DSC 680

Inman, Gracie

White Paper

05/05/24

Obesity Prediction

In the United States, there is an entire industry dedicated to weight loss. Obesity is a disease that affected 1 out of 8 people in the world in 2022 (World Health Organization, 2024). Obesity is often determined using the height and weight of an individual to calculate their BMI. According to the World Health Organization, “Obesity is a chronic complex disease defined by excessive fat deposits that can impair health. Obesity can lead to increased risk of type 2 diabetes and heart disease, it can affect bone health and reproduction, and it increases the risk of certain cancers” (2024). There are numerous risk factors for obesity. The obesity dataset was obtained from Kaggle and contains several factors. According to Fatemeh Mehrparvar (2022), the factors contained within this dataset are:

- Gender: Feature, Categorical, "Gender"
- Age: Feature, Continuous, "Age"
- Height: Feature, Continuous
- Weight: Feature Continuous
- family_history_with_overweight: Feature, Binary, " Has a family member suffered or suffers from overweight? "
- FAVC: Feature, Binary, " Do you eat high caloric food frequently? "
- FCVC: Feature, Integer, " Do you usually eat vegetables in your meals? "
- NCP: Feature, Continuous, " How many main meals do you have daily? "

- CAEC: Feature, Categorical, " Do you eat any food between meals? "
- SMOKE: Feature, Binary, " Do you smoke? "
- CH2O: Feature, Continuous, " How much water do you drink daily? "
- SCC: Feature, Binary, " Do you monitor the calories you eat daily? "
- FAF: Feature, Continuous, " How often do you have physical activity? "
- TUE: Feature, Integer, " How much time do you use technological devices such as cell phones, video games, television, computers, and others? "
- CALC: Feature, Categorical, " How often do you drink alcohol? "
- MTRANS: Feature, Categorical, " Which transportation do you usually use? "
- NObeyesdad: Target, Categorical, "Obesity level"

This covers a wide range of topics that will be key in analysis. The data contains 77% synthetically generated data and 23% web-obtained data (Mehrparvar, 2024). Using the above dataset, KNN, Naïve Bayes, Logistic Regression, and Random Forest models will be trained to determine if someone is likely to be obese or become obese. Before the models can be trained, the data must be prepped for analysis.

Data preparation is a key step in the analysis. The first step in the analysis was visualizing each attribute. Shown below in figures 1-3 are the plots of the weight, height, and obesity levels.

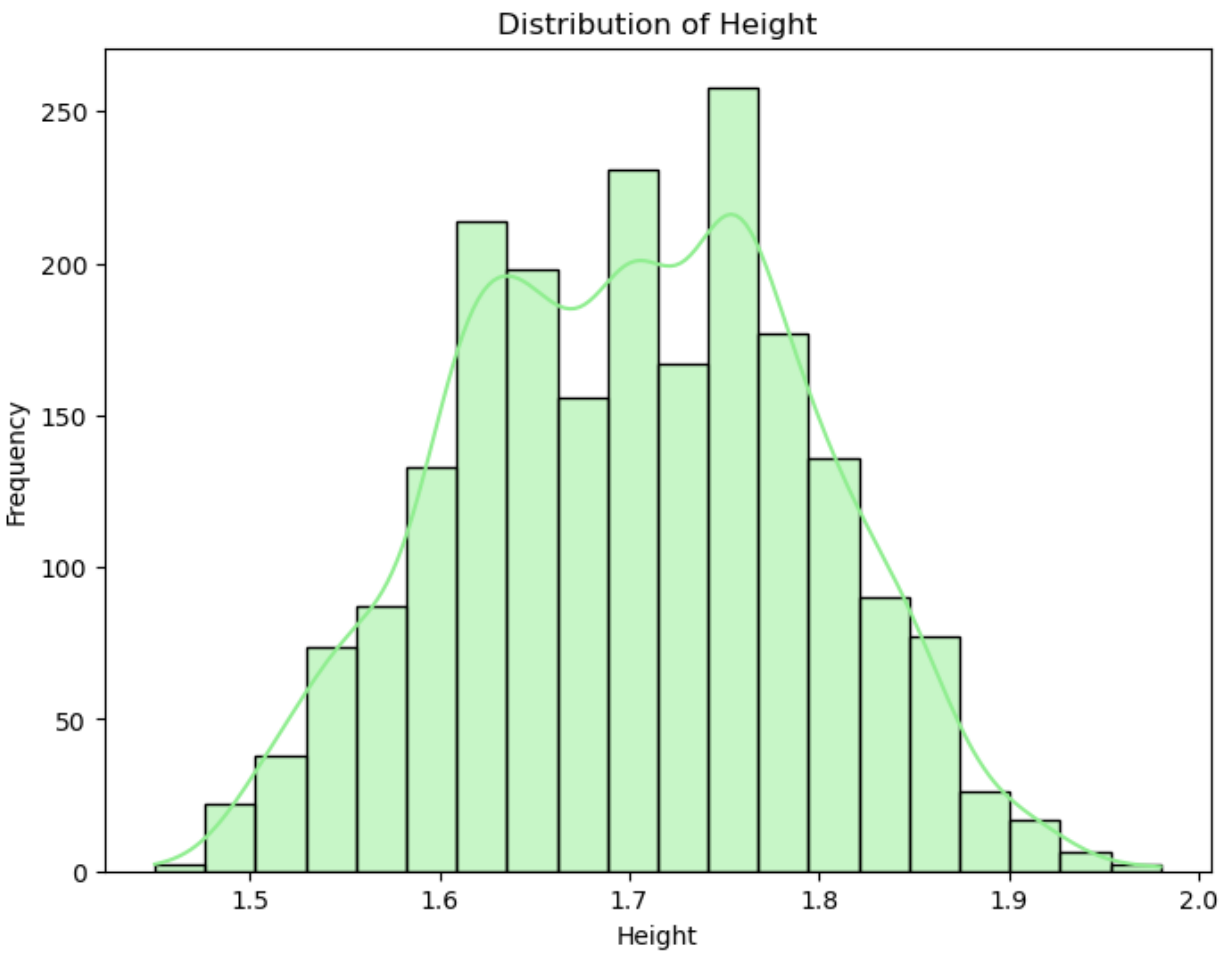

Figure 1: Distribution of height

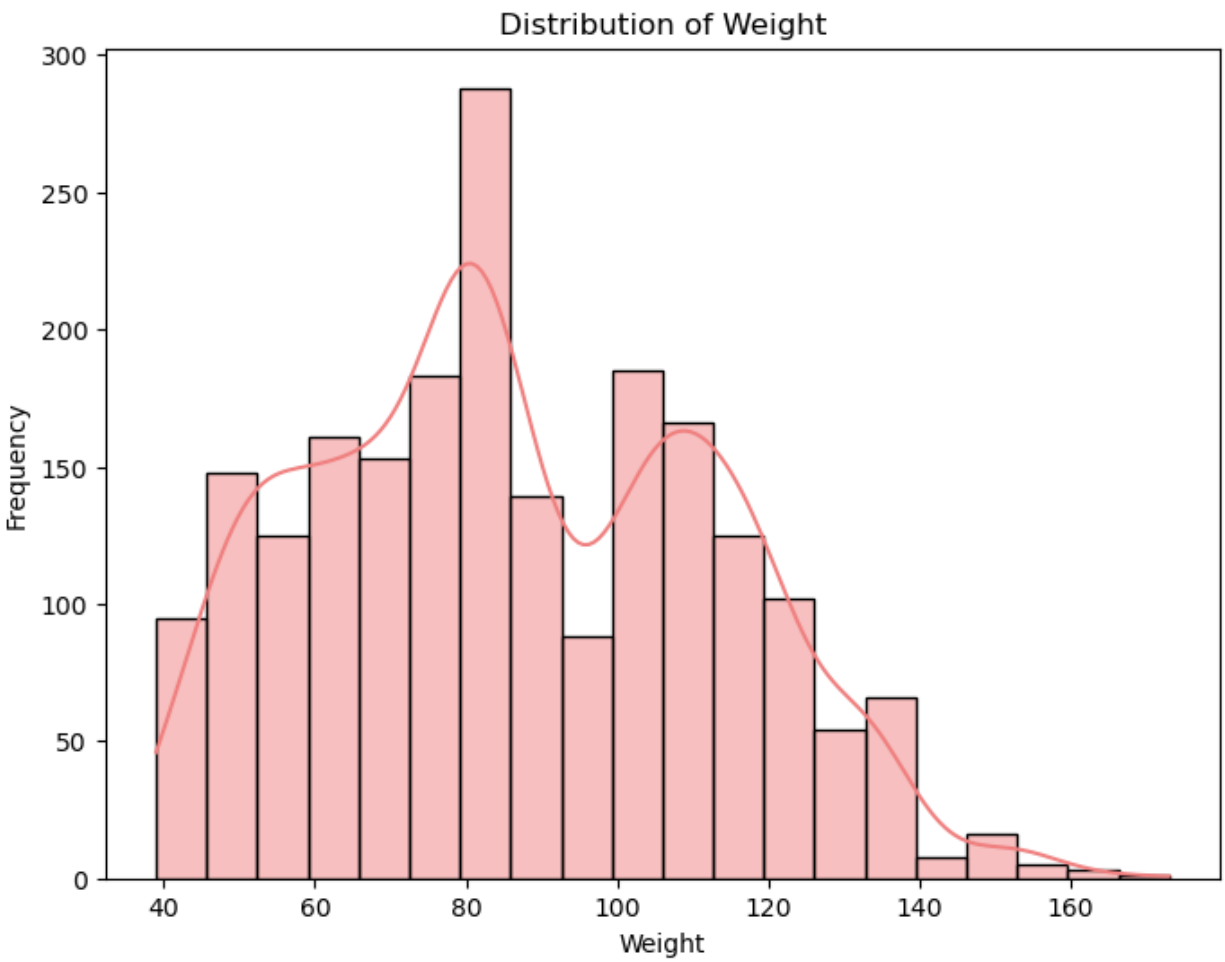

Figure 2: Distribution of weight.

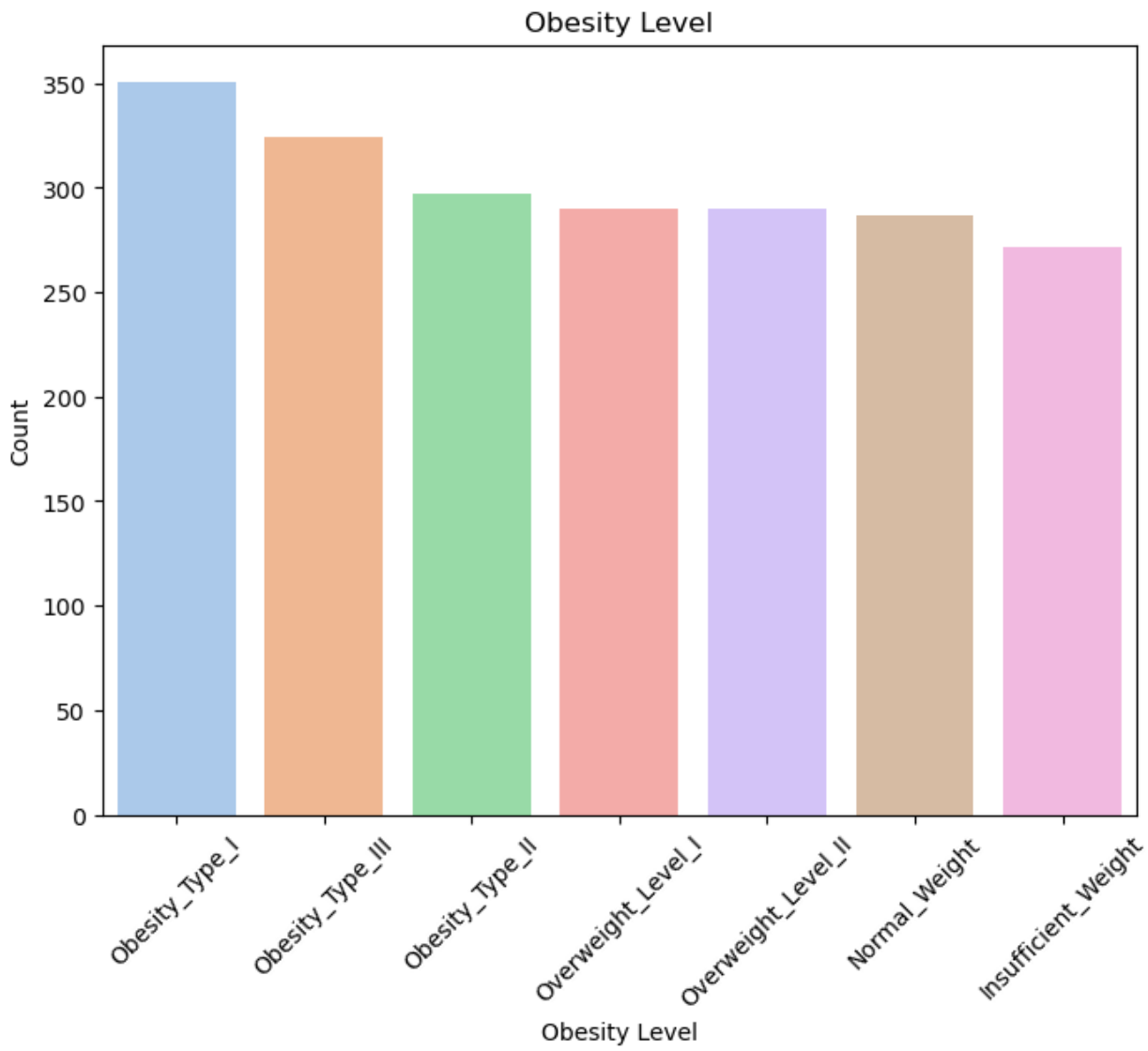

Figure 3: Distribution of obesity level.

After each column was visualized, the data was checked for missing and unique values. There were no missing values present in the dataset. The numerous obesity and overweight variables were combined into one overweight and one obesity variable. The unique categorical variables were encoded to allow for analysis. The encoded variables can be seen in appendix B. Once converted, the data was split into training and testing, and the KNN, Naïve Bayes, Logistic Regression, and Random Forest models were trained.

After performing the initial analysis, it was determined that the model had the characteristics of being overfit. After looking over the attributes, it was determined that since height and weight are used to determine someone's BMI it could be causing the model to be overfit. Height and weight will be excluded from the analysis to prevent overfitting and the models were retrained. This also allows the model to focus on lifestyle factors versus weight. The following are the obtained metrics from each model.

Table 1: KNN				
	Precision	**Recall**	**F1-score**	**Support**
0	0.68	0.91	0.78	56
1	0.63	0.19	0.30	62
2	0.80	0.96	0.87	199
3	0.81	0.67	0.73	106
Accuracy			0.77	423
Macro avg	0.73	0.68	0.67	423
Weighted avg	0.76	0.77	0.74	423

Table 2: Naïve Bayes				
	Precision	**Recall**	**F1-score**	**Support**
0	0.39	0.66	0.49	56
1	0.42	0.18	0.25	62
2	0.71	0.94	0.81	199
3	0.58	0.21	0.31	106
Accuracy			0.61	423
Macro avg	0.52	0.50	0.46	423
Weighted avg	0.59	0.61	0.56	423

Table 3: Logistic Regression				
	Precision	**Recall**	**F1-score**	**Support**
0	0.52	0.54	0.53	56
1	0.50	0.24	0.33	62
2	0.72	0.90	0.80	199
3	0.52	0.42	0.47	106
Accuracy			0.64	423
Macro avg	0.56	0.53	0.53	423
Weighted avg	0.61	0.64	0.61	423

Table 4: Random Forest				
	Precision	**Recall**	**F1-score**	**Support**
0	0.95	0.95	0.95	56
1	0.71	0.73	0.72	62
2	0.91	0.94	0.93	199
3	0.86	0.78	0.82	106
Accuracy			0.87	423
Macro avg	0.86	0.85	0.85	423
Weighted avg	0.87	0.87	0.87	423

Based on the evaluation metrics, it was determined that the random forest model was the best model for the given dataset. It had the highest precision, recall, and F1 score. Figure 4 shown below shows the confusion matrix for the random forest model. The model supports the evaluation metrics findings and shows the model can predict someone's obesity level using their lifestyle.

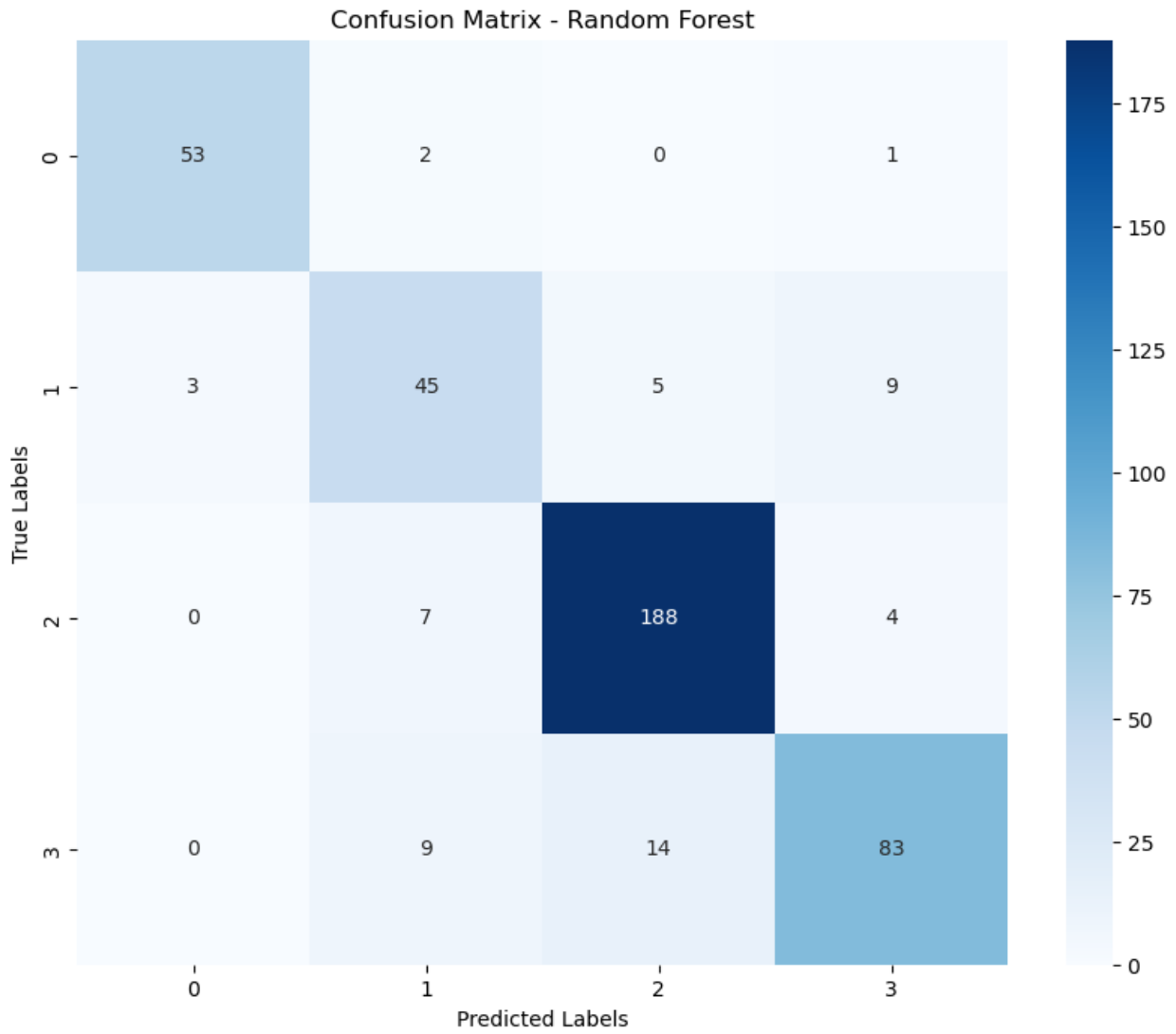

Figure 4: Confusion matrix of the Random Forest Model.

Assumptions are always important to consider in each analysis. Throughout the analysis, I assumed that the generated data would accurately reflect the actual data. This could cause misrepresentation throughout the analysis. Originally, I assumed that the weight and height columns would not cause overfitting. This assumption was incorrect, and the analysis had to be redone without the weight and height columns. I also am assuming that the model will perform well on non-generated data when the model is deployed. The key assumption I am making is that

someone with similar lifestyle habits can become or be in the same obesity category as someone with similar lifestyle habits.

A limitation of this dataset is that it is primarily synthetic data. This can cause the model to not perform as well on real-world data. A challenge will be obtaining a substantial amount of real-world data to test and optimize the model. This model can be used for customer targeting in the weight loss industry. It could also be used as a risk assessment or quiz for public use.

For the model to be implemented, the data should be tested on unseen real-world data. This could be obtained through surveys. Once tested and optimized, it could be incorporated into the customer targeting systems or used to create a risk assessment quiz.

Ethical concerns are always something that needs to be considered. One concern is this is health-related data. With health-related data, there is always a concern about privacy and confidentiality as informed consent. There were no names or identifying information linked to individual cases which reduces the confidentiality concerns with this data. I am also concerned about the discrimination of individuals labeled as obese by this model. Another concern is the data is 77% synthetic, this can cause misrepresentation if the synthetic data does not accurately represent real-world data.

References

Centers for Disease Control and Prevention. (2022, May 17). *Adult obesity facts*. Centers for Disease Control and Prevention. https://www.cdc.gov/obesity/data/adult.html

Mehrparvar, F. (2024, April 7). *Obesity levels*. Kaggle. https://www.kaggle.com/datasets/fatemehmehrparvar/obesity-levels

World Health Organization. (2024, March 1). *Obesity and overweight*. World Health Organization. https://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight#:~:text=Overview,the%20risk%20of%20certain%20cancers.

Appendix

A. Questions

1. Why did the data contain synthetic data?
    i. The data likely contained synthetic data because there wasn't enough internet data to perform a large analysis.
2. When you say risk assessment quiz what do you mean?
    i. By using the data and having people pick their lifestyle habits, it can determine which weight category they are likely to fall into now or in the future
3. How did you know height and weight were causing the model to be overfit?
    i. I checked feature importance, and I had a hunch because that is what is used to calculate BMI.
4. How would you go about getting additional real-world data?
    i. The most likely source would be to conduct an anonymous survey.
5. Why did you combine the obesity variables into one?
    i. For this study, I did not feel as though all of the additional weight categories were necessary.
6. Which model did you think would perform the best?
    i. I thought KNN would perform the best
7. Why did you encode the variables?
    i. This is necessary to ensure the optimal performance of models
8. You mention privacy concerns as a possible ethics issue. Can you elaborate on why you are and are not concerned?

i. I am not concerned because there is no identifying information in any of the data. In addition, 77% of the data is synthetic and not obtained from health records or survey data.

9. Why did you remove height and weight and not just weight?
    i. Both played a role in BMI, and I felt after removing weight the model was still overfit.
10. Can you explain why there were more cases of obesity correctly identified in the confusion matrix?
    i. This is likely due to the fact that there were more cases of obesity located in the dataset.

B. Encoded Variables

```
Encoded values in 'Gender' column:
[0 1]
Original values:
['Female' 'Male']

Encoded values in 'CALC' column:
[3 2 1 0]
Original values:
['no' 'Sometimes' 'Frequently' 'Always']

Encoded values in 'FAVC' column:
[0 1]
Original values:
['no' 'yes']

Encoded values in 'SCC' column:
[0 1]
Original values:
['no' 'yes']

Encoded values in 'SMOKE' column:
[0 1]
Original values:
['no' 'yes']

Encoded values in 'family_history_with_overweight' column:
[1 0]
Original values:
```

```
['yes' 'no']

Encoded values in 'CAEC' column:
[2 1 0 3]
Original values:
['Sometimes' 'Frequently' 'Always' 'no']

Encoded values in 'MTRANS' column:
[3 4 0 2 1]
Original values:
['Public_Transportation' 'Walking' 'Automobile' 'Motorbike' 'Bike']

Encoded values in 'NObeyesdad' column:
[1 3 2 0]
Original values:
['Normal_Weight' 'Overweight' 'Obesity' 'Insufficient_Weight']
```