the programming language used for all analysis and implantation of the dataset. Pandas and Numpy were used for data frame creation and manipulation as well as basic statistical analysis. Matplotlib and Seaborn were used to create various graphs and plots. SKLearn was used to create confusion matrices, logistic regression model, and logistic regression data processing. Statsmodels was used to get a deeper statistical understanding of the dataset. Many of these libraries were used in combination with each other throughout the process of analysis and implementation.

V. RESULTS

A. Classification Measures/ Accuracy measure

Experiment one had an overall f1-score of 0.74 with obesity class 3 having an f1-score of 1.00. This experiment had the highest f1-score out of all the experiments. Overweight level 2 had the lowest f1-score with a score of 0.33. F1-scores towards the extremes (Obesity Class 3 and underweight) have much higher scores than scores in the middle. This model was also able to recall 100% of the obesity class 3 values with a precision of 94%. **Table 3** shows all precision, recall, and f1-scores for all values.

Table 3: 80/20 Raw Classification Report

80/20 Raw Classification Report				
	Precision	*Recall*	*F1-Score*	*Support*
Under Weight	0.86	0.84	0.85	57
Normal Weight	0.64	0.64	0.64	55
Overweight Level 1	0.66	0.65	0.65	62
Overweight Level 2	0.58	0.33	0.33	55
Obesity Class 1	0.62	0.67	0.67	63
Obesity Class 2	0.73	0.95	0.95	55
Obesity Class 3	0.94	1.00	1.00	76
Accuracy			0.74	423
Macro Avg	0.72	0.72	0.71	423
Weighted Avg	0.73	0.74	0.72	423

The confusion matrix of experiment one (figure 5) shows that predictions tend to be over the weight of the actual weight of the individual. The confusion matrix has the most correct guesses at the polar ends of the obesity spectrum and seems to guess a higher weight than lower. Overweight level 2 has the most wrong guesses for the weight and

guesses are spread out between being underweight and overweight of target weight. Overweight level 2 has the most incorrect guesses out of any of the other obesity levels.

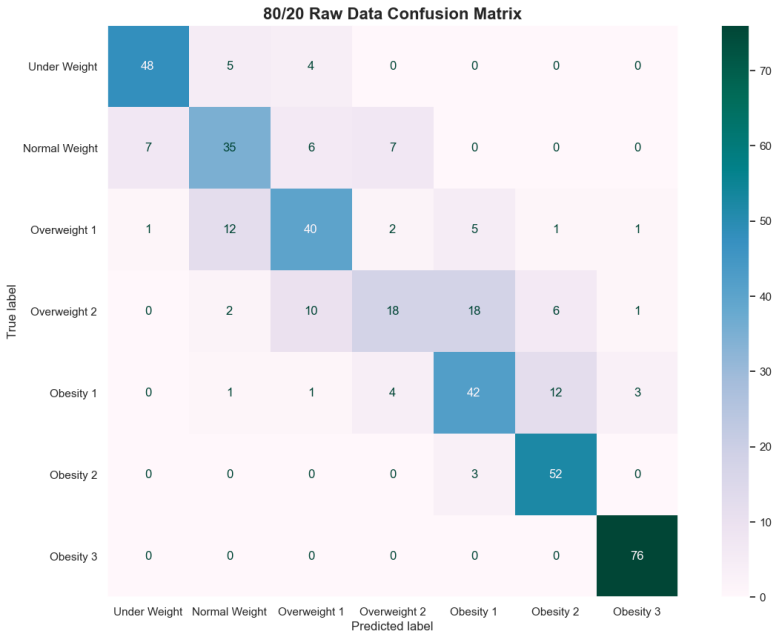

Figure 5: 80/20 Raw Confusion Matrix

Experiment two had the second highest f1-score out of the experiments with a score of 0.74. This score is 0.01 off from the f1-score from experiment one. Obesity class 3 had an f1-score of 0.97 and underweight had an f1-score of 0.80. Overweight level 2 was again the lowest f1-score of 0.47. The f1-score in experiment 2 is 0.14 more than the score in experiment one. Overall, the scores from experiment two are just slightly less than experiment one. The full list of precision, recall, and f1-scores for experiment 2 are in **Table 4**.

Table 4: 70/30 Raw Classification Report

70/30 Raw Classification Report				
	Precision	*Recall*	*F1-Score*	*Support*
Under Weight	0.77	0.84	0.80	80
Normal Weight	0.62	0.57	0.59	81
Overweight Level 1	0.71	0.67	0.69	90
Overweight Level 2	0.60	0.38	0.47	81
Obesity Class 1	0.66	0.61	0.63	102
Obesity Class 2	0.68	0.94	0.79	86
Obesity Class 3	0.93	1.00	0.97	114
Accuracy			0.73	634
Macro Avg	0.71	0.71	0.70	634
Weighted Avg	0.72	0.73	0.72	634

Experiment two's confusion matrix (figure 6) is extremely similar to the confusion matrix of experiment one (figure 5). This matrix follows the same trends as experiment one's confusion matrix with accurate predictions at the poles and losses accuracy at the middle of the matrix. This matrix is also more likely to predict that someone weighs more than they actually do. The middle of this matrix also tends to have more overweight guesses than underweight for the target weight.

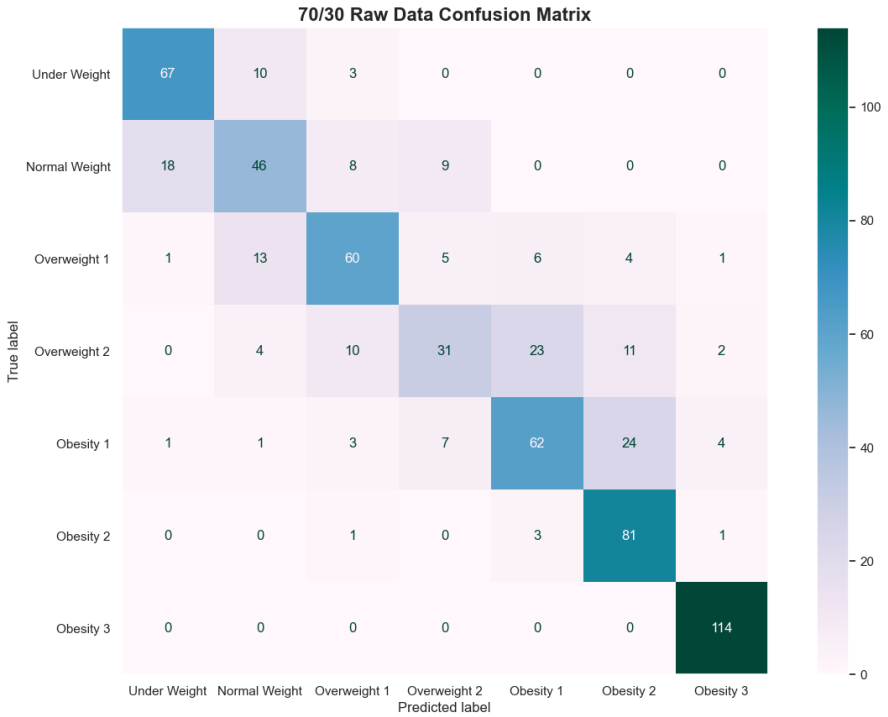

Figure 6: 70/30 Raw Confusion Matrix

Experiment three had the third highest f1-score of 0.63. The scores from the lifestyle choices data have a much lower f1-score across the board than the raw dataset scores. However, it is important to note that obesity class 3 has an f1-score of 0.93. Underweight has a f1-score of 0.69 which is a 0.16 lower f1-score than experiment one. Overall, the f1-scores are all lower for experiment 3 except for the recall for obesity class 3. This experiment had the lowest f1-score for overweight level 2 at 0.21. All recall, precision, and f1-scores for experiment 3 are in **Table 5**.

Table 5: 80/20 Lifestyle Classification Report

80/20 Lifestyle Classification Report				
	Precision	*Recall*	*F1-Score*	*Support*
Under Weight	0.73	0.65	0.69	57
Normal Weight	0.64	0.49	0.56	55
Overweight Level 1	0.62	0.42	0.50	62
Overweight Level 2	0.38	0.15	0.21	55
Obesity Class 1	0.51	0.68	0.59	63
Obesity Class 2	0.52	0.91	0.66	55
Obesity Class 3	0.87	1.00	0.93	76
Accuracy			0.63	423
Macro Avg	0.61	0.61	0.59	423
Weighted Avg	0.62	0.63	0.61	423

The confusion matrix for experiment 3 (figure 7) shows similar results to experiment one and two but with lower correlation values. Obesity class 3 still has an extremely high correct predictions of weight. However, the other weights all drop drastically in obesity level correct predictions. Overweight level 2 has more incorrect predictions

than correct predictions. Most predictions for overweight level 2 are in the obesity class 2 category. Underweight also has more incorrect predictions than in the last experiment.

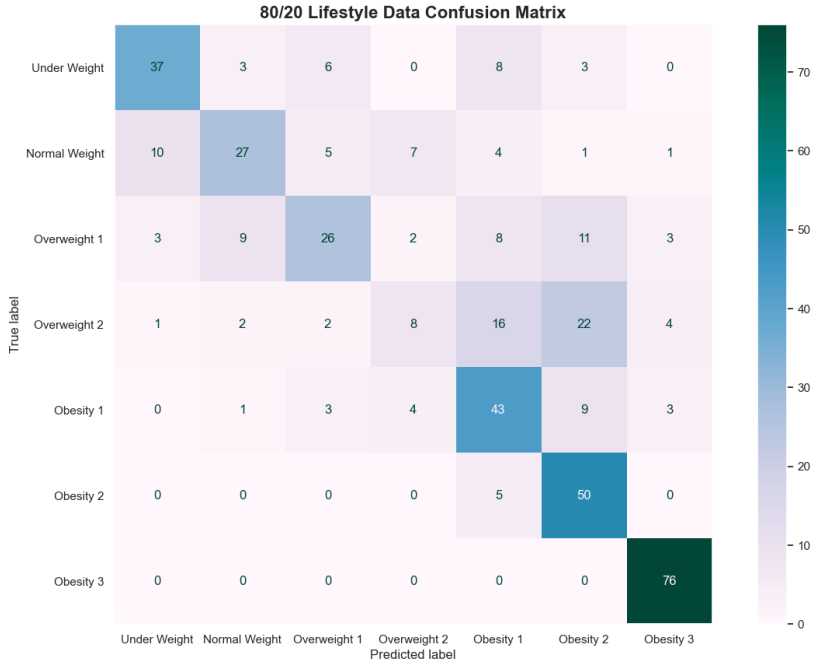

Figure 7: 80/20 Lifestyle Confusion Matrix

The final experiment had the lowest scores across the board compared to all other experiments. **Table 6** shows that all f1-scores, precision, and recall are the same or lower than all previous experiments. Obesity class 3 still has an extremely high f1-score of 0.93 and has remained consistently high through all experiments. Another important statistic from **Table 6** is that overweight level 2 has an f1-score rating of 0.25.

Table 6: 70/30 Lifestyle Classification Report

70/30 Lifestyle Classification Report				
	Precision	*Recall*	*F1-Score*	*Support*
Under Weight	0.64	0.68	0.65	80
Normal Weight	0.63	0.44	0.52	81
Overweight Level 1	0.57	0.34	0.43	90
Overweight Level 2	0.44	0.17	0.25	81
Obesity Class 1	0.52	0.61	0.56	102
Obesity Class 2	0.5	0.91	0.65	86
Obesity Class 3	0.86	1.00	0.93	114
Accuracy			0.61	634
Macro Avg	0.60	0.59	0.57	634
Weighted Avg	0.61	0.61	0.59	634

The final experiment's confusion matrix (figure 8) shows like experiment 3's confusion matrix. The difference is that experiment 4 has more incorrect predictions from underweight to obesity class 2. This matrix is similar in overweight level 2 to the rest of the matrix as it has the lowest f1-score and has more incorrect guesses than correct ones. Overall, this confusion matrix was the least accurate at guessing weight than across all other experiments.

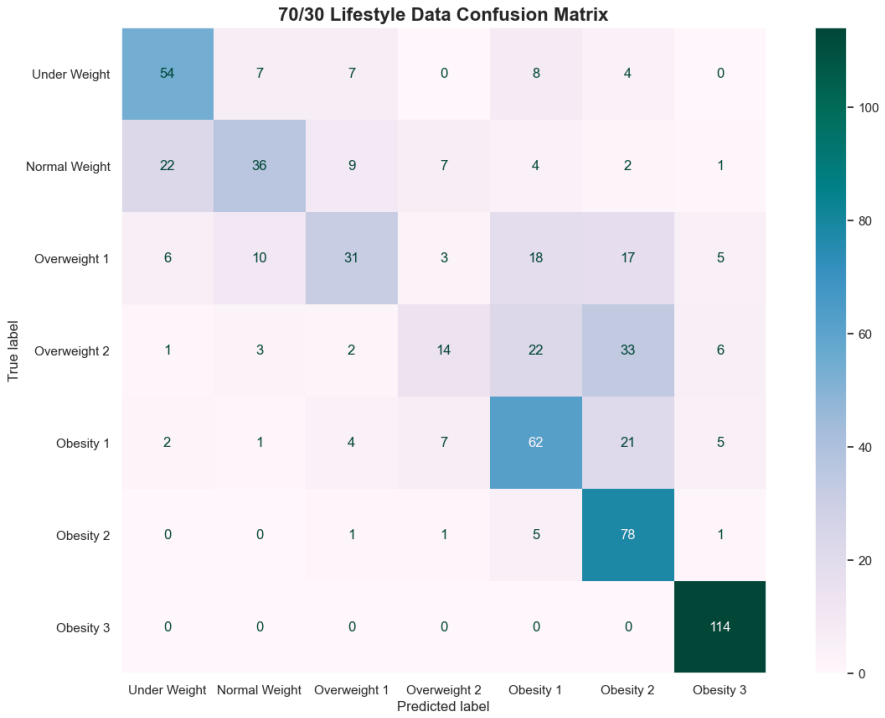

Figure 8: 70/30 Lifestyle Confusion Matrix

B. Discussion of Results

The model for experiment 1 proved to be the best classification out of any of the other models. While it still did have an f1-score that I would think would not be usable in a medical setting, it was a decent model. The model is not good at predicting mid to overweight range weights in individuals but excels at underweight and obesity class 3. I thought that experiment 3 provided good insight into external factors that predict obesity. It provides good insight since the f1-score for obesity class 3 was like model 1 and was extremely accurate. Considering that obesity class 3 was accurately predicted across all 4 models this dataset can possibly give us an idea of what some of the factors for obesity are. 99% of people from this data set with obesity class 3 sometimes drink alcohol, take public transportation, are women, have a high caloric intake, and do not smoke. 100% of participants in obesity class 3 do not do any form of calorie monitoring, eat 3 meals a day, and eat 3 vegetables a day, have a family history of family members being overweight. People in obesity class 3 have similar averages on the rest of the independent variables compared to the average participant in this data set. It seems that model 1 would be the best for predicting the weight of any individual since it had the highest correct predictions and f1-score. Model 3 would be a good indication based on lifestyle choices to see if you have obesity class 3. Across the raw data and the lifestyle choice data the 80% training data and 20% testing data had better f1-scores and correct predictions than the 70% training and 30% testing models.

C. Problems Encountered

When obtaining the dataset there were initial problems in finding a good dataset that I could perform a regression analysis on. Many of the datasets didn't have a clear independent variable to predict or were missing too much data. With this dataset specifically the translation from Spanish to English caused many problems with understanding variables. Many of the variables had weird numbers that didn't make sense and had no frame of reference for what the possible values could mean. While preparing the data I ran into the problem of the independent variables being encoding in a strange way. Overweight level 1 and 2 were both at the end of the index and not in the middle. This made the confusion matrix hard to understand when looking at the graph. I was able to fix the encoding problem by using the OrdinalEncoder from SKLearn. Changing the y value to be ordinally encoded meant I had to adjust the learning models to have the ravel function be in the actual training model. The final notable problem I encountered when creating the learning model was that the data would not train. The console told me to raise the iteration value higher and higher. I eventually had up to 10,000 iteration value but the code would still not run. To get the model to run correctly I had to use scaling to range normalization for the training sets to be able to train the data.

D. Limitations of Implementation

The limitations of this model are that it cannot predict people who are overweight well at all and the dataset that was used. This model works well in the extremes but works poorly in the middle and is not generalizable to many other people. Another issue is that the dataset seems to have a significant amount of internal validity. The data is skewed towards overweight people leading me to believe that this survey was targeted toward overweight people. 77% of the data is synthetically made meaning the generalizability of this data is far less as most of the data isn't collected from the outside world.

E. Improvements/Future Work

To improve the model in the future I would like to do find a different dataset. I think this dataset had too many instances of information being left up to me to decide what it meant. There is too much room for error in interpreting this dataset. Another improvement would be to see if changing some of the one hot encoding variables to be ordinally encoded would help bring up the f1-scores of the overweight values. I think also adding a BMI variable would help with the accuracy of the models.

VI. CONCLUSION

In summary, obesity is a problem that is faced all around the world and the factors can be hard to predict. The logistic regression model 1 had the best f1-score accuracy at 0.74. All models were able to accurately predict obesity class 3 in individuals, even based on lifestyle choices alone. I think that overall, this model needs more work to be effective and useful in the real world. There is too much error in the overweight level 1 and 2 for this model to be used in the real world. Using the 80% training set and 20% testing set performed the best across both models. If this model is improved upon, I think predicting obesity based on lifestyle choices will be a more realistic goal.

REFERENCES

- Kaggle Data set
- Information about obesity
- Definition of obesity
- https://www.geeksforgeeks.org/adding-new-column-to-existing-dataframe-in-pandas/
- https://stackoverflow.com/questions/51865367/cannot-convert-the-series-to-class-int?newreg=d8e32be70e3541379c7722d61065bf11
- https://stackoverflow.com/questions/13148429/how-to-change-the-order-of-dataframe-columns
- https://medlineplus.gov/ency/patientinstructions/000348.htm#:~:text=These%20ranges%20of%20BMI%20are,to%20or%20greater%20than%2040.0
- https://stackoverflow.com/questions/11346283/renaming-column-names-in-pandas
- https://www.freecodecamp.org/news/how-to-rename-a-column-in-pandas/
- https://python-graph-gallery.com/101-make-a-color-palette-with-seaborn/#:~:text=You%20can%20control%20the%20colors,a%20parameter%20to%20set_palette%20function.
- https://www.geeksforgeeks.org/how-to-change-seaborn-legends-font-size-location-and-color/
- https://stackoverflow.com/questions/45201514/how-to-edit-a-seaborn-legend-title-and-labels-for-figure-level-functions
- https://stackoverflow.com/questions/28468584/seaborn-factorplot-set-series-order-of-display-in-legend
- https://www.geeksforgeeks.org/how-to-change-order-of-items-in-matplotlib-legend/
- https://matplotlib.org/stable/gallery/color/named_colors.html
- https://matplotlib.org/3.1.1/gallery/pie_and_polar_charts/pie_features.html
- https://www.w3schools.com/python/matplotlib_pie_charts.asp
- https://developers.google.com/machine-learning/data-prep/transform/normalization
- https://www.youtube.com/watch?v=Fw5iiyIHzew&ab_channel=SoumilShah
- https://machinelearningmastery.com/one-hot-encoding-for-categorical-data/
- https://community.alteryx.com/t5/Alteryx-Designer-Desktop-Discussions/Get-output-in-specific-columns-by-skipping-some-columns/td-p/927253
- https://stackoverflow.com/questions/1987694/how-do-i-print-the-full-numpy-array-without-truncation

- https://www.w3schools.com/python/matplotlib_grid.asp
- https://learnpython.com/blog/loop-over-multiple-lists/
- https://scales.arabpsychology.com/stats/how-to-use-bold-font-in-matplotlib-with-examples/
- https://www.youtube.com/watch?v=15uClAVV-rI&ab_channel=RyanNolanData
- https://stackoverflow.com/questions/20625582/how-to-deal-with-settingwithcopywarning-in-pandas
- https://www.youtube.com/watch?v=15uClAVV-rI&ab_channel=RyanNolanData
- https://stackoverflow.com/questions/48342098/how-to-check-python-anaconda-version-installed-on-windows-10-pc