Obesity Levels in Mexico, Peru and Columbia

CS982: Big Data Fundamentals

Word Count: 2,748 words

Contents Page

List of Tables and Figures

1. Introduction

The World Health Organisation (WHO) defines obesity as; someone who's Body Mass Index (BMI) is greater than or equal to thirty (WHO, 2023a). BMI is calculated by taking an individual's weight (kg) and dividing it by their height (m) squared (WHO, 2023b). Caused by an energy imbalance between calories consumed and calories expended, obesity is typically associated with an increase in energy dense, high fat and sugar foods and a decrease in physical activity (WHO, 2023b). In 2016, around 13% of adults globally were obese. Being overweight and obese can have a major negative effect on an individual's health with an increased risk for chronic diseases including; cardiovascular diseases, diabetes and kidney disease (WHO and FAO, 2004). It is imperative to study the causes of obesity in order to combat and reduce the increased risk to health.

With obesity most prevalent in states and communities with lower average income and where there is less access to highly nutritional foods, obesity can be onset by societal and environmental factors, such as: agriculture, transport, distribution among many others (WHO, 2023c). Approximately 57% of the population in Latin America are overweight and 19% are obese (Garcia-Garcia, 2022) . This report analyses data from a survey across Mexico, Peru and Colombia, detailing various lifestyle and dietary factors that may affect obesity of an individual. Through analysing potential factors that relate to obesity and employing machine learning methods to categorised obesity levels, this analysis will provide results that can be used to combat and reduce levels of obesity.

2. Dataset:

2.1 Source

We selected our data to investigate the potential factors that contribute to obesity. Palecho and de la Manotans (2019) compiled data based on respondents to a survey on obesity in Latin America. The dataset is freely available on the UC Irvine Machine Learning Repository Website. Their data was collected from an online survey to identify respondents' physical health status. Questions involved demographic information on the respondents including gender, age, height, weight and family history with obesity. Further data involved respondents' eating habits; Frequency of high calorific food intake, vegetable intake, number of meals per day and if they snacked between meals. The final set of data was collected in regard to respondents' general health involving; whether they smoked, drank alcohol, daily water intake, most common form of transport used, time spent on technology devices and how frequently they exercised. This data allows for an intense analysis of obesity levels and their relationship to eating habits and physical conditions. Also allowing for categorising obesity levels based on genetic and lifestyle factors.

Only 23% of the data was collected from respondents to the survey and the following 77% was generated synthetically through the Weka tool and the SMOTE filter. The authors preprocessed their data before publishing, adding a variable holding respondents calculated BMI and sorting them into seven categories based on their obesity levels, in relation to the WHO and Mexican Normativity, these are categorised as follows:

Underweight: BMI less than 18.5

Normal: BMI 18.5 to 24.9

Overweight: BMI 25.0 to 29.9

Obesity I: BMI 30.0 to 34.9

Obesity II: BMI 35.0 to 39.9

Obesity III: BMI higher than 40

2.2 Dataset Challenges

The dataset provided statistics with those of ‘normal’ and ‘insufficient weights’. We wanted to focus on overweight and obese individuals, therefore we dropped respondents with a calculated BMI as ‘normal weight’ or ‘insufficient weight’ from the dataset.

One of the biggest challenges within this dataset was the number of categorical variables. Transforming the variables from categorical to numeric made it possible to conduct different statistical analyses of variables and conduct machine learning methods. It was very time consuming work to transform manually ourselves, so we transformed these categorical categories into numerical ones using code found from statology (Statology, 2021). This assigned every value within a categorical variable a number, however there was no way to tell what each number meant apart from comparing it by eye to the dataset before it was recoded. We combatted this by creating figures with categorical variables before transforming them to numeric values. We also self recoded our target variable (obesity level) to ensure we could pull the correct results. Our recoding resulted in the following values being assigned to each weight category;

0: Overweight Level I

1: Overweight Level II

2: Obesity Level I

3: Obesity Level II

4: Obesity Level III

This made it easy to read and extract meaningful conclusions from our results without the worry we were drawing incorrect findings. This also helped when interpreting the output and confusion matrix in our supervised methods.

3. Analysis:

We first performed preliminary analyses of our variables to understand the relationship to discover what potential factors could cause obesity.

Table 3.1: Summary statistics of demographic numeric variables

Variable	**Mean**	**Standard Deviation**	**Min**	**Max**
Age (years)	25.58	6.50	15.00	56.00
Height (cm)	1.71	0.09	1.45	1.98
Weight (kg)	97.53	21.09	53.00	173.00

Using these averages of height and weight, we can calculate that the average BMI of the respondents who were classed as overweight or obese was: 33.35, falling into the Level I Obesity category. This is coherent with figure 3.2, which indicates Obesity Level I as containing the highest number of respondents.

3.1.1 Genetic and Demographic factors

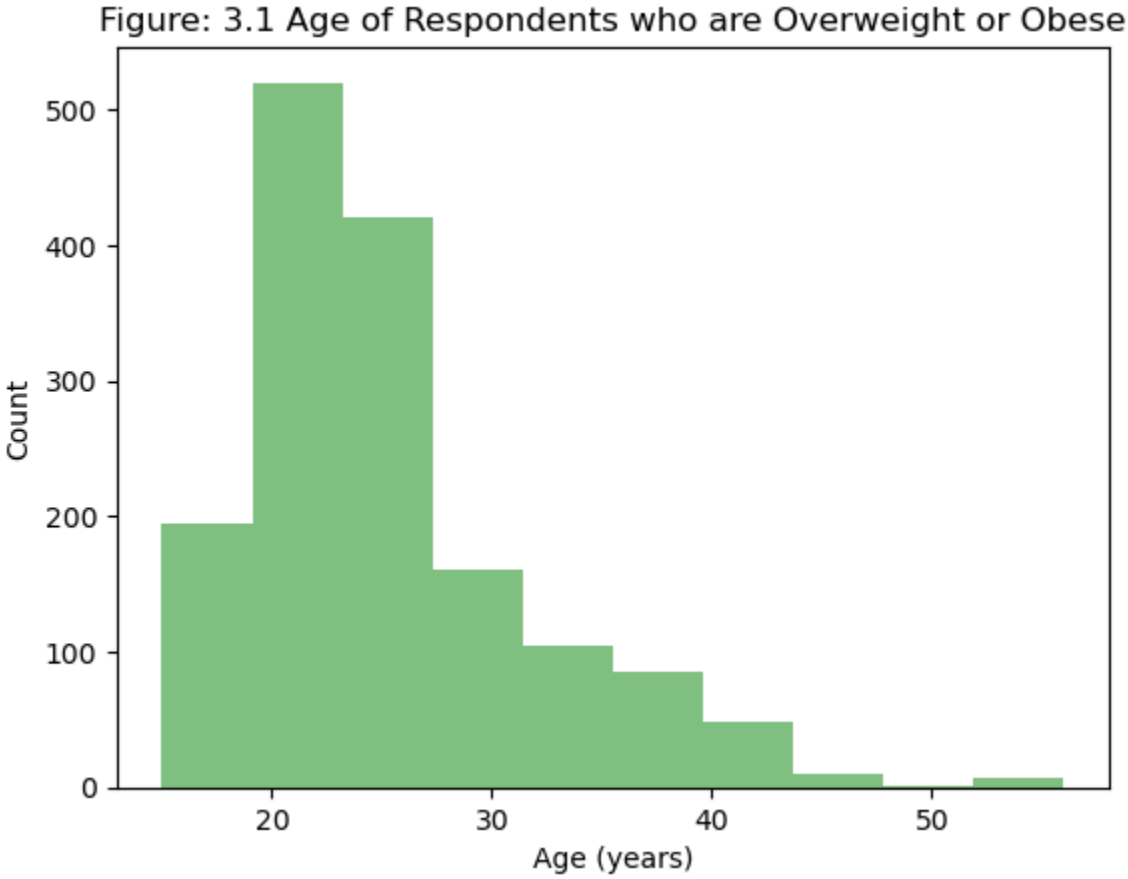
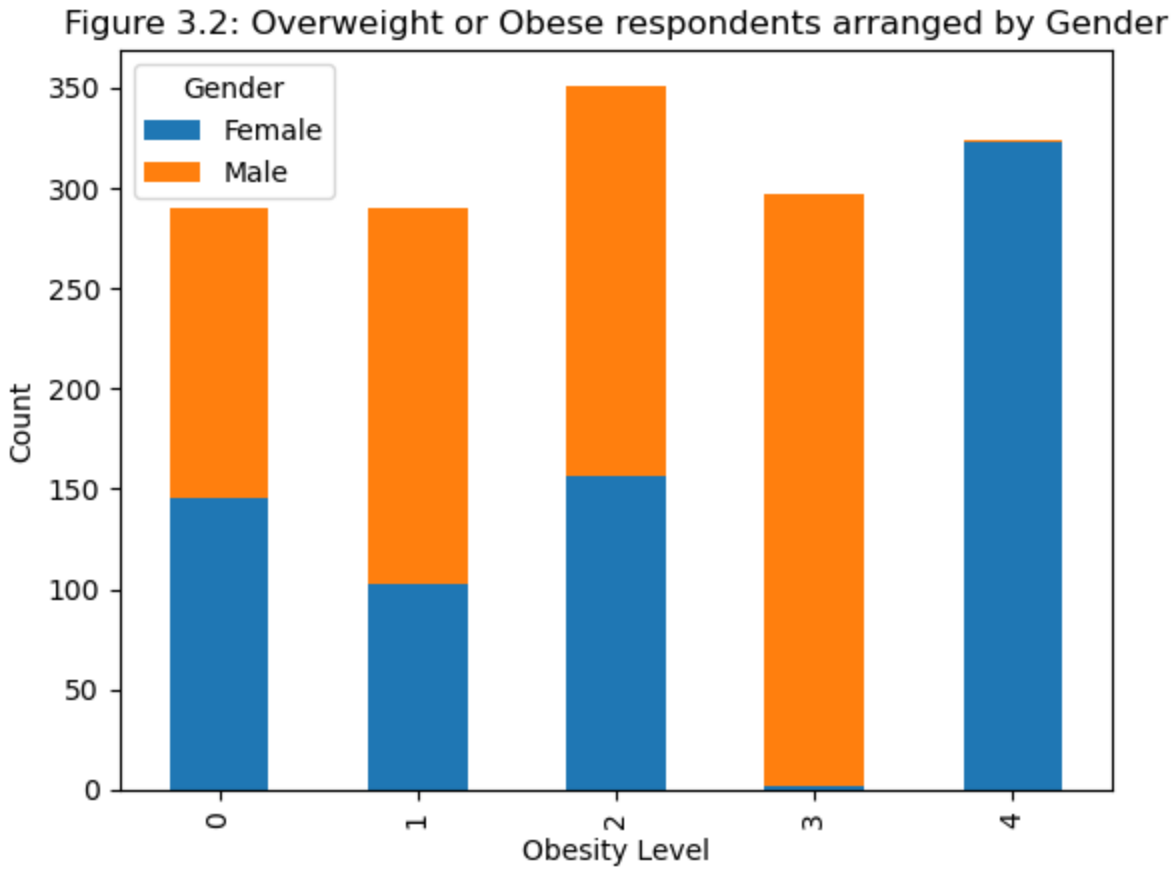
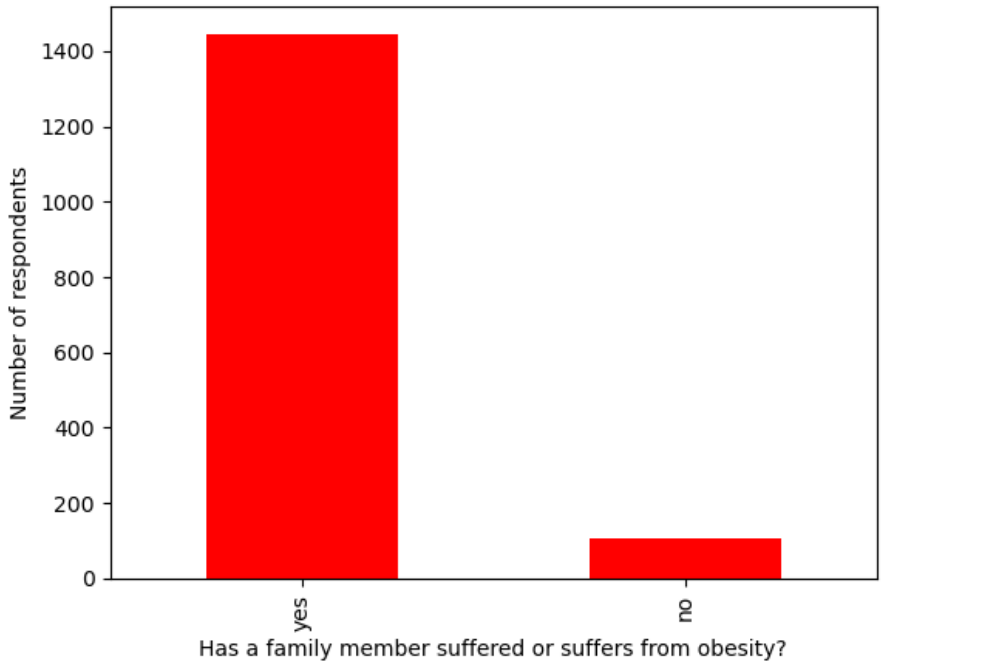


Figure 3.1 shows a histogram of the respondents who are classed as overweight or obese by a calculation of their BMI. The graph is left-skewed with the majority of respondents who are overweight or obese falling below age 30. This suggests that the younger generation are more likely to be obese or overweight, however this could be due to the lack of respondents over the age of 40.



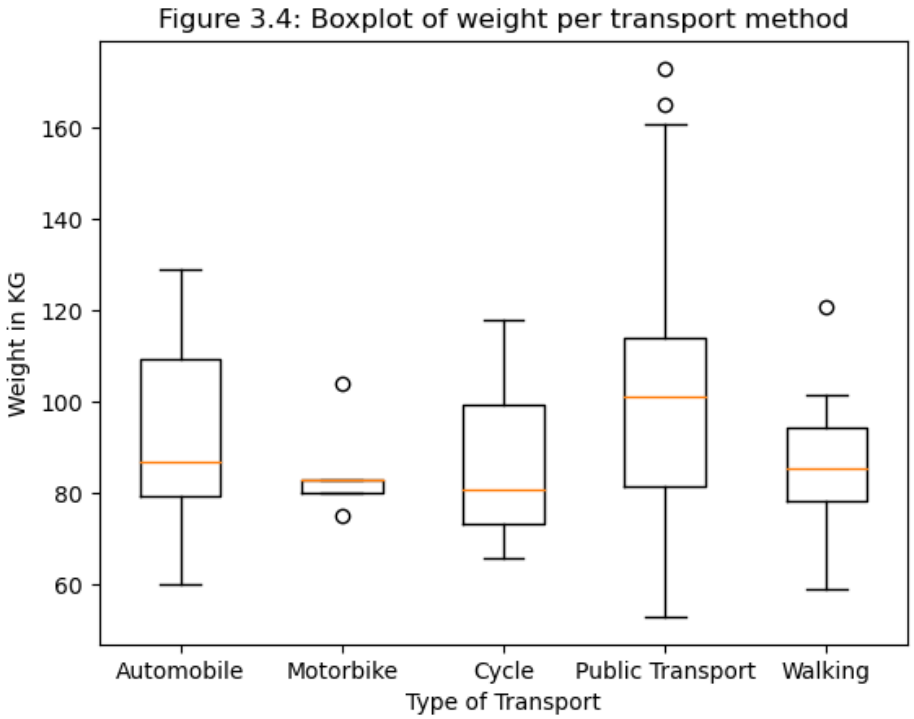
Moreover, figure 3.2 shows the distribution of male and female respondents in each level of obesity. Most respondents fell into Obesity Level I with a count of 350. The most apparent finding is that the majority of respondents in Obesity Level II, were male, in stark contrast the bulk of individuals that fell into Obesity Level III were female. This implies that women are more likely to have a larger BMI than men.

Figure 3.3: Overweight or Obese Respondents with a Family History of Obesity



Shifting away from demographics, figure 3.3 highlights the significant number of overweight or obese respondents who have a family member who was or is obese. Genetics and lifestyle are major factors to why this could play a part in obesity. Firstly, genetics that are related to obesity can be passed down to children. Moreover, growing up with an obese family member may have distorted an individual's perception of a healthy diet. This alongside a lack of access to healthy food could explain these results.

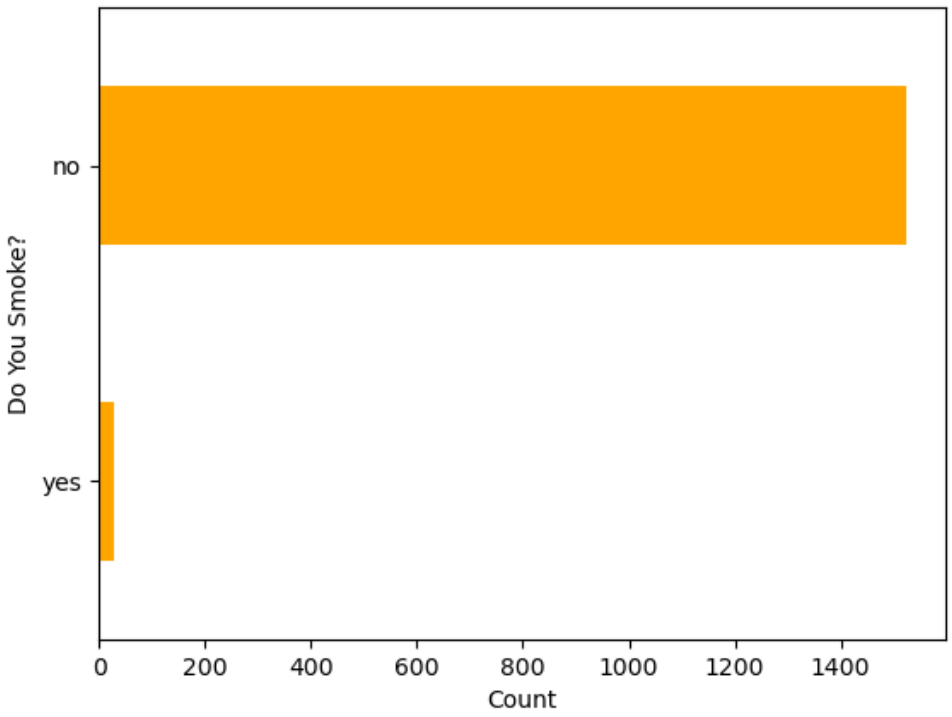
3.1.2 Dietary and Lifestyle factors



In figure 3.4 we can see that the mode of transport that has the largest mean weight is public transport. This makes sense as this is a less physically active method of transport from some of the others, such as cycling. The lowest mean weight was cyclists, this is logical per the data as it is the most energy burning mode of transport. Perhaps in order to combat this infrastructure to encourage cycling more frequently should be put in place.

Figure 3.5 displays the amount of obese and overweight respondents who do and do not smoke. As can be seen from the bar graph that a large majority of the obese and overweight respondents do not smoke. So whilst smoking is an unhealthy lifestyle choice it does not necessarily contribute to obesity in Latin America based on these findings.

Figue 3.5 Bar Chart of Whether or Not Obese Respondents Smoke



Figue 3.6 Bar Chart of Whether or Not Obese Respondents Regularly Consume High Calorie Foods

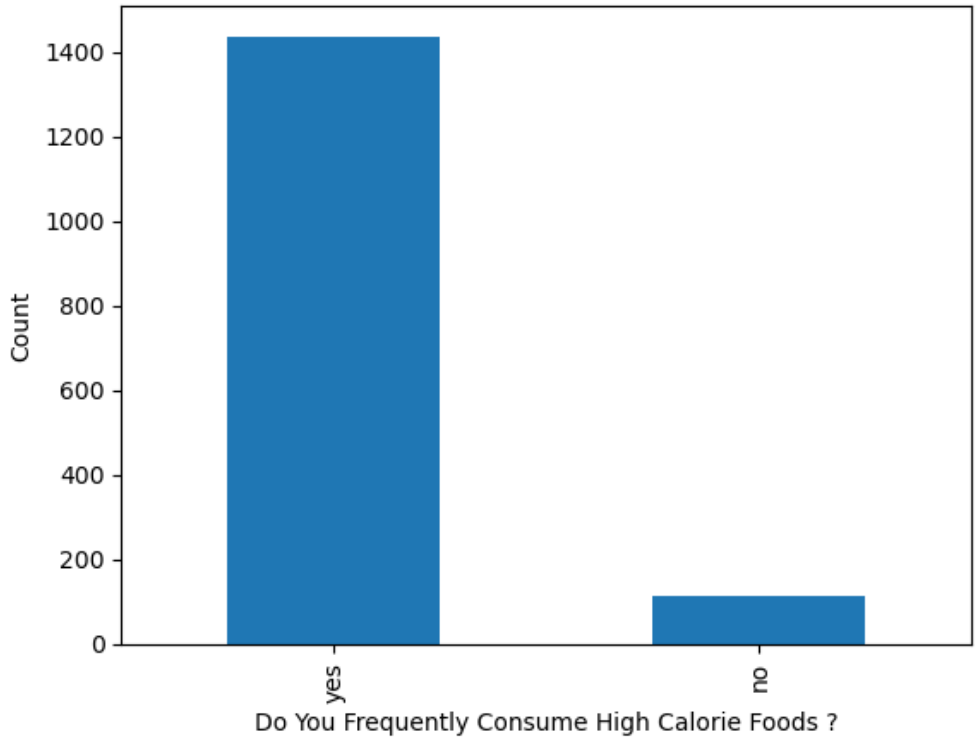
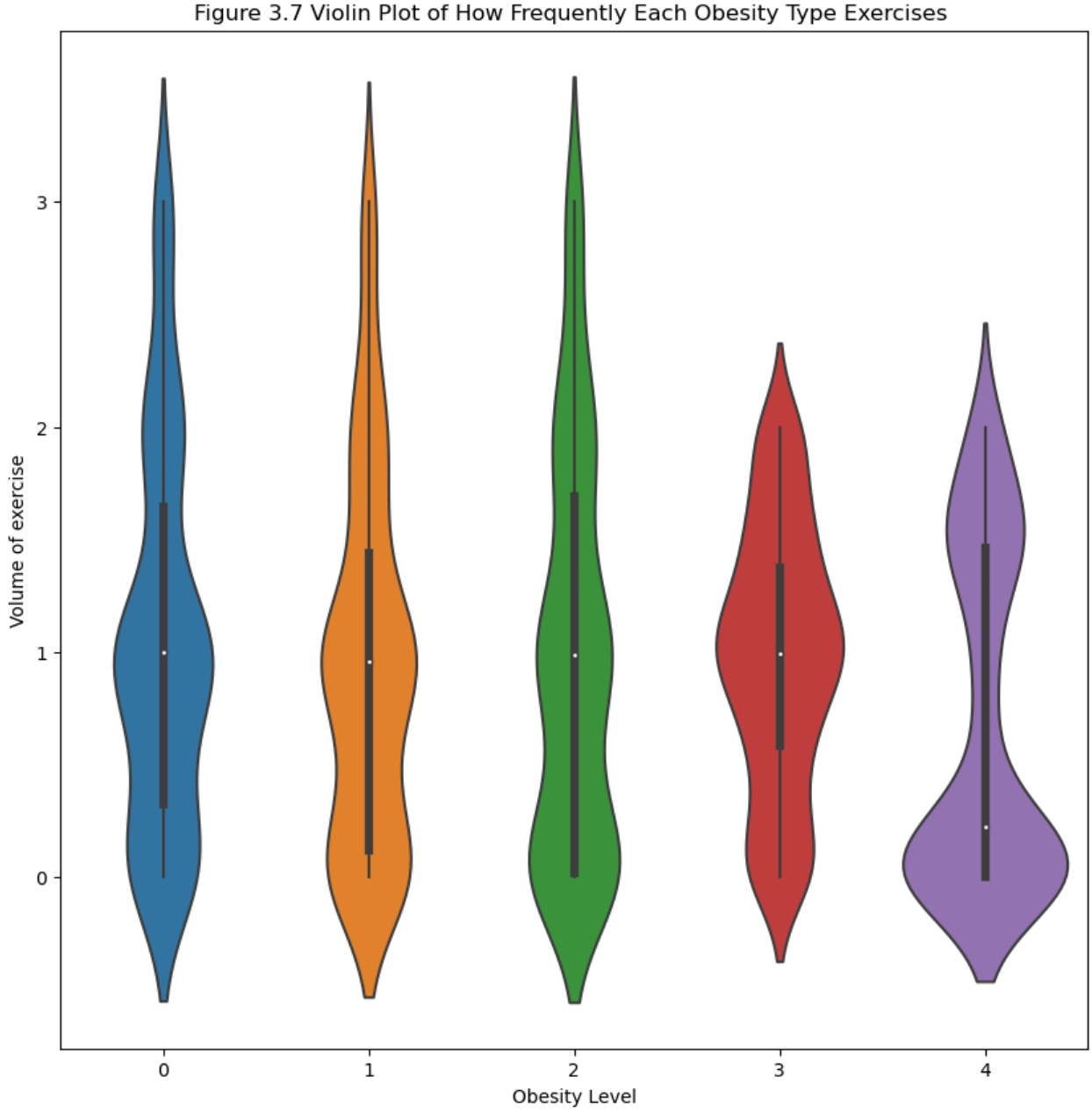


Figure 3.6 displays whether or not obese and overweight respondents consume food that is high in calories. A large proportion of the respondents to this question did, this is logical as the more frequently an individual consumes high calorie foods the more likely their BMI will be in the overweight or obese categories. Those who responded ‘no’ perhaps have other factors that may lead to them having a high BMI, such as lack of physical exercise.

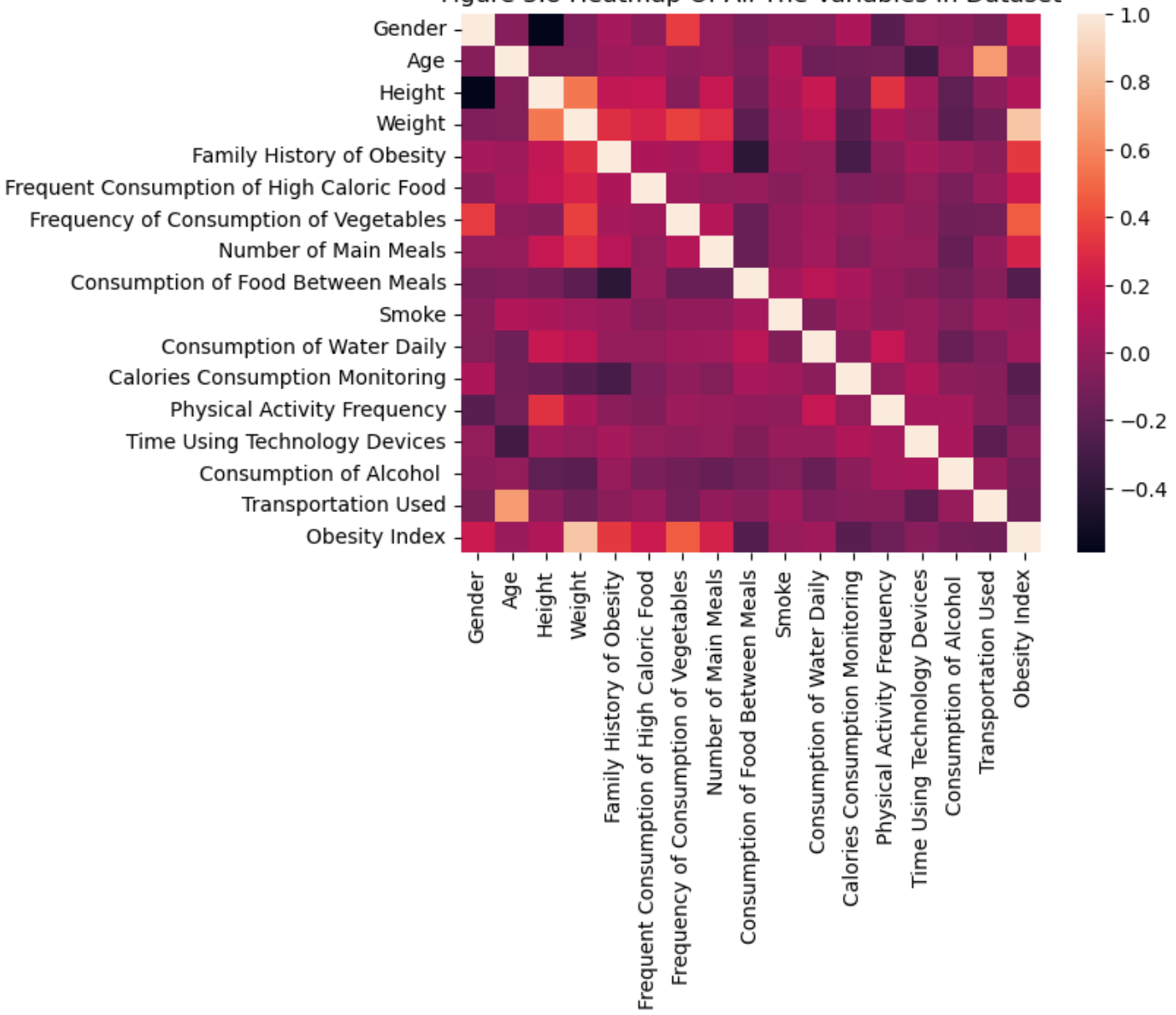
Figure 3.7 Violin Plot of How Frequently Each Obesity Type Exercises



In figure 3.7 we can see the different types of obesity and how often they exercise. We can clearly see that the higher the category of obesity a respondent is in the less frequently they engage in physical exercise. In the question on physical exercise respondents were asked on a four point scale how often they exercised, with 0 being not at all and 3 being 4 to 5 days per week. Obesity type III is the most unhealthy category of obesity and not a single respondent engaged in the highest category of frequency of physical exercise. The two lowest categories of BMI, Overweight levels I and II both have the largest distribution and the most respondents who engaged in the highest category of physical exercise. This makes logical sense as those with higher BMI would be engaging in physical exercise less frequently.

Overall in regards to lifestyle and dietary factors there is clear correlation between some of the factors explored and obesity. The heatmap in figure 3.8 highlights some peculiar results insinuating that the factors that we analysed are not necessarily associated with higher obesity. As we purely analysed obese or overweight respondents, this graph does not provide evidence of what factors are positively correlated with obesity.

Figure 3.8 Heatmap Of All The Variables In Dataset



4. Unsupervised Methods - Clustering

Unsupervised methods are the result of algorithms analysing, clustering and discovering patterns in unlabeled data without any instruction. Clustering is a form of unsupervised learning in which data is grouped together based on their similarities (Delua, 2021). In this paper we will employ agglomerative clustering to perform our analysis as our data was well suited to clustering due to its categorical values.

4.1 Agglomerative Clustering

Agglomerative clustering is a method of hierarchical clustering in which the system groups observations in the data together by their similarities. How similar the data points are to each other is determined by calculating the **distance** between them. There are four types of distance measures commonly used in clustering: Cosine, Manhattan, Hamming and Euclidean. We employed all measures of distance to see if there were differences between them and discover the best metric for clustering. Moreover, the closeness of the data is determined by the measure of **linkage** between the two data points, we will utilise the average measure of distance between data points as this provides a robust and standard approach across all metrics.

Determining how accurate the clustering is determined by three metrics:

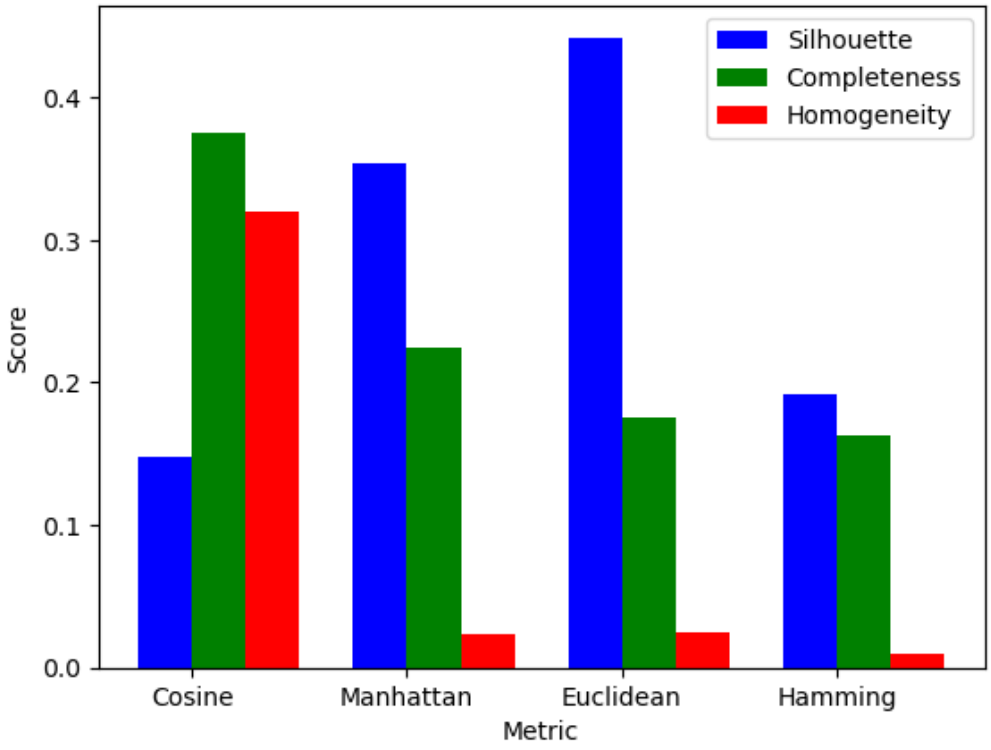
Silhouette scores determine how similar a data point is to its own cluster compared to the others, a high value determines appropriateness and a well matched cluster.

Completeness scores measure whether data of a specific class is all within the same cluster or if it is spread over multiple clusters. A high score implies that all data of the same class is within one cluster.

Homogeneity scores evaluate whether data within clusters are of the same class. A high score indicates that the data within a cluster are all of the same class.

The results are shown in Figure 4.1. No metric performed extremely well with all scores being less than 0.5. Despite performing well on completion and homogeneity scores, the cosine metric produced the lowest silhouette score. Therefore, it seems that the euclidean metric performed the best overall, with a high silhouette score and average completeness and homogeneity scores. This suggests that data was similar to other data within its cluster, but the class of the data within the cluster was not the same.

Figure 4.1 Silhouette, Completeness and Homogeneity Scores for Each Metric



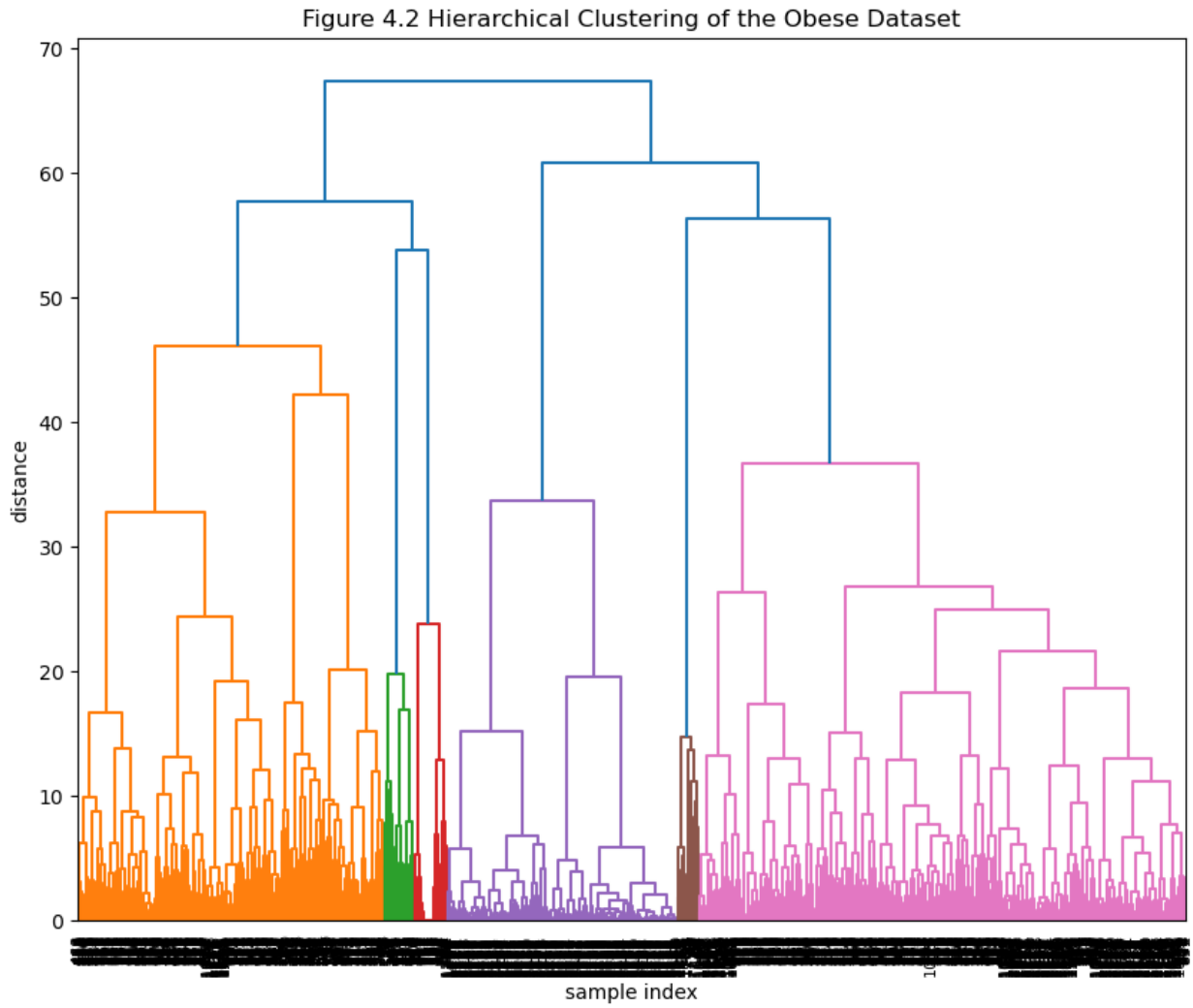


Figure 4.2 shows a dendrogram of clusters. Despite there only being five categories within our target variable, we can see that there are six distinct clusters within the graph, with the blue cluster representing the entire dataset. This could be due to the method of agglomerative hierarchical clustering employed, as it simply takes the closest data points and clusters them together, creating clusters that aren't a direct result of one of the categories in our target variable.

5. Supervised methods

Supervised learning involves training a system to learn from labelled data to produce predictions (IBM, 2023). We will use decision trees to categorise levels of obesity.

5.1 Decision Tree Model

Decision Trees are models of supervised learning that allow the classification or prediction of data. There are two types of decision tree models, classification trees and regression trees. In this case, we will utilise the classification method. Classification Decision Tree modelling was well suited to our data as it allowed for the data to be separated into distinct categories which fitted well to our target variable.

Similar to the process of human decision making, decision trees contain root nodes, branches and leaf nodes. Root nodes hold an attribute, which branches show the decisions or rules and the leafs show the result of the decision (Patel and Prajapati 2018). The model categorises each data point by memorising which category it was placed into in the training set and using this to base its decision off of. We ran the model five times to improve its learning.

Table 5.1: Results of Decision Tree Classifier

Obesity Level	**Precision**	**Recall**	**f1**
0.0	0.95	0.99	0.97
1.0	0.90	0.91	0.90
2.0	0.94	0.93	0.94
3.0	0.98	0.95	0.96
4.0	1.00	0.99	1.00

Our Decision Tree Classifier produced an accuracy of 95%. Table 5.1 shows the results of the model.

Precision calculates how many of the system's guesses were actually correct, we can see that the model produced perfect precision for those in the Obesity Level III category. **Recall** is graded on how accurate the model is at identifying true positives, we can see that again Obesity Level III produced the highest figure and Obesity Type I produced the lowest recall score (0.93) . However, this is still a hugely accurate score and thus an extremely accurate model. **F1 score** is the weighted average of precision and recall, again the model performed well for assigning each obesity level.

Figure 5.1 Visualisation of The Supervised Decision Tree Model

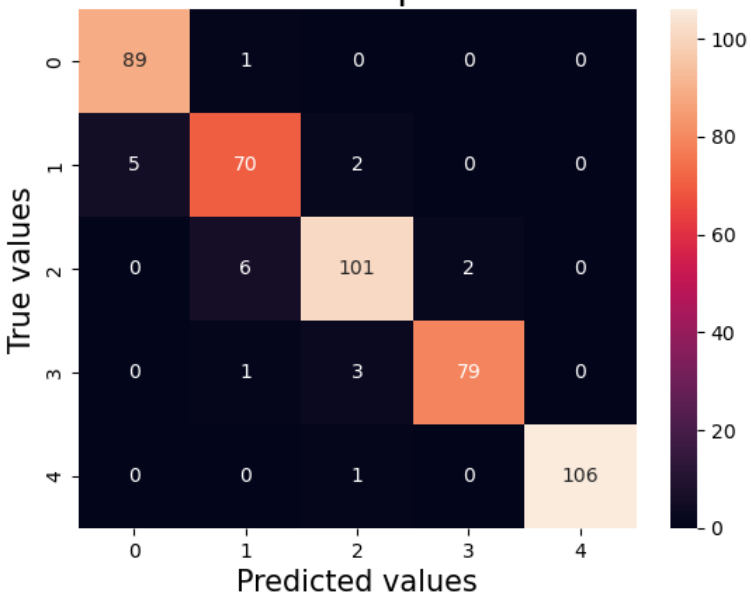


Figure 5.1 shows a confusion matrix of our decision tree model. We can see that the model wrongly categorised some values in all categories. The model found Obesity Level III the easiest to identify. Whilst struggling the most to categorise Obesity Level I correctly.

Reflection

Upon completing this paper, the selection of a dataset with a high number of categorical variables made it difficult to perform advanced statistical analysis. Despite transforming these to numerical values, not knowing which category became which value made it difficult to comprehend what our results showed. We were able to transform our dependent variable manually which made interpreting our findings easier, however if applied for all variables, we could have conducted more statistical analysis and found it easier to interpret results.

Restricting our analysis to solely focus on respondents who were overweight and obese allowed us to view correlations between variables, however did not allow a comparison of those who were of a normal weight or insufficient weight. For example in figure 3.8 unhealthy eating and exercise habits were negatively correlated with obesity levels. Perhaps if we had included all weights we would have returned results that showed which factors affect obesity levels.

BMI is not a wholly accurate measure. Gender, ethnicity and age all impact the calculation of an individual's BMI. Therefore, focusing on a generic BMI calculation for all respondents is problematic. This could have indicated any conclusions as invalid due to the inaccurate calculation of BMI and not providing a unique individual calculation based on a respondents age, ethnicity and gender. Perhaps creating a new category of individual calculation of a person's BMI would have honed more accurate results.

Finally, as 77% of data was synthetically constructed through the use of the Decision Tree estimation method, this may indicate why the Decision Tree produced high levels of accuracy. Using another

supervised machine learning method would have been interesting to investigate if it produced a more accurate result.

Conclusion

Our analysis of different factors contributing to the level of obesity in Latin America has provided some findings however they do not contribute much. We found some interesting results in our preliminary analysis, however these did not match the strong positive correlations we expected to find in figure 3.8.

As the dataset was made largely using supervised machine learning methods it worked effectively in categorising values. Conversely, unsupervised methods had fairly low returns in terms of the various scores of each pathway with the most effective returns coming from Euclidean.

In conclusion, while there have been some correlations found there is nothing groundbreaking to add to this field of study, every correlation both positive and negative have been found previously in the study of obesity. In order to combat this the education of consumers, the infrastructure to encourage positive lifestyle changes and food suppliers reducing prices on nutritious low calorie foods must all be put in place.

Appendix

Environment

Language: Python 3.22.6

Dataset from:

Palechor, F.M. and 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, 25, p.104344.

https://archive.ics.uci.edu/dataset/544/estimation+of+obesity+levels+based+on+eating+habits+and+physical+condition

Packages used:

- pandas
- numpy
- Matplotlib.pyplot
- seaborn
- sklearn
- cluster from sklearn
- metrics from sklearn
- scale from sklearn.preprocessing
- model_selection from sklearn
- DecisionTreeClassifier from sklearn.tree

Bibliography

Garcia-Garcia, G. (2022). Obesity and overweight populations in Latin America. *The Lancet Kidney Campaign*.

How to convert categorical variable to numeric in Pandas (2021) *Statology*. Available at: https://www.statology.org/convert-categorical-variable-to-numeric-pandas/ (Accessed: 16 Oct 2023).

IBM (2023) *What is Supervised Learning?* International Business Machines. Available at: https://www.ibm.com/topics/supervised-learning (Accessed: 24 Oct 2023)

Joint WHO/FAO Expert (2003). *Diet, Nutrition and the Prevention of Chronic Diseases, Report of a Joint WHO and FAO Expert Consultation.* Geneva: World Health Organisation. Available at: http://health.euroafrica.org/books/dietnutritionwho.pdf (Accessed: 22 Oct 2023).

Popkin, B. M., and Reardon, T. (2018) Obesity and the food system transformation in Latin America. *Obesity Reviews*, 19: 1028–1064

Palechor, F.M. and 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, 25, p.104344.

Revilla,M and Höhne,J,K (2020) Comparing the participation of Millennials and older age cohorts in the CROss-National Online Survey panel and the German Internet. *Survey Research Methods*, 14: 499-513

WHO (2023a) Obesity. World Health Organisation. Available at: https://www.who.int/health-topics/obesity/#tab=tab_1 (Accessed: 22 Oct 2023)

WHO (2023b). *Body Mass Index (BMI)*. World Health Organisation. Available at: https://www.who.int/data/gho/data/themes/topics/topic-details/GHO/body-mass-index (Accessed: 22 Oct 2023)

WHO (2023c). *Obesity and Overweight.* World Health Organisation. Available at: https://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight (Accessed: 22 Oct 2023)