# Understanding the relationship between adult lifestyle factors and risk of obesity

By: Sashini Kosgoda, Jinyu Luo, Zhi Lin Zhou

## Introduction

The goal of our project was to understand the relationship between lifestyle choices of adults and their risk of obesity. More precisely, we wanted to see if individual lifestyle factors can predict BMI. Previous research found that age, gender and dietary habits were important indicators of BMI (Platikanova et al., 2022). Additionally, people between the ages of 51 and 69 had higher BMI and, men had higher BMI compared to women (Platikanova et al., 2022). If individual lifestyle factors can indeed predict BMI levels, our findings can guide decision-makers in developing efficient surveillance and intervention strategies to prevent obesity.

## Data Engineering Process

We began the data engineering process by checking for missing values and encoding the 'Gender' variable. The exploratory data analysis revealed that the columns for fast food frequency and screen time were ordinal while others, except gender, were numerical. Although numerous outliers were observed in BMI values, we chose to retain them, postulating that they might represent unique cases or implications.

Next, two new variables were created for later classification tasks. The first variable represented BMI levels, comprising six categories: Underweight, Normal, Pre-obesity, Obese class 1, 2, and 3 (WHO, 2010). While 'Normal' and 'Pre-obesity' had a significant number of individuals, other categories are notably underrepresented. This imbalance could potentially affect subsequent analyses or model performances. The second variable was a binary indicator of obesity, categorizing individuals with a BMI below 25 as 'no risk' and those above this threshold as 'high risk'. The population was evenly distributed between these two groups.

## Analysis

KNN was chosen for our diverse dataset since it did not assume any specific data distribution. Using GridSearch, the hyperparameter tuning process was automated and found it superior to the elbow method. The elbow method displayed a downward trend that plateaued at n = 600 with no ‘bend’ and a comparable accuracy to n=15 chosen by GridSearch. Notably, GridSearch favored the Manhattan distance over the Euclidean, which made sense given our varied feature types and the method’s reduced sensitivity to outlier features. Despite optimizing the KNN classifier using GridSearch and stratifying BMI levels to both two and six categories, the performance remained unsatisfactory. The average precision across six BMI levels was 16.3% with a 20% weighted accuracy, while the obesity risk classification accuracy achieved around 50%. Such results suggested that the KNN model did not capture the intricate relationships and patterns present in the data. This necessitated exploring an alternative approach.

Then, a Decision Tree model was chosen. In contrast to KNN, the Decision Tree was a structured and hierarchical method that made decisions based on explicit rules. It automatically highlighted significant variables and their interactions, which provided a clearer understanding of the decision-making process. With our dataset, a decision tree could discern patterns linked to each gender without requiring separate models. If gender was influential in determining BMI levels, the tree would use it as a primary decision node. Using a single decision tree for the entire dataset was not only computationally

efficient but also ensured that any interaction between gender and other features was captured effectively.

**Findings**

Using KNN modelling we found that the precision, recall and F1 score was higher when only two categories were used for BMI rather than six categories (**Table 1**). The accuracy of the model with two BMI categories was 0.52 compared to 0.18 for the model with six categories. The decision tree model produced a higher precision, F1 score and accuracy of 0.74. Looking at the correlation matrix of all the variables analyzed we found that all correlations were quite weak, suggesting that there were no strong linear relationships between these variables.

**Conclusion**

In our study, we first fitted multiple KNN models and utilized a variety of techniques to tune hyper parameters and address imbalance issues. However, all KNN model variations produced poor performance. The decision tree model with six BMI categories produced the best results with an accuracy of 74%. This led us to conclude that first, the relationship between BMI and lifestyle factors were not linear. Second, the lifestyle variables we selected were adequate in predicting BMI class. BMI and obesity are quite complex health conditions in which a variety of genetic and lifestyle factors play a part. Further research should focus on incorporating other lifestyle and genetic factors to more accurately predict BMI.

**Individual Contributions:** All team members contributed greatly to the coding and write up process.

- **Sashini:** KNN modeling, SMOTE utilization and figure creation.
- **Jinyu:** EDA, data engineering, hyperparameter tuning, model fitting, analysis.
- **Zhi Lin:** KNN modeling, data cleaning and google slide creation.

**Google slides link:** https://docs.google.com/presentation/d/e/2PACX-1vTLF4PKtYQcmqs1SnoYdkyyRJCV90RwV0LMoV1qK9r8GRhcFQGMMtBQ-XcKM7Id3feehNnLbcXnk04N/pub?start=true&loop=true&delayms=60000

**Github Repository:** https://github.com/Jinyu-Luo/CHL5230-Datathon1.git

## Appendix

**Table 1.** Model performance of different models. All the classifiers were fitted with data that dropped Height, Weight, and BMI.

| Model | Description | BMI | Accuracy | Precision | Recall | F1-score |
|---|---|---|---|---|---|---|
| 1 | **KNN model**<br>Class Label: BMI Levels<br>Metric: Manhattan distance<br>5 neighbors | 0 – Underweight | 0.18 | 0.11 | 0.14 | 0.13 |
| | | 1 – Normal | | 0.33 | 0.21 | 0.26 |
| | | 2 – Pre-obesity | | 0.26 | 0.18 | 0.22 |
| | | 3 – Obesity Class1 | | 0.16 | 0.16 | 0.16 |
| | | 4 – Obesity Class2 | | 0.08 | 0.18 | 0.11 |
| | | 5 – Obesity Class3 | | 0.04 | 0.15 | 0.06 |
| 2 | **KNN model**<br>Class Label: Obesity<br>Metric: Euclidean distance<br>75 neighbors | 0 – Non-obese | 0.52 | 0.47 | 0.25 | 0.33 |
| | | 1 – Obese | | 0.54 | 0.75 | 0.63 |
| 3 | **Decision tree**<br>Class Label: BMI Levels | 0 – Underweight | 0.74 | 0.85 | 0.99 | 0.91 |
| | | 1 – Normal | | 0.73 | 0.80 | 0.76 |
| | | 2 – Pre-obesity | | 0.47 | 0.31 | 0.37 |
| | | 3 – Obesity Class1 | | 0.76 | 0.90 | 0.82 |
| | | 4 – Obesity Class2 | | 0.90 | 1.00 | 0.95 |
| | | 5 – Obesity Class3 | | 0.57 | 0.46 | 0.51 |

## Reference

Blagus, R., Lusa, L. SMOTE for high-dimensional class-imbalanced data. BMC Bioinformatics 14, 106 (2013). https://doi.org/10.1186/1471-2105-14-106

Ferdowsy, F., Rahi, K. S. A., Jabiullah, M. I., & Habib, M. T. (2021). A machine learning approach for obesity risk prediction. Current Research in Behavioral Sciences, 2, 100053. https://doi.org/10.1016/j.crbeha.2021.100053

Platikanova, M., Yordanova, A., & Hristova, P. (2022). Dependence of Body Mass Index on Some Dietary Habits: An Application of Classification and Regression Tree. Iranian Journal of Public Health, 51(6), 1283–1294. https://doi.org/10.18502/ijph.v51i6.9672

World Health Organization. (2010, May 6). *A healthy lifestyle - who recommendations*. World Health Organization. https://www.who.int/europe/news-room/fact-sheets/item/a-healthy-lifestyle---who-recommendations