Obesity Prediction based on lifestyle and demographic factors - Milestone 1

Weiqi Zhou
UC San Diego
wez092@ucsd.edu

Feiyang Jiang
UC San Diego
fejiang@ucsd.edu

Junhan Chen
UC San Diego
juc102@ucsd.edu

Fanyuanhang Zhang
UC San Diego
faz007@ucsd.edu

Yiqian Liu
UC San Diego
yil381@ucsd.edu

1 Problem Description and Motivation

We study the problem of predicting an individual's obesity category using demographic and lifestyle features. Obesity is a major public-health concern linked to increased risk of chronic diseases and reduced quality of life. Risk factors such as diet, physical activity, sedentary behavior, and family history interact in complex, probabilistic ways rather than simple linear relationships. Our goal is to build a probabilistic model that can not only predict obesity class, but also reveal how different factors contribute to obesity risk and support "what-if" reasoning (e.g., how changes in activity level affect risk).

2 Dataset Description

For our study, we will be using the dataset - "Dataset for estimation of obesity levels based on eating habits and physical condition in individuals from Colombia, Peru and Mexico"[Palechor and la Hoz Manotas, 2019].

This dataset include data for the estimation of obesity levels in individuals from the countries of Mexico, Peru and Colombia, based on their eating habits and physical condition. The data contains 17 attributes and 2111 records, the records are labeled with the class variable NObesity (Obesity Level), that allows classification of the data using the values of Insufficient Weight, Normal Weight, Overweight Level I, Overweight Level II, Obesity Type I, Obesity Type II and Obesity Type III.

3 Methodology

Overall Idea
We use a Bayesian network to model the joint distribution of demographic, diet, and lifestyle variables together with an intermediate BMI node and the final obesity class. All continuous features (e.g., age, BMI, physical activity level, screen time, water intake) will be discretized into a small number of bins (such as low/medium/high) so that all variables are discrete and the conditional probability tables (CPTs) remain manageable.

3.1 Variables in the Belief Network

We will include the following features from the dataset as variables in the belief network:

- Demographic & family history: Gender, Age_bin, family_history

Preprint. Under review.

- Diet-Related Factors: FAVC (high-calorie food frequency), FCVC (vegetable intake), NCP (meals per day), CAEC (snacking), CH20_bin (water intake)
- Lifestyle / Activity: SMOKE, FRAF_bin (physical activity), TUE_bin (screen time), CALC (alcohol), MTRANS (transport mode)
- Intermediate Node: BMI_bin
- Outcome: Obesity

3.2 Domain-Informed Structure

We will construct the structure of BN based on domain knowledge of dependencies among selected variables. Specifically, the structure flow will be:

- Demographic and family history $\rightarrow$ Lifestyle and dietary traits: due to (1) Age and gender influence activity pattern, calorie intake, and eating routine; and (2) Family history affects both BMI and lifestyle behaviors
- Lifestyle and diet $\rightarrow$ BMI_bin: due to eating habits (FAVC, FCVC, NCP, CAEC, CH2O_bin) and lifestyle variables (FAF_bin, TUE_bin, SMOKE, CALC, MTRANS) influence body mass.
- BMI_bin $\rightarrow$ Obesity: due to BMI serves as an intermediate physiological indicator directly determining the final obesity class.

To avoid unmanageably large CPTs, we will cap the number of parents per node. When multiple variables influence BMI, we will optionally introduce a Lifestyle_Index node that aggregates major behaviors (e.g., physical activity, screen time, calorie frequency) into a discretized summary variable. This reduces the number of parents for BMI_bin while preserving model interpretability.

3.3 Parameter Learning and Inference

For parameter learning, we plan to use maximum likelihood estimation to obtain stable conditional probability tables once the network structure has been determined from the previous step. Since our dataset is complete and fully observed, MLE allows us to directly estimate these probabilities from the data without relying on iterative procedures like EM. After the CPTs are learned, we will perform inference on the network using different inference methods for comparison of result: exact inference with variable elimination, and approximate inference methods like Markov Chain Monte Carlo. For prediction step, we will compute $P(Obesity|features)$ and select the most likely obesity category as the model's output. After prediction on the test set, we will assess the performance of the model through metrics like accuracy, F1-score, etc. In addition to performance evaluation, we will run several "what-if" queries (ex. adjusting activity level or caloric intake) to highlight the interpretability benefits of using a Bayesian network and to better understand how specific lifestyle factors influence obesity risk.

References

Fabio Mendoza Palechor and Alexis De la Hoz Manotas. Dataset for estimation of obesity levels based on eating habits and physical condition in individuals from colombia, peru and mexico. *Data in Brief*, 25, 2019. URL `https://api.semanticscholar.org/CorpusID:201195793`.