Introduction

College students often face a variety of lifestyle challenges that can impact their overall health. Balancing academics, part-time jobs, social obligations, and extracurricular commitments can lead to irregular sleep patterns, poor dietary habits, reduced physical activity, and increased stress. These behaviors may contribute to serious health outcomes, such as obesity. Understanding how daily lifestyle choices affect individual health can help identify risk factors early and support more informed health decisions. By identifying patterns in real world behavioral data, we can uncover which combination of habits most strongly predict health outcomes.

We used Decision trees and clustering techniques to examine the connection between lifestyle factors and health outcomes, with a focus on BMI. The data mining goals of this project are, Health Risk Prediction; identify individuals at higher risk of obesity issues based on patterns in their habits, Behavioral Analysis; determine which specific habits contribute most to BMI variation, Visual Modeling; create intuitive visualizations that map lifestyle patterns to health risk categories, and Lifestyle Grouping; cluster individuals into groups based on lifestyle similarities to assess varying BMI outcomes.

Data Description

The dataset used, from Kaggle (Health & Lifestyle), includes a range of lifestyle and health-related variables. It contains approximately 1,000 records that include both demographic and behavioral attributes, along with health outcome indicators.

Our response variables were BMI Category that is derived from continuous BMI values and classified as Underweight, Normal, Overweight, or Obese based the CDC guidelines. As for our predictor variables we had**,** Age, Gender, Hours of Sleep per Day, Daily Steps, Exercise Hours per Week, Calories Consumed per Day, Alcohol Consumption per Week, and Smoking Status. This dataset is well-suited for data mining applications, as it includes a mix of numerical and categorical variables and reflects real-world health behavior patterns across diverse individuals.

Methodology

This project employed supervised and unsupervised machine learning methods using JMP Pro 18 to analyze relationships between daily lifestyle behaviors and health outcomes. The main techniques used were decision tree modeling for prediction and k-means clustering for behavioral grouping. Luckily for the analysis we were doing there was not much we needed to clean up except for deleting the ID of each case.

Decision Trees

To identify health risk factors and predict outcomes, the partition platform in JMP Pro 18 was used to create decision trees for one response variable**.** BMI Category, the target variable, was derived from the original continuous BMI field and categorized into Underweight, Normal, Overweight, and Obese. Predictor variables included Age, Gender, Hours of Sleep, Daily Steps, Exercise Hours, Calories, Alcohol Consumption, and Smoking Status. Each tree was grown using recursive binary splits based on logworth and G^2 statistics, with R-squared reported to indicate model performance. Pruning was applied to prevent overfitting and limit the number of splits to meaningful levels.

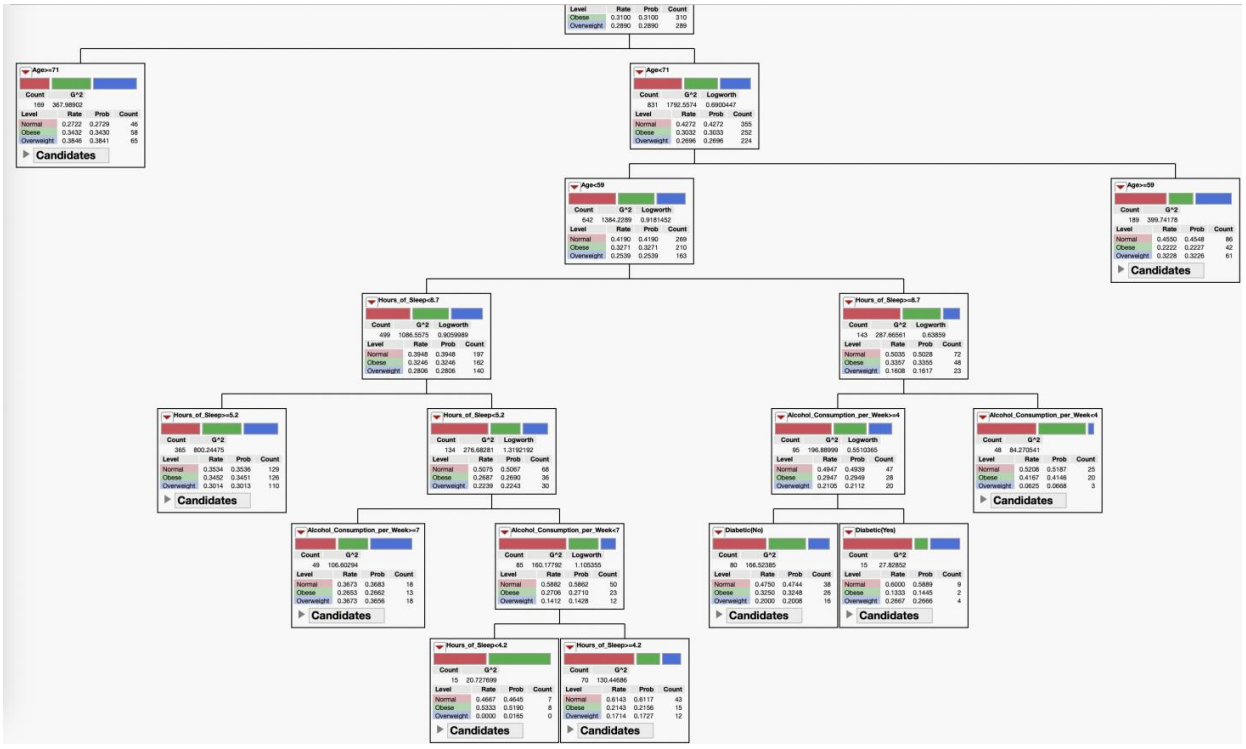
Clustering Analysis

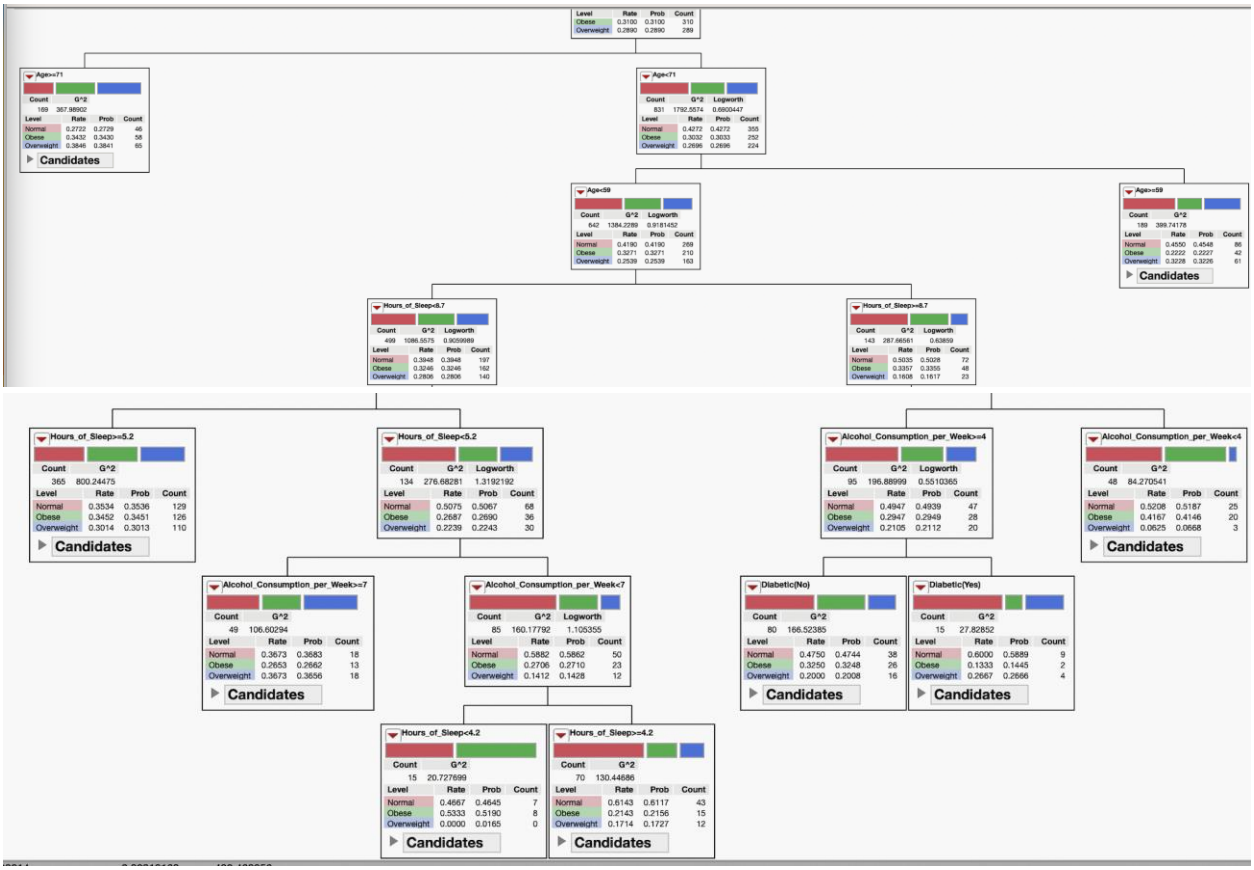
To support the goal of lifestyle grouping, a k-means clustering analysis was performed. Behavioral variables such as sleep duration, exercise hours, and calories consumed were included to identify natural groupings of individuals. The optimal number of clusters was selected based on interpretability and internal validation measures within JMP.

Results and Interpretation

Health Risk Prediction

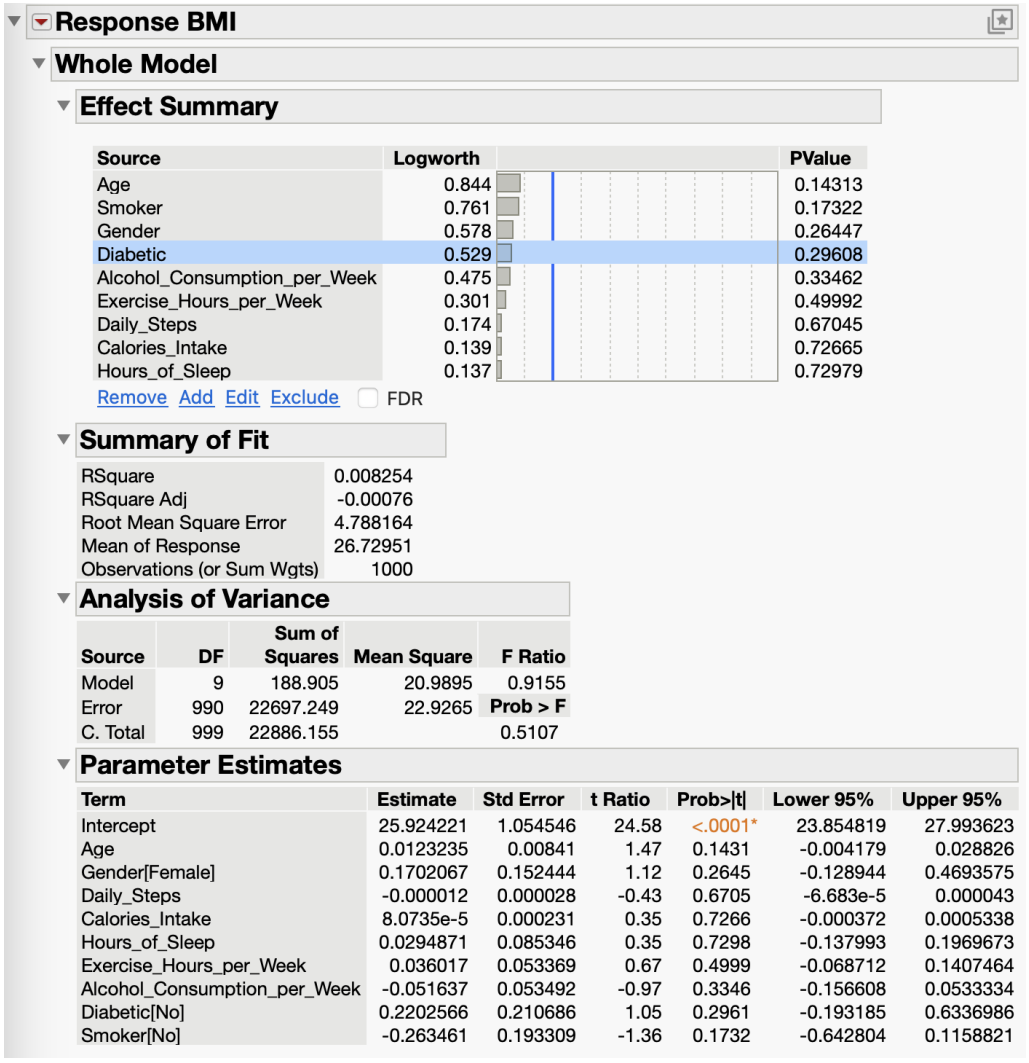
RSquare	N	Number of Splits
0.037	1000	8


Level Rate Prob Count
Obese 0.3100 0.3100 310
Overweight 0.2890 0.2890 289
Age>=71
Count G^2
169 367.98902
Level Rate Prob Count
Normal 0.2722 0.2729 46
Obese 0.3432 0.3430 58
Overweight 0.3846 0.3841 65
Candidates
Age<71
Count G^2 Logworth
831 1792.5574 0.6920447
Level Rate Prob Count
Normal 0.4272 0.4272 355
Obese 0.3032 0.3033 252
Overweight 0.2696 0.2696 224
Age<59
Count G^2 Logworth
642 1384.2289 0.9181452
Level Rate Prob Count
Normal 0.4190 0.4190 269
Obese 0.3271 0.3271 210
Overweight 0.2539 0.2539 163
Age>=59
Count G^2
189 399.74178
Level Rate Prob Count
Normal 0.4550 0.4548 86
Obese 0.2222 0.2227 42
Overweight 0.3228 0.3226 61
Candidates
Hours_of_Sleep<8.7
Count G^2 Logworth
499 1086.5575 0.9059989
Level Rate Prob Count
Normal 0.3948 0.3948 197
Obese 0.3246 0.3246 162
Overweight 0.2806 0.2806 140
Hours_of_Sleep>=8.7
Count G^2 Logworth
143 287.66561 0.63859
Level Rate Prob Count
Normal 0.5035 0.5028 72
Obese 0.3357 0.3355 48
Overweight 0.1608 0.1617 23
Hours_of_Sleep>=5.2
Count G^2
365 800.24475
Level Rate Prob Count
Normal 0.3534 0.3536 129
Obese 0.3452 0.3451 126
Overweight 0.3014 0.3013 110
Candidates
Hours_of_Sleep<5.2
Count G^2 Logworth
134 276.68281 1.3192192
Level Rate Prob Count
Normal 0.5075 0.5067 68
Obese 0.2687 0.2690 36
Overweight 0.2239 0.2243 30
Alcohol_Consumption_per_Week>=4
Count G^2 Logworth
95 196.88999 0.5510365
Level Rate Prob Count
Normal 0.4947 0.4939 47
Obese 0.2947 0.2949 28
Overweight 0.2105 0.2112 20
Alcohol_Consumption_per_Week<4
Count G^2
48 84.270541
Level Rate Prob Count
Normal 0.5208 0.5187 25
Obese 0.4167 0.4146 20
Overweight 0.0625 0.0668 3
Candidates
Alcohol_Consumption_per_Week>=7
Count G^2
49 106.60294
Level Rate Prob Count
Normal 0.3673 0.3683 18
Obese 0.2653 0.2662 13
Overweight 0.3673 0.3656 18
Candidates
Alcohol_Consumption_per_Week<7
Count G^2 Logworth
85 160.17792 1.105355
Level Rate Prob Count
Normal 0.5882 0.5862 50
Obese 0.2706 0.2710 23
Overweight 0.1412 0.1428 12
Diabetic(No)
Count G^2
80 166.52385
Level Rate Prob Count
Normal 0.4750 0.4744 38
Obese 0.3250 0.3248 26
Overweight 0.2000 0.2008 16
Candidates
Diabetic(Yes)
Count G^2
15 27.82852
Level Rate Prob Count
Normal 0.6000 0.5889 9
Obese 0.1333 0.1445 2
Overweight 0.2667 0.2666 4
Candidates
Hours_of_Sleep<4.2
Count G^2
15 20.727699
Level Rate Prob Count
Normal 0.4667 0.4645 7
Obese 0.5333 0.5190 8
Overweight 0.0000 0.0165 0
Candidates
Hours_of_Sleep>=4.2
Count G^2
70 130.44686
Level Rate Prob Count
Normal 0.6143 0.6117 43
Obese 0.2143 0.2156 15
Overweight 0.1714 0.1727 12
Candidates


Level Rate Prob Count
Obese 0.3100 0.3100 310
Overweight 0.2890 0.2890 289
Age>=71
Count G^2
169 367.98902
Level Rate Prob Count
Normal 0.2722 0.2729 46
Obese 0.3432 0.3430 58
Overweight 0.3846 0.3841 65
Candidates
Age<71
Count G^2 Logworth
831 1792.5574 0.6920447
Level Rate Prob Count
Normal 0.4272 0.4272 355
Obese 0.3032 0.3033 252
Overweight 0.2696 0.2696 224
Age<59
Count G^2 Logworth
642 1384.2289 0.9181452
Level Rate Prob Count
Normal 0.4190 0.4190 269
Obese 0.3271 0.3271 210
Overweight 0.2539 0.2539 163
Age>=59
Count G^2
189 399.74178
Level Rate Prob Count
Normal 0.4550 0.4548 86
Obese 0.2222 0.2227 42
Overweight 0.3228 0.3226 61
Candidates
Hours_of_Sleep<8.7
Count G^2 Logworth
499 1086.5575 0.9059989
Level Rate Prob Count
Normal 0.3948 0.3948 197
Obese 0.3246 0.3246 162
Overweight 0.2806 0.2806 140
Hours_of_Sleep>=8.7
Count G^2 Logworth
143 287.66561 0.63809
Level Rate Prob Count
Normal 0.5035 0.5028 72
Obese 0.3357 0.3355 48
Overweight 0.1608 0.1617 23
Hours_of_Sleep>=5.2
Count G^2
365 800.24475
Level Rate Prob Count
Normal 0.3534 0.3536 129
Obese 0.3452 0.3451 126
Overweight 0.3014 0.3013 110
Candidates
Hours_of_Sleep<5.2
Count G^2 Logworth
134 276.68281 1.3192192
Level Rate Prob Count
Normal 0.5075 0.5067 68
Obese 0.2687 0.2690 36
Overweight 0.2239 0.2243 30
Alcohol_Consumption_per_Week>=4
Count G^2 Logworth
95 196.88999 0.5510365
Level Rate Prob Count
Normal 0.4947 0.4939 47
Obese 0.2947 0.2949 28
Overweight 0.2105 0.2112 20
Alcohol_Consumption_per_Week<4
Count G^2
48 84.270541
Level Rate Prob Count
Normal 0.5208 0.5187 25
Obese 0.4167 0.4146 20
Overweight 0.0625 0.0668 3
Candidates
Alcohol_Consumption_per_Week>=7
Count G^2
49 106.60294
Level Rate Prob Count
Normal 0.3673 0.3683 18
Obese 0.2653 0.2662 13
Overweight 0.3673 0.3656 18
Candidates
Alcohol_Consumption_per_Week<7
Count G^2 Logworth
85 160.17792 1.105355
Level Rate Prob Count
Normal 0.5882 0.5862 50
Obese 0.2706 0.2710 23
Overweight 0.1412 0.1428 12
Diabetic(No)
Count G^2
80 166.52385
Level Rate Prob Count
Normal 0.4750 0.4744 38
Obese 0.3250 0.3248 26
Overweight 0.2000 0.2008 16
Candidates
Diabetic(Yes)
Count G^2
15 27.82852
Level Rate Prob Count
Normal 0.6000 0.5889 9
Obese 0.1333 0.1445 2
Overweight 0.2667 0.2666 4
Candidates
Hours_of_Sleep<4.2
Count G^2
15 20.727699
Level Rate Prob Count
Normal 0.4667 0.4645 7
Obese 0.5333 0.5190 8
Overweight 0.0000 0.0165 0
Candidates
Hours_of_Sleep>=4.2
Count G^2
70 130.44686
Level Rate Prob Count
Normal 0.6143 0.6117 43
Obese 0.2143 0.2156 15
Overweight 0.1714 0.1727 12
Candidates

We used a classification decision tree to model BMI Category based on behavioral variables. The tree generated eight splits, with an R-squared value of approximately 0.034, indicating modest explanatory power. The most important finding was that age was the strongest initial splitter, with individuals aged 71 and older showing a higher prevalence of overweight and obesity. Sleep duration also played a key role. Very short sleep (less than 5.2 hours) and long sleep (more than 8.7 hours) were both associated with elevated obesity risk. Additionally, exercise exceeding 7.8 hours per week was linked to healthier BMI outcomes, while high alcohol consumption (seven or more drinks per week), particularly when combined with low activity, was associated with obesity. These results suggest that negative behaviors can compound health risks, while balanced sleep and regular exercise support better weight outcomes.

Behavioral Analysis



Response BMI

Whole Model

Effect Summary

Source	Logworth	PValue
Age	0.844	0.14313
Smoker	0.761	0.17322
Gender	0.578	0.26447
Diabetic	0.529	0.29608
Alcohol_Consumption_per_Week	0.475	0.33462
Exercise_Hours_per_Week	0.301	0.49992
Daily_Steps	0.174	0.67045
Calories_Intake	0.139	0.72665
Hours_of_Sleep	0.137	0.72979

Remove Add Edit Exclude FDR

Summary of Fit

RSquare	0.008254
RSquare Adj	-0.00076
Root Mean Square Error	4.788164
Mean of Response	26.72951
Observations (or Sum Wgts)	1000

Analysis of Variance

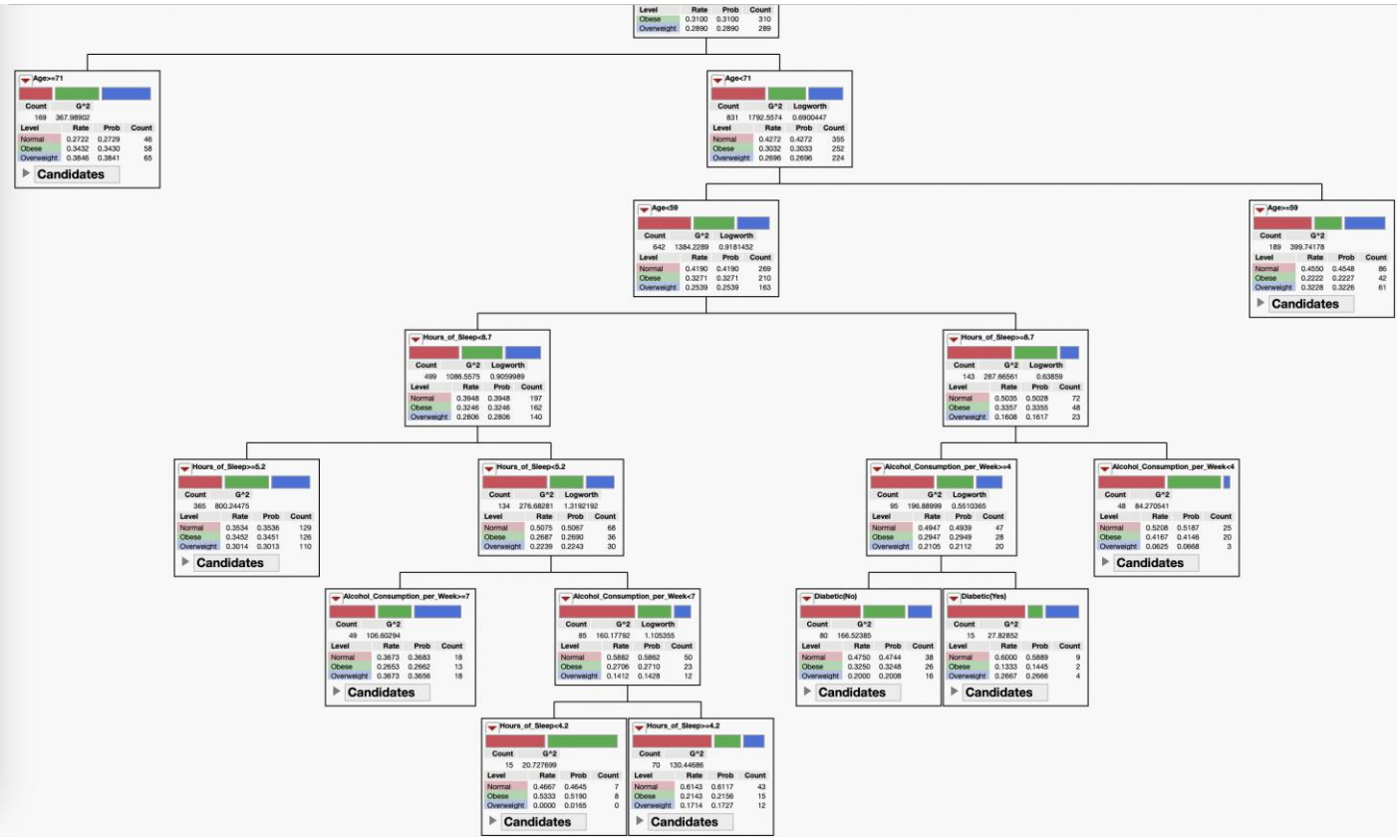
Source	DF	Sum of Squares	Mean Square	F Ratio
Model	9	188.905	20.9895	0.9155
Error	990	22697.249	22.9265	Prob > F
C. Total	999	22886.155		0.5107

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|	Lower 95%	Upper 95%
Intercept	25.924221	1.054546	24.58	<.0001*	23.854819	27.993623
Age	0.0123235	0.00841	1.47	0.1431	-0.004179	0.028826
Gender[Female]	0.1702067	0.152444	1.12	0.2645	-0.128944	0.4693575
Daily_Steps	-0.000012	0.000028	-0.43	0.6705	-6.683e-5	0.000043
Calories_Intake	8.0735e-5	0.000231	0.35	0.7266	-0.000372	0.0005338
Hours_of_Sleep	0.0294871	0.085346	0.35	0.7298	-0.137993	0.1969673
Exercise_Hours_per_Week	0.036017	0.053369	0.67	0.4999	-0.068712	0.1407464
Alcohol_Consumption_per_Week	-0.051637	0.053492	-0.97	0.3346	-0.156608	0.0533334
Diabetic[No]	0.2202566	0.210686	1.05	0.2961	-0.193185	0.6336986
Smoker[No]	-0.263461	0.193309	-1.36	0.1732	-0.642804	0.1158821

We applied multiple linear regression to assess which behaviors most influenced BMI as a continuous variable. The model's R-squared was 0.008, indicating very low predictive power. No behavioral variables showed statistically significant effects (all p-values > 0.14), although Age, Smoker, and Gender had higher relative Logworth. This suggests that BMI variation may not be well explained linearly by individual variables alone, or that interaction effects are more meaningful.

Visual Modeling of Habit Combinations



We used partition plots to visualize how lifestyle combinations influence BMI outcomes. The plot clearly displayed BMI category regions across splits involving age, hours of sleep, alcohol consumption, and exercise. Dense clusters of obesity were seen in groups with less than 5 hours of sleep and high alcohol use. Healthier BMI zones were concentrated where individuals maintained 5–8 hours of sleep and higher physical activity. These visuals support earlier interpretations and make risk patterns easier to communicate.

Lifestyle Grouping (Clustering)

Iterative Clustering

Cluster Comparison

Method	NCluster	CCC	Best
K Means Cluster	3	-7.6203	Optimal CCC
K Means Cluster	4	-12.34	

Columns Scaled Individually

Control Panel

K Means NCluster=3

Columns Scaled Individually

Cluster Summary

Cluster	Count	Step	Criterion
1	330	20	0
2	365		
3	305		

Cluster Means

Cluster	Hours_of_Sleep	Exercise_Hours_per_Week	Calories_Intake
1	6.03969697	7.74090909	2299.45152
2	8.82876712	4.65315068	2357.19726
3	5.55508197	2.59868852	2321.05246

Cluster Standard Deviations

K Means NCluster=4

Columns Scaled Individually

Cluster Summary

Cluster	Count	Step	Criterion
1	257	10	0
2	236		
3	272		
4	235		

Cluster Means

Cluster	Hours_of_Sleep	Exercise_Hours_per_Week	Calories_Intake
1	8.19416342	6.61984436	1772.37354
2	5.7470339	7.54152542	2741.0678
3	5.43014706	2.95220588	2014.43015
4	8.38595745	3.24	2880

Cluster Standard Deviations

Cluster	Hours_of_Sleep	Exercise_Hours_per_Week	Calories_Intake
1	1.1812103	2.21957054	377.748234
2	1.25064038	1.66049953	493.661042
3	1.03967754	1.89123236	536.347316
4	1.11398014	2.20919163	422.462256

We performed K-means clustering using hours of sleep, exercise per week, and calorie intake. The 3-cluster solution was optimal. Cluster 1, size of 330, had 6 hours of sleep, high exercise (7.7 hours), and moderate calorie intake. Cluster 2, size of 365, had 8.8 hours of sleep, moderate exercise (4.6 hours), and high calorie intake. Cluster 3, size of 305, had 5.6 hours of sleep, low exercise (2.6 hours), and high calorie intake. Although the cubic clustering criterion (CCC) was negative (–7.62), the clusters revealed distinct behavioral profiles, supporting the use of clustering for exploratory grouping.

Conclusion

This project explored how daily lifestyle habits relate to health outcomes, with a focus on predicting and understanding BMI variation. Through decision tree modeling, regression, and clustering, we examined both individual and combined behavioral patterns.

Key findings showed that age, sleep duration, alcohol consumption, and exercise habits interact in meaningful ways to influence BMI categories. Although no single behavior strongly predicted BMI in isolation, the decision tree revealed important combinations that elevated risk. Clustering further revealed distinct lifestyle profiles such as high-activity individuals with moderate sleep versus low activity, short sleep individuals which may support targeted health recommendations. While model performance was modest, the results offer interpretable patterns and a strong foundation for future studies that may include more health outcomes or psychosocial variables. These findings can help inform interventions aimed at improving student wellness through balanced behavioral habits.