A Bayesian Network Analysis of Behavioral and Genetic Determinants of Obesity Severity

By Lashana Narayan and Christian Lopez

Introduction

Obesity emerges from complex genetic predisposition, dietary habits, and lifestyle choices, and is rarely the result of a single isolated factor. This report investigates the Estimation of Obesity Levels Based on Eating Habits and Physical Condition dataset, sourced from the UCI Machine Learning Repository. The data consist of 2,111 observations from individuals in Mexico, Peru, and Colombia. It offers a unique hybrid composition: 23% was collected via real-world web surveys, while the remaining 77% was synthetically generated using the SMOTE algorithm to ensure class balance. Since the synthetic data were statistically indistinguishable, this analysis treats the findings as a descriptive map of statistical patterns within this cohort rather than as a direct estimate of general population-level effects.

The methodological backbone of this report is a Bayesian Network (BN) approach, utilizing Hill Climb Search and the Bayesian Information Criterion (BIC) to automatically learn the most likely connections between variables. Unlike standard "black-box" machine learning models, this probabilistic framework uses Variable Elimination and PyMC parameter learning to treat probabilities as distributions rather than fixed points. This allows for a more nuanced understanding of how specific behavioral switches interact with genetic history.

This report aims to identify the critical points at which lifestyle changes most effectively mitigate or exacerbate the risk of progression to clinical obesity.

Data

The data used in this study comes from the UCI Machine Learning Repository, specifically the *Estimation of Obesity Levels Based on Eating Habits and Physical Condition* dataset. The dataset contains observations on 2,111 individuals and was designed to study the relationship between lifestyle behaviors and obesity outcomes.

Each observation includes demographic, behavioral, and physical attributes, as well as a labeled obesity category. The original dataset comprises both real-world data (23%) collected via web surveys conducted in Mexico, Peru, and Colombia, and synthetically generated data (77%) created using the SMOTE algorithm to address class imbalance.

The synthetic data are indistinguishable from the real data. Because synthetic observations are not identifiable, analyses are conducted on the full dataset with cautious interpretation (associations within the augmented dataset, not population-level effects).

Outcome Variable

The primary outcome is obesity level, derived from body mass index (BMI) and categorized in the original dataset into multiple classes. For this study, these categories were consolidated into three ordered levels:

- Normal, N = 287 (15.6%)
- Overweight, N = 580 (13.5%)
- Obese, N = 972 (52.9%)

There are 272 records for those individuals identified as having insufficient weight. These records will be excluded from the analysis. This outcome is treated as an ordinal variable, reflecting the increasing severity of obesity status.

Predictor Variables

Predictor variables fall into three broad groups:

1. Demographic variables: age, gender, height, and weight
2. Eating behavior variables: frequency of high-calorie food consumption, vegetable intake, number of daily meals, snacking behavior, alcohol consumption, and water intake
3. Lifestyle and physical activity variables: physical activity frequency, time spent using technology, transportation mode, and family history of overweight

Variables include a mix of continuous, ordinal, and categorical measures, with ordinal variables representing self-reported frequency scales.

Data Preprocessing

We transformed the raw obesity dataset into a format suitable for probabilistic modeling through four main steps.

1. Cleaned and renamed the features to be more intuitive, filtered out ‘Insufficient Weight’ to focus on weight gain progression.

2. Performed manual binning, where continuous numbers like age and physical activity frequency were categorized into meaningful categories (eg, ‘Low’, ‘Medium’, ‘High’) using specific clinical cut points.
3. Dropped height and weight due to their direct relationship to how the obesity state is determined, forcing the model to find deeper behavioral patterns.
4. Converted all features to strings to correctly calculate the conditional dependencies between lifestyle habits and obesity outcome.

Exploratory Data Analysis

This analysis explores the complex relationship between dietary habits, physical condition, and obesity levels using the *Estimation of obesity levels based on eating habits and physical condition* dataset, sourced from the UCI Machine Learning Repository (Palechor & de la Hoz Manotas, 2019). The dataset contains approximately 2,111 records and 17 features, with *NObeyesdad* serving as the categorical target variable representing seven distinct levels of weight and obesity.

Recent studies using this dataset have highlighted its utility for benchmarking machine learning algorithms, particularly for personalized health interventions (Safaei et al., 2021). A key characteristic of this dataset is the "Artificial Balance" of the target variable. Unlike real-world medical data, where "Normal" cases typically outnumber "Obese" cases, this dataset exhibits a nearly equal distribution of participants across all categories, ranging from Insufficient Weight to Obesity Type III. This balance was likely achieved using the Synthetic Minority Over-sampling Technique (SMOTE), a method shown in recent comparative analyses to significantly reduce class-imbalance bias and improve the F1-scores of predictive models such as Random Forests (Muliawan et al., 2024). However, this implies that the data do not perfectly reflect the general population's demographics.

A detailed examination of data quality reveals specific artifacts attributed to the synthetic generation process. Several features that represent integer values, such as the Number of Main Meals (NCP) and Frequency of Vegetable Consumption (FCVC), are stored as floating-point numbers (e.g., 2.3 or 1.7). This confirms the dataset was augmented using SMOTE, which interpolates between existing data points to create new samples. While these values should ideally be rounded for strict descriptive reporting, retaining them as floats preserves the synthetic variance necessary for robust machine learning training (Palechor & de la Hoz Manotas, 2019; Zheng et al., 2025).

For the demographic and lifestyle features, the age distribution is notably right-skewed, with the dominant group aged 18-26 years (See EDA Figure 1). This concentration indicates that the data primarily represent young adults, limiting its generalizability to older populations (Zheng et al., 2025). While height follows a relatively normal distribution, weight displays a multimodal distribution with multiple peaks corresponding to distinct weight classes. In terms of lifestyle, a significant majority of participants reported a family history of being overweight, which consistently emerges as a dominant predictor. Transportation analysis (MTRANS) indicates that public transportation and automobiles are the dominant modes. In contrast, walking and biking are rare, consistent with the sedentary nature often observed in higher obesity classes (Khater et al., 2024).

Bivariate analysis highlights clear clusters and drivers of obesity. Visualizing Weight vs. Height serves as a proxy for BMI, revealing a positive correlation, with the Obesity_Type_III class forming a distinct cluster characterized by high weight and average-to-high height. Although the dataset population is young, the Obesity Type II and III categories show slightly higher mean ages than the Normal or Insufficient Weight categories, suggesting a trend of weight accumulation over time (See EDA Figure 2).

Standard correlation analysis, supported by recent feature importance studies, identifies three primary drivers. First, Family History is King: Zheng et al. (2025) found that nearly all severely obese individuals in the dataset reported a positive family history, making it a critical risk marker. Second, Snacking (CAEC): A high frequency of snacking between meals is strongly associated with higher obesity prevalence. Finally, Physical Activity (FAF) shows a strong inverse correlation; higher frequency of physical activity is significantly associated with Normal and Insufficient Weight categories, a finding corroborated by Khater et al. (2024) in their optimization of Random Forest models.

The EDA confirms that, although the dataset is clean, balanced, and prepared for modeling, it remains synthetic and heavily skewed toward young adults; future modeling efforts should account for this demographic bias.

Model Selection

The goal of this study is to understand the most likely predictors of the three obesity states: (1) Normal, (2) Overweight, and (3) Obese. To do this, the analysis uses Bayesian Network (BN) libraries in Python: HillClimbSearch with the Bayesian Information Criterion (BIC) score to automatically learn the most likely connections (the 'structure') between your variables. Next, it ensures that every variable in the dataset is included in the model, even if it has no connections, to prevent errors during computation. Then, it performs 'Parameter Learning' using a BayesianEstimator to calculate the probability tables (CPDs) for each node based on your data. Finally, it uses VariableElimination to perform inference, specifically to calculate the baseline probabilities for 'obesity_level_3cat' and to save the results to a text file.

The model uses pgmpy to discover the ‘skeleton’ (the connections) and PyMC for advanced parameter learning. Using PyMC, probabilities are treated as distributions rather than fixed points, allowing for robust estimates of complex lifestyle profiles, as shown in Figure 1 below.

This model selection is appropriate because Bayesian Networks explicitly represent conditional dependencies among correlated lifestyle and genetic factors while remaining interpretable and robust to mixed data types, making them well suited for exploratory risk pathway analysis in obesity research.

Model Assumptions

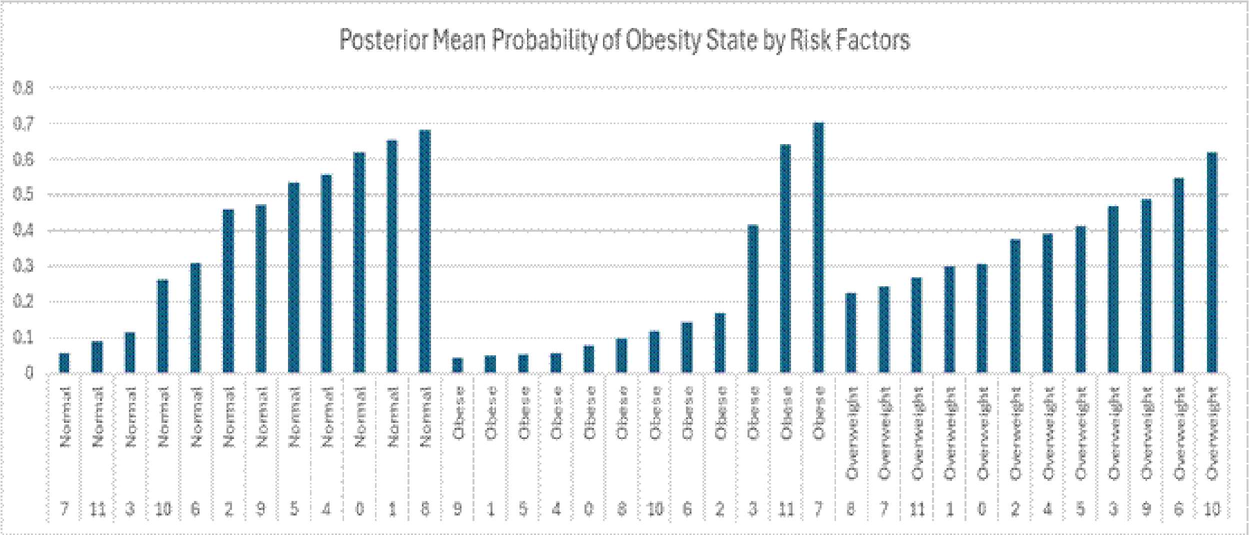
The Bayesian network (BN) model assumes that obesity severity and its associated demographic and behavioral features can be represented as a discrete, directed acyclic graph (DAG), where the joint probability distribution factorizes according to conditional dependencies among variables. All predictors were treated as categorical; continuous variables such as age and frequency-based measures were discretized into bins (e.g., Young/Mid/Senior/Elder; Low/Medium/High), which assumes that individuals within each bin are probabilistically similar and that the chosen cut points meaningfully capture variation. Height and weight were explicitly excluded from the BN because obesity status is derived from body mass index, which is a deterministic function of those variables; including them would create a trivial definitional pathway (e.g., Weight → Obesity) and obscure behavioral structure. The original multilevel obesity categories were collapsed into three ordered groups (Normal, Overweight, Obese) to stabilize conditional probability tables (CPTs), improve interpretability, and avoid sparse cells; the insufficient weight category was excluded to focus on severity progression. Structural constraints were imposed through expert knowledge: demographic variables (age, gender, family history) were treated as root nodes with no incoming edges, and obesity severity was restricted from having outgoing edges to prevent reverse-causation artifacts. Selected biologically plausible relationships (e.g., family history → obesity, diet/activity → obesity) were encouraged to ensure the learned graph reflected substantive knowledge rather than purely data-driven correlations. The model further constrained the maximum in-degree (e.g., ≤3 parents per node) to limit CPT dimensionality and reduce overfitting. Finally, because the dataset contains a substantial synthetic component (SMOTE-generated observations), the BN is interpreted as a 'descriptive map of statistical patterns' within the specific dataset, rather than definitive population-level causal mechanisms.

The model, using the config class in Python, made a strategic choice to focus on the three features that were most important: physical activity, family history, and high-calorie intake. In BN, if a node has too many parents, its probability table grows exponentially. By selecting the three most significant behavioral and genetic drivers identified during the structure-learning phase, the analysis is focused, interpretable, and computationally efficient while still capturing the most influential predictors of obesity outcomes.

Model Analysis

Based on the BN structure using the three outcomes (normal weight, overweight, and obese) and three most significant predictors, including amount of physical activity (low, medium, high), family history of overweight (yes, no), and high calorie food (yes, no), and their associated posterior probabilities were calculated to determine most likely predictors of obesity and overweight. The results for normal weight, as illustrated in Figure N, indicate that risk factors 2, 9, 5, 4, 0, 1, and 8 are likely to be associated with normal weight. This is further supported by the posterior probability of obesity for the same risk factors. For obesity, the risk factors are 3, 11, and 7 (low probability of normal weight).

Figure 1: Posterior Mean Probability of Obesity State by Risk Factors.



Note: The risk factors are the numbers below the obesity state. The risk factors are the 12 combinations of the 3 levels of physical activity, a binary indicator for family history of overweight, and a binary indicator for high-calorie food consumption. The risk factor indices and their definitions are provided in the appendix.

Further analysis of the risk factors and their posterior probability of obesity revealed three key drivers of obesity:

1. Family History - The dominant driver
2. Activity vs. Diet - There are behavioral synergies between these
3. Sedentary Lifestyle - Low Physical Activity + Family History + High Calorie Food Frequency

The 'Dominant' Driver: Family History of Overweight emerged as the most significant structural predictor. Regardless of lifestyle habits, individuals with a family history have a much higher 'baseline' probability of being in the Overweight or Obese categories. In fact, for the most high-risk profile (Low Activity + High Calorie), having a family history pushes the probability of obesity to over 70%.

Behavioral Synergies: Activity vs. Diet - Our analysis reveals a powerful interaction between physical activity and dietary habits:

- High Calorie Frequency: This acts as a 'tipping point.' Even for those with Medium' activity levels, moving from low to high-calorie food frequency shifts the highest probability mass from 'Normal' to 'Overweight.'

- Physical Activity Protection: For individuals without a family history, maintaining 'Medium' to 'High' physical activity acts as a strong buffer, keeping the probability of staying in a 'Normal' weight state between 62% and 68%.

The Sedentary Risk Profile: The 'highest risk' profile identified by the model consists of Low Physical Activity + Family History + High Calorie Food Frequency. This combination resulted in a Lift of 3.29x, indicating that individuals in this group are more than three times as likely to be obese as the average person in the dataset. For those with a family history, low physical activity is a critical failure point, leading to a 70.2% probability of obesity. Integrated programs that combine metabolic monitoring with activity are essential here.

The 'Overweight' state appears as a transitional zone. Unlike 'Normal' or 'Obese' states, which have very clear 'best' predictors, the Overweight state has high probabilities across several risk categories, as shown in Figure 1. This suggests that for many, the

'Overweight' state is highly sensitive to small behavioral changes in either direction. 'High Calorie Food Frequency' was identified as a primary 'tipping point' into the Overweight category. Behavioral therapy and apps should focus specifically on reducing the *frequency* of high-calorie meals rather than just total caloric count, as frequency showed a stronger structural dependency in our learned graph. Our analysis shows the 'Overweight' category is highly sensitive to lifestyle changes, meaning they represent the most 'preventable' cases for progressing into clinical obesity. The probability of a 'Normal' weight outcome decreases by ~13% when moving from Medium to Low activity; reaching this minimum threshold (i.e., Medium activity) offers the highest 'return on investment' for weight maintenance.

The data used in this analysis is based on a survey conducted in Peru, Colombia, and Mexico. The diets in Peru, Colombia, and Mexico can be very different from those in other Western countries. As such, the results may not generalize to other regions of the world.

Conclusion and Recommendations

This study evaluates the association between obesity status and physical and behavioral features reported in survey data collected in Peru, Colombia, and Mexico. The Bayesian Network analysis reveals that obesity risk is not driven by a single isolated factor but by the complex interplay between genetic predisposition (ie, family history of overweight) and specific behavioral choices. Family history remains the most powerful structural predictor in the network, significantly shifting the baseline risk toward overweight and obese states regardless of other behaviors. Physical activity and high-calorie food consumption act as the primary switches that can either mitigate or exacerbate this genetic risk. Our model identified a sweet spot at the medium physical activity level, which provided the highest probability of maintaining a normal weight status for the general population, suggesting that consistency is more impactful than peak intensity.

Public health interventions should prioritize individuals with a known family history of overweight. They are also most sensitive to sedentary risk. For general wellness messaging, the 'Medium' activity threshold (0.5–2.0 sessions per week) should be the primary target because it offers the highest statistical protection against weight gain. Focus nutrition counseling on reducing the frequency of high-calorie food events rather than just total calories, as frequency was the primary driver in the causal model.

Limitations

Because a majority of observations were synthetically generated using SMOTE and cannot be distinguished from the original records, the findings should be interpreted as identifying associative patterns within the augmented dataset rather than as estimating population-level effects. While the Bayesian Network effectively identifies structural relationships and behavioral archetypes in the data, the exact probability estimates (e.g., 70.2%) reflect the specific augmented dataset used to train the model, and real-world clinical probabilities may vary.

References

- Khater, A., Hibbert, D., & Silver, D. (2024). Optimization of machine learning classifiers for obesity levels estimation using physical habits and dietary data. *World Scientific News*, 198, 326–353.

- Muliawan, A., Fauziah, D. A., & Afrianto, E. (2024). Obesity risk prediction using a random forest based on eating habit parameters. *INSIDE Journal*, 2(1), 13–18.

- Palechor, F. M., & de la Hoz Manotas, A. (2019). *Estimation of obesity levels based on eating habits and physical condition* [Data set]. UCI Machine Learning Repository. https://doi.org/10.24432/C5H31Z

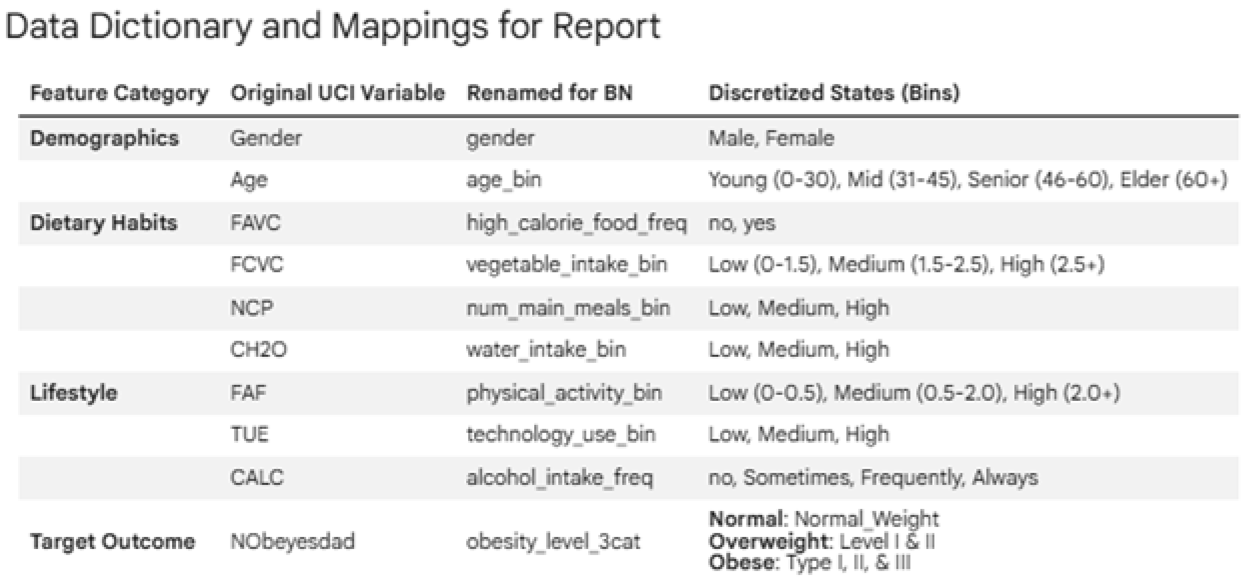
- Safaei, M., Sundararajan, E. A., Driss, M., Boulila, W., & Shapi'i, A. (2021). A systematic literature review on obesity: Understanding the causes and consequences of obesity and reviewing various machine learning approaches used to predict obesity. *Computers in Biology and Medicine*, 136, 104754. https://doi.org/10.1016/j.compbiomed.2021.104754

- Zheng, X., & Collaborators. (2025). Machine learning analysis of obesity levels based on lifestyle and physical features. *E3S Web of Conferences*, 385, 05037. https://doi.org/10.1051/e3sconf/202455305037

Appendix

The code notebook can be found on GitHub: Christian_Lashana_Github.

DATA



Data Dictionary and Mappings for Report

Feature Category	Original UCI Variable	Renamed for BN	Discretized States (Bins)
Demographics	Gender	gender	Male, Female
	Age	age_bin	Young (0-30), Mid (31-45), Senior (46-60), Elder (60+)
Dietary Habits	FAVC	high_calorie_food_freq	no, yes
	FCVC	vegetable_intake_bin	Low (0-1.5), Medium (1.5-2.5), High (2.5+)
	NCP	num_main_meals_bin	Low, Medium, High
	CH2O	water_intake_bin	Low, Medium, High
Lifestyle	FAF	physical_activity_bin	Low (0-0.5), Medium (0.5-2.0), High (2.0+)
	TUE	technology_use_bin	Low, Medium, High
	CALC	alcohol_intake_freq	no, Sometimes, Frequently, Always
Target Outcome	NObeyesdad	obesity_level_3cat	**Normal**: Normal_Weight **Overweight**: Level I & II **Obese**: Type I, II, & III

EDA

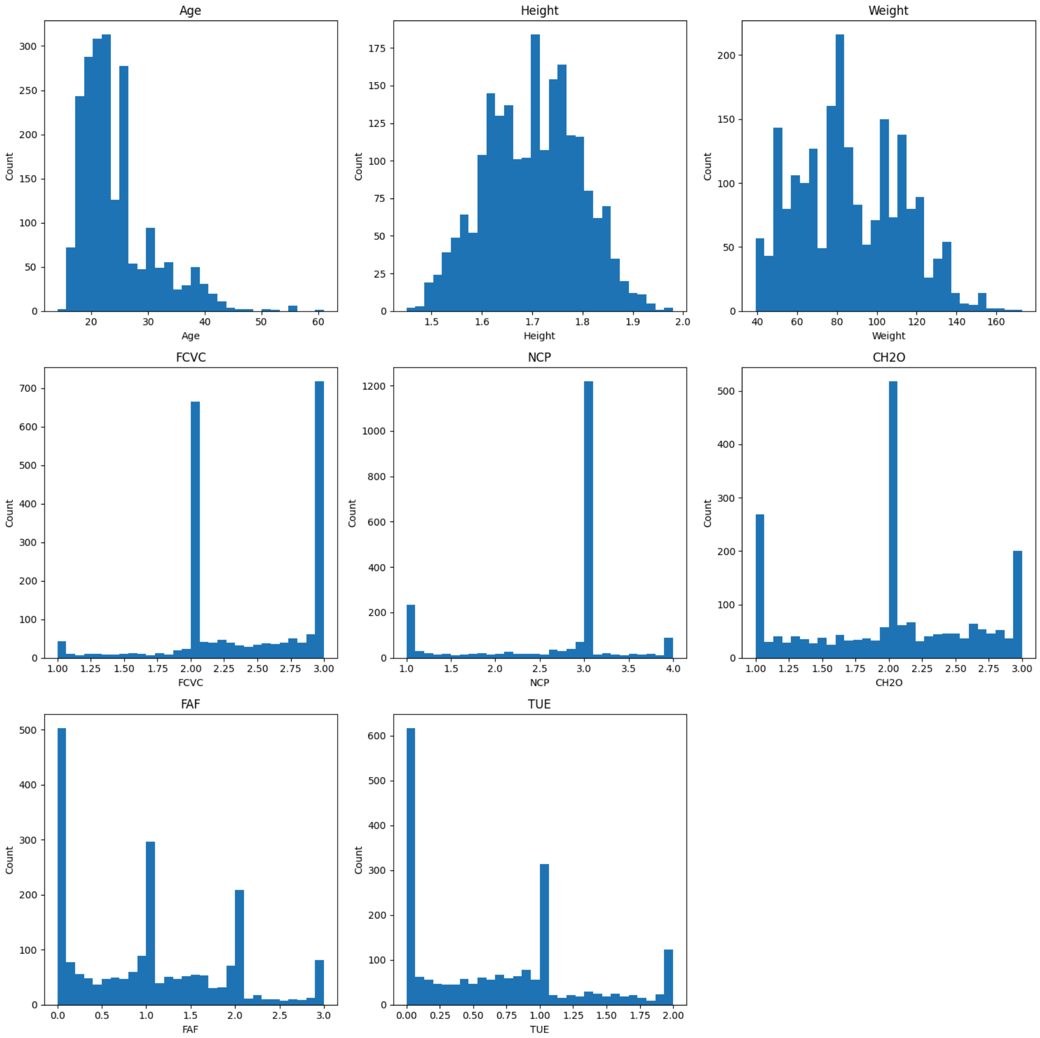

Age
Count
0
50
100
150
200
250
300
20
30
40
50
60
Age
Height
Count
0
25
50
75
100
125
150
175
1.5
1.6
1.7
1.8
1.9
2.0
Height
Weight
Count
0
50
100
150
200
40
60
80
100
120
140
160
Weight
FCVC
Count
0
100
200
300
400
500
600
700
1.00
1.25
1.50
1.75
2.00
2.25
2.50
2.75
3.00
FCVC
NCP
Count
0
200
400
600
800
1000
1200
1.0
1.5
2.0
2.5
3.0
3.5
4.0
NCP
CH2O
Count
0
100
200
300
400
500
1.00
1.25
1.50
1.75
2.00
2.25
2.50
2.75
3.00
CH2O
FAF
Count
0
100
200
300
400
500
0.0
0.5
1.0
1.5
2.0
2.5
3.0
FAF
TUE
Count
0
100
200
300
400
500
600
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
2.00
TUE

Figure 1.

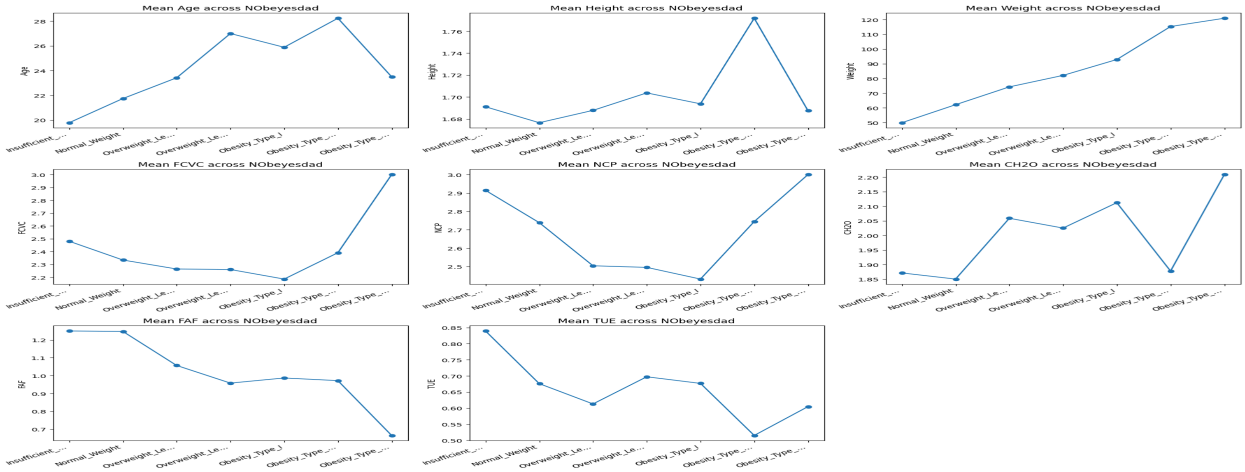

Mean Age across NObeyesdad
Age
20
22
24
26
28
Mean Height across NObeyesdad
Height
1.68
1.70
1.72
1.74
1.76
Mean Weight across NObeyesdad
Weight
50
60
70
80
90
100
110
120
Mean FCVC across NObeyesdad
FCVC
2.2
2.3
2.4
2.5
2.6
2.7
2.8
2.9
3.0
Mean NCP across NObeyesdad
NCP
2.5
2.6
2.7
2.8
2.9
3.0
Mean CH2O across NObeyesdad
CH2O
1.85
1.90
1.95
2.00
2.05
2.10
2.15
2.20
Mean FAF across NObeyesdad
FAF
0.7
0.8
0.9
1.0
1.1
1.2
Mean TUE across NObeyesdad
TUE
0.50
0.55
0.60
0.65
0.70
0.75
0.80
0.85
Insufficient_...
Normal_Weight
Overweight_Le...
Overweight_Le...
Obesity_Type_I
Obesity_Type_...
Obesity_Type_...

Figure 2.

MODEL

Figure A-1: Map of the BN structure.

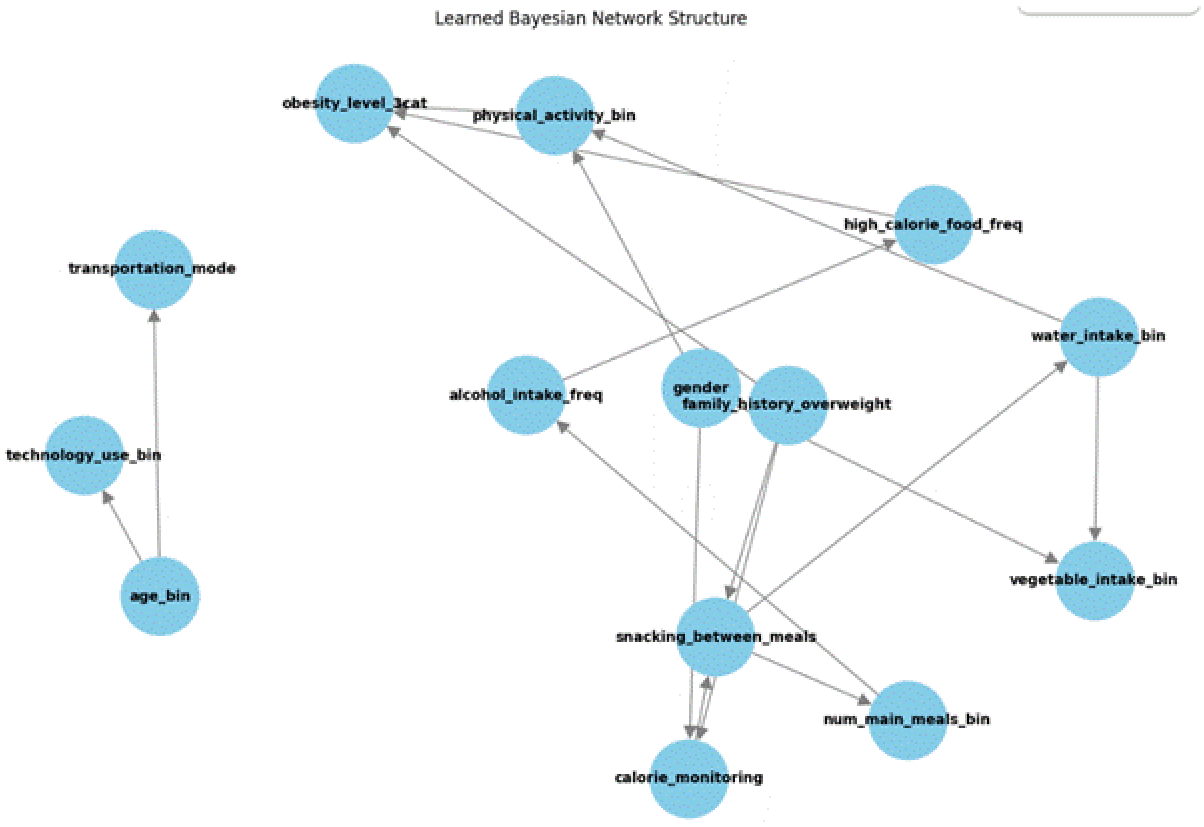
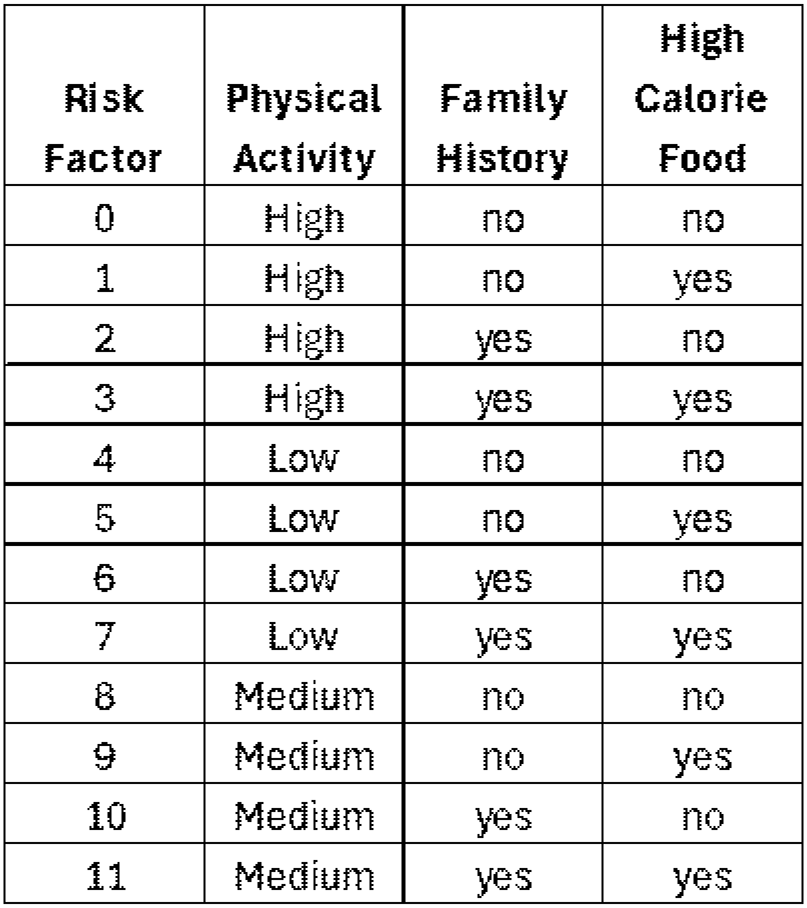


Figure A-2:



Risk Factor	Physical Activity	Family History	High Calorie Food
0	High	no	no
1	High	no	yes
2	High	yes	no
3	High	yes	yes
4	Low	no	no
5	Low	no	yes
6	Low	yes	no
7	Low	yes	yes
8	Medium	no	no
9	Medium	no	yes
10	Medium	yes	no
11	Medium	yes	yes