Data Mining for Lifestyle Risk Factors Associated with Overweight and Obesity among Adolescents

Anthony Pochini, Yitian Wu, Gongzhu Hu
Department of Computer Science
Central Michigan University
Mount Pleasant, MI 48859
(pochi1as, wu6y, hu1g)@cmich.edu

***Abstract*—Data mining techniques have been applied to many areas in the business world and our daily life, including healthcare and clinical health services. One of the mostly watched health problems is obesity and overweight, particularly for children and adolescents. In this paper, we try to find the most significant lifestyle risk factors associated with overweight and obesity among high school students in the US. Lifestyle survey data from the 2011 National Youth Risk Behavior Survey (YRBS) was used with the students' body weight statuses, overweight or obesity, considered as two target variables. Both logistic regression models and decision tree models were created for each target variable. Both the logistic regression and decision tree method show that frequently doing physical activity and having breakfast everyday were protective factors against being overweight or obese. Smoking and drinking sugar-sweeten beverage frequently were found to be associated with an increased risk to be obese.**

***Index Terms*—overweight and obesity, data mining, logistic regression, decision tree**

I. Introduction

Obesity as a global epidemic has widely spreaded in recent years that has become a major health problem in many regions of the world [13], [16], [17]. According to a comparitive study [10] about the overweight and obesity prevalence in school-aged youth, some countries such as the United States had over 25% of youth overweight and obese and increased to over 30% in 2013 in the US [2].

People believe that when risk factors associated with overweight and obesity are well understood, it is possible to educate the public to reduce their obesity risk by avoiding those risk behaviors. Such a health education might make more significant effect for children and adolescents because those poor habits and risk behaviors are not established or consolidated during childhood. Thus, identification of the associated risk factors is a top of priority to reduce the prevalence of overweight and obesity among children and adolescents.

Previous studies [8] show that the potential obesity risk factors in the young population can be divided into three groups: individual characteristics, lifestyle factors, and family/environment factors. Individual characteristics relate to the child or adolescent himself, such as health, age and gender. Lifestyle risk factors related to their lifestyle, such as sedentary habits, physical activities, daily nutrition, sleep duration, etc.. Family/ environment related to genetic cause and living environment. Although the risk factors from the other categories are not to be ignored. The lifestyle behavior is more likely to be closely related to the prevalence of overweight and obesity problems in the population [6]. Furthermore, the lifestyle modifications are more reliable in the management of body weight. Since the pattern of lifestyle risk factors among children and adolescents could change over years, it is necessary to monitor the current trend based on the latest updated data.

In this study, the data from the 2011 Youth Behavior Risk Survey (YBRS) [4] is used to examine the prevalence of overweight and obesity among the high school students in US. Logistic regression models were built to investigate whether these factors were associated with overweight and obesity in this population in 2011. Decision trees based on the same factors were also built, and results were compared between the two types of models.

II. Methodology

A. Study Sample

The target population was all regular public, Catholic, and other private school students, in grades 9 through 12, in 50 states and the District of Columbia of the USA. A three-stage clustered sampling design was used to produce a representative sample of 9th through 12th grade students. A total of 194 schools were selected systematically with probability proportional to enrollment in grades 9 through 12 using a random start. Classes in 9-12 grades in those selected schools were sampled with systematic equal probability. All students attending in the selected classes were encouraged to submit the responses to the questionnaire. A representative sample of 17,672 students were surveyed, and 15425 (87%) of them provided complete usable data and were included in the present analysis.

B. Data Collection

The data was obtained from the 2011 National Youth Risk Behavior Survey (YRBS). In the field survey, questionnaires were distributed to the students and no anthropometric measurement or medical examination was performed. All information was based on the students responses to the questionnaires. The data that came out from the questionnaires was processed

978-1-4799-4173-5/14 $31.00 © 2014 IEEE
DOI 10.1109/IIAI-AAI.2014.175

and cleaned, including identifying the logical confliction to set as missing values. Among the collected data, only those about demographic characteristics, alcohol and tobacco use, dietary behaviors, and physical activity were considered in our study.

1) Variables considered: Only the variables related to the lifestyle of the high school students were considered as factors in the present study. All factors and two target variables are binary, where the result is *yes* or *no*, with a third option allowed where the value was missed. These variables are coded 1 for *yes* and 0 for *no*.

2) Targets: The two target variables represent whether or not a student is overweight (denoted by QNOWT) or obesity (denoted by QNOBESE), respectively. They are determined by the Body Mass Index (*BMI*) that is calculated using the height H (in kilogram) and weight W (in meter) of a person in the following formula:

$$BMI = \frac{W}{H^2}$$

The two target variables are defined as:

$$\text{NNOWT} = \begin{cases} 1, \text{ if } 85\% \leq \textit{BMI} \text{ percentile} < 95\% \\ 0, \text{ otherwise} \end{cases}$$

$$\text{QNOBESE} = \begin{cases} 1, \text{ if } \textit{BMI} \text{ percentile} \geq 95\% \\ 0, \text{ otherwise} \end{cases}$$

If BMI was missing, then QNOWT and QNOBESE were set to missing, these two target variables were mutually exclusive.

TABLE I
FACTOR VARIABLES

Variable	Meaning
QN42	whether the student had 1+ drinks past 30 days.
QN72	whether the student drank fruit juice past 7 days
QN78	whether the student drank soda 1+ times/day past 7 days
QN91	whether the student ate breakfast on all of the past 7 days
QN96	whether the student gets 8+ hours sleep
QNANYTOB	whether the student used any tobacco past 30 days
QNFRVG	whether the student ate 5+ fruits/vegetables/day 7 days
QNPA7DAY	whether the student is physically active 7 of past 7 days
TV_Game	whether the student watches TV or playing video/computer games over 3 hours each day

The following Pearson Correlation Statistics were obtained for the factor variable.

C. Statistical Analysis

Binary logistic regression analysis was performed to examine the associations between overweight and obesity with students lifestyle factors. Odds Ratios (ORs) and 95% confidence intervals (CIs) were calculated. Statistical significance was set at P_0.05. Data were analyzed with SPSS.

TABLE II
PEARSON CORRELATION STATISTICS

	Overweight	Obesity
Used tobacco past 30 days	0.007554	0.036373
Soda 1+ time per day past 7 days	-0.008633	0.023361
5+ fruits/vegetables /day past 7 days	0.015055	0.020395
TV / video games over 3 hours / day	-0.013247	0.004126
1+ Alcoholic Drinks past 30 days	-0.004209	-0.005558
Get 8+ hours of sleep per night	0.014203	-0.011265
Drank fruit juice past 7 days	-0.015976	-0.017209
Ate breakfast all of past 7 days	-0.020326	-0.033645
Physically active 7 of past 7 days	-0.024551	-0.038662

1) Logistic Regression Results: The following table presents the prevalence of overweight and obesity among US high school students in 2011. 13.6% were obese (students who were $\geq$ 95th percentile for body mass index), and 15.3% were overweight (students who were $\leq$ 95th and $\geq$ 85th percentile for body mass index). Both BMI percentiles were based on sex- and age-specific reference data from the 2000 CDC growth charts.

TABLE III
PREVALENCE OF OVERWEIGHT AND OBESITY AMONG US HIGH SCHOOL STUDENTS IN 2011

	Overweight		Obesity	
	Frequency	Percent	Frequency	Percent
0	7699	84.7	7852	86.4
1	1390	15.3	1237	13.6
Total	9089	100.0	9089	100.0

TABLE IV
PREVALENCE (%) OF LIFESTYLE BEHAVIOR FACTORS AMONG HIGH SCHOOL STUDENTS IN 2011 IN US

Variable	Description for Variable	Percent (%)
Tv_Game	Sedentary behavior (Screen time $\geq$ 3h per day)	66.5
QNANYTOB	Smoking Consumption	22.9
QN42	Alcohol Consumption	39.1
QNPA7DAY	Physical Activity	28.9
QN72	Fruit Juice Consumption	81.6
QNFRVG	Frequent Fruit/Vegetable Consumption	22.1
QN78	Suger-Sweeten Beverage Consumption	27.6
QN91	Breakfast	37.1
QN96	Sleep Duration	31.4

The school health policies and programs in 2006 recommends all school-age children do moderate or vigorous physical activity for at least 60 minutes per day. Nearly 71% of high school students did not meet the recommendation. Nearly 23% of the students smoked at least once in the past 30 days. 39% of students reported consuming alcoholic beverages more than once in the past 30 days. On average 67% of students spent three hours or more per day watching TV, or using

computer and/or playing video games. Only 31% of the survey participants had adequate hours of sleep based on the US National Sleep Foundations recommendation. Nearly 37% of the students ate breakfast each day on the past whole week. Over 80% drank fruit juice once and 22% of students ate fruits or vegetables more than 5 times during the past 7 days. Nearly 28% drank a can, bottle, or glass of soda or pop more than once per day during the 7 days before the survey.

a) Creation of Logistic regression model: Using logistic regression analysis the association between a set of lifestyle behavior factors with overweight and obesity, respectively was assessed. Only frequent eating breakfast and physical inactivity were significantly associated with overweight. Compared with those who eat breakfast each day, students who reported infrequently eating breakfast were 15% more likely to be overweight (overweight OR: 0.85, 95% CI: 0.7520.966). In this study, physical inactivity was defined as time of doing any kind of physical activity that increased their heart rate and made them breathe hard was less than once per day. Physical activity was associated with a lower odds of overweight (OR: 0.864, 95% CI: 0.7560.987).

The other logistic regression analysis identified three lifestyle risk behavior factors and three protective behavior factors related to obesity. Students who frequently drank sugar-sweeten beverage (obesity OR: 1.19, 95% CI: 1.041.36) and having smoking habit (obesity OR: 1.26, 95% CI: 1.081.48) were more likely to be obese than the others. Frequent eating breakfast (obesity OR: 0.83, 95% CI: 0.720.94), consumption of fruit juice (obesity OR: 0.86, 95% CI: 0.731.00) and adequate physical activity (obesity OR: 0.74, 95% CI: 0.640.86) were significant protective factors for obesity. However, alcohol consumption, adequate sleep and spending over three hours or more watching TV, or playing computer / video game were not significant for obesity. Frequently having fruit/vegetable was positively associated with obesity, which is different with previous study and common sense.

D. Decision Tree Models

Two decision tree models were built. One used obesity as the target variable and the other used overweight as the target variable. The trees were generated using SAS Enterprise Miner with entropy used as the splitting rule. They were pruned using the `Decision` option, where the tree with the largest average profit and smallest average loss was selected. The data was partitioned randomly, with 70 percent put into the training set and 30 percent put into the validation set.

1) Obesity Decision Tree: The decision tree in Figure 1 was obtained using obesity as the target variable.

The tree shows that if the variable physically active 7 of past 7 days has a value of 1, then the student is most likely not obese. If they are not physically active and they used tobacco in the past 30 days, then they were most likely obese. If they were not physically active and did not use tobacco, then they were most likely to be obese if they did not eat breakfast all of the past 7 days and not obese if they did.

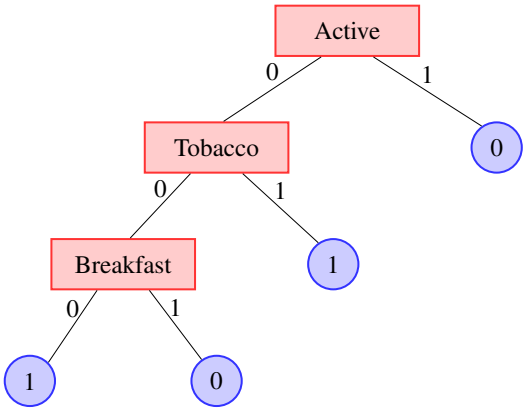

Fig. 1. Decision tree for obesity

The variable importance measures for this tree are given in the following table with higher values indicating that the variable is more significant in predicting the target variable.

TABLE VII
IMPORTANCE OF VARIABLES FOR OBESITY

Obesity	
Variable	**Importance**
Active	1.000
Tobacco	0.934
Breakfast	0.731

Using the validation set, the model was scored and the score rankings matrix plot was obtained, as shown in Figure 2.

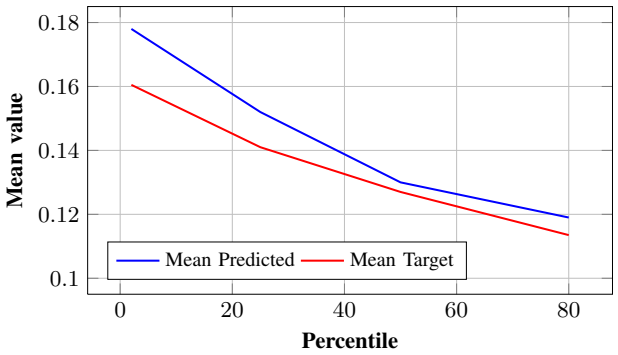

Fig. 2. Predicted and actual mean values (obesity)

The plot shows that the mean value predicted by the model is relatively close to the actual mean value, which suggests the model is pretty good.

2) Overweight Decision Tree: A second decision tree was generated using all of the same data and settings as the first, but with overweight used as the target variable. After pruning this tree, all of the branches were removed, which shows that no tree gives a sufficient level of average profit with

TABLE V
ASSOCIATION BETWEEN LIFESTYLE BEHAVIOR FACTORS AND OVERWEIGHT AMONG HIGH SCHOOL STUDENTS IN 2011 IN US

Variable	Description of Variable	Overweight			
		OR	p-value	C.I. for OR	
				Lower	Upper
Tv_Game	Sedentary behavior (Screen time≥3h per day)	0.957	0.478	0.848	1.080
QN42	Alcohol Consumption	0.937	0.342	0.820	1.071
QN72	Fruit Juice Consumption	0.880	0.088	0.759	1.019
QN78	Suger-Sweeten Beverage Consumption	0.931	0.289	0.816	1.062
QN91	Breakfast	0.852	0.012	0.752	0.966
QN96	Sleep Duration	1.085	0.204	0.957	1.229
QNANYTOB	Smoking Consumption	1.060	0.460	0.908	1.236
QNFRVG	Frequent Fruit/Vegetable Consumption	1.127	0.102	0.977	1.301
QNPA7DAY	Physical Activity	0.864	0.032	0.756	0.987
Constant		0.222	0.000		

TABLE VI
ASSOCIATION BETWEEN LIFESTYLE BEHAVIOR FACTORS AND OBESITY AMONG HIGH SCHOOL STUDENTS IN 2011 IN US

Variable	Description of Variable	Obesity			
		OR	p-value	C.I. for OR	
				Lower	Upper
Tv_Game	Sedentary behavior (Screen time≥3h per day)	0.989	0.865	0.870	1.124
QN42	Alcohol Consumption	0.897	0.129	0.779	1.032
QN72	Fruit Juice Consumption	0.857	0.050	0.734	1.000
QN78	Suger-Sweeten Beverage Consumption	1.192	0.010	1.044	1.361
QN91	Breakfast	0.826	0.005	0.723	0.943
QN96	Sleep Duration	1.010	0.880	0.884	1.155
QNANYTOB	Smoking Consumption	1.265	0.004	1.080	1.481
QNFRVG	Frequent Fruit/Vegetable Consumption	1.361	0.000	1.175	1.576
QNPA7DAY	Physical Activity	0.744	0.000	0.644	0.859
Constant		0.181	0.000		

overweight as the target variable. The unpruned decision tree is shown in Figure 3.

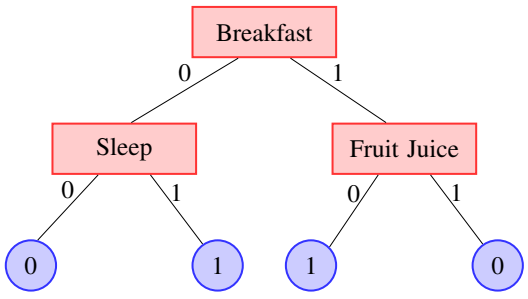

Fig. 3. Decision tree for overweight

This tree shows that the most significant factor towards determining if a student is overweight or not is whether he or she ate breakfast all of the past 7 days or not. For students that did not eat breakfast, those that got 8+ hours of sleep were more likely overweight while those that did not were more likely not overweight. For students that did eat breakfast, those that drank fruit juice the past 7 days we more likely to not be overweight while those that did not were more likely to be overweight. The variable importance measures are given below.

TABLE VIII
IMPORTANCE OF VARIABLES FOR OVERWEIGHT

Overweight	
Variable	**Importance**
Active	1.000
Tobacco	0.978
Breakfast	0.784

Using the validation set, the model was scored and the following score rankings matrix plot was obtained.

The plot shows that the mean value predicted by the model is very different from the actual mean value. This suggests that the model is not very good. This is expected because all of the branches were removed during the pruning process.

III. EXPERIMENTAL RESULTS

The ranked variable importance for the for models are shown in Table IX, with the variables for obesity in IX(a) and for overweight in IX(b). Tmost significant variables for predicting the target variable shown first.

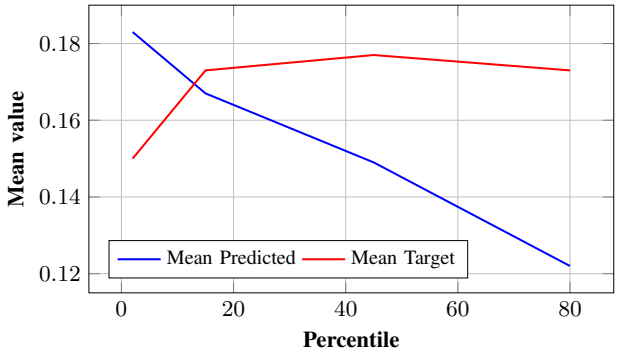

Fig. 4. Predicted and actual mean values (overweight)

TABLE IX
MOST SIGNIFICANT VARIABLES BY MODEL

(a) Significant Variables for Obesity

Logistic Regression	Decision Tree
Fruit/Vegetable	Physically Active
Tobacco	Tobacco
Physically Active	
Breakfast	
Fruit Juice	
Soda	

(b) Significant Variables for Overweight

Logistic Regression	Decision Tree
Breakfast	Breakfast
Physically Active	Fruit Juice
Breakfast	Sleep

The table shows that for predicting obesity, both the logistic regression model and the decision tree model determined that the significant lifestyle factors included being physically active for seven out of seven days, using tobacco within 30 days, and eating breakfast everyday. The only factor found by both models to significantly effect students being overweight was eating breakfast everyday.

IV. Related Work

The problem of overweight and obesity prevalence in children and adolescents has been studied for many years, both in the healthcare/medical and the data mining communities, as the problem itself is interdisciplinary in nature across the two communities.

In the healthcare and medical field, people studied the trend of overweight and obesity [15], [17], the prevalence of overweight and obesity in various countries (US [9], Canada [7], China [6], India [5], etc.), assessment [11], and risk factors in categories of socioeconomic, individual characteristics, as well as lifestyle [1], [10]. These studies revealed that the overweight and obesity prevalence, psecially in children and adolescents, has been getting worst in the last twi decades and sperading over many parts of the world. All the risk factors identified need real close attention by the governments, healthcare providers, and the public.

In the data mining community, various machine learning methods have been applied to find patterns and predict overweight and obesity [3]. Regression [18], Bayesian classification [14], predictive modeling [8], and decision tree [12] approaches are some of the commonly used techniques. These general data mining techniques are quite effective applied to the health-related datasets.

V. Conclusion

Using lifestyle survey data for high school students, along with logistic regression and decision tree models, the most significant factors contributing to or preventing childhood overweight and obesity were determined. It was found that physical activity and eating breakfast were significant in preventing obesity, while tobacco use is a factor significantly leading to obesity. The only factor found to significantly contribute to childhood overweight was eating breakfast, which was negatively associated with overweight. As a result, it is recommended that students be encouraged to be physically active, eat breakfast, and avoid tobacco.

References

[1] Hazzaa M Al-Hazzaa, Nada A Abahussain, Hana I Al-Sobayel, Dina M Qahwaji, and Abdulrahman O Musaiger. Lifestyle factors associated with overweight and obesity among saudi adolescents. *BMC public health*, 12(1):354, 2012.

[2] American Heart Association. Statistical fact sheet 2013 update: Overweight & obesity. http://www.heart.org/idc/groups/heart-public/@wcm/@sop/@smd/documents/downloadable/ucm_319588.pdf, 2013.

[3] Riccardo Bellazzi, Fulvia Ferrazzi, and Lucia Sacchi. Predictive data mining in clinical medicine: a focus on selected methods and applications. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 1(5):416–430, 2011.

[4] Center for Disease Control and Prevention. Youth risk behavior surveillance system. http://www.cdc.gov/HealthyYouth/yrbs/data/index.htm, 2011.

[5] Ramesh K Goyal, Vitthaldas N Shah, Banshi D Saboo, Sanjiv R Phatak, Navneet N Shah, Mukesh C Gohel, Prashad B Raval, and Snehal S Patel. Prevalence of overweight and obesity in indian adolescent school going children: its relationship with socioeconomic status and associated lifestyle factors. *The Journal of the Association of Physicians of India*, 58:151–158, 2010.

[6] Xiaofan Guo, Liqiang Zheng, Yang Li, Xiaoyu Zhang, Shasha Yu, Hongmei Yang, Xingang Zhang, Zhaoqing Sun, and Yingxian Sun. Prevalence and risk factors of being overweight or obesese among children and adolescents in northeast china. *Pediatric Research*, 74(4):443–449, 2013.

[7] Anthony JG Hanley, Stewart B Harris, Joel Gittelsohn, Thomas MS Wolever, Brit Saksvig, and Bernard Zinman. Overweight among children and adolescents in a native canadian community: prevalence and associated factors. *The American journal of clinical nutrition*, 71(3):693–700, 2000.

[8] Muhamad Hariz, B. Muhamad Adnan, Wahidah Husain, and Nur'Aini Abdul Rashid. Parameter identification and selection for childhood obesity prediction using data mining. *2nd International Conference on Management and Artificial Intelligence IPEDR*, 35, 2012.

[9] Allison A Hedley, Cynthia L Ogden, Clifford L Johnson, Margaret D Carroll, Lester R Curtin, and Katherine M Flegal. Prevalence of overweight and obesity among us children, adolescents, and adults, 1999-2002. *Jama*, 291(23):2847–2850, 2004.

[10] Ian Janssen, Peter T Katzmarzyk, William F Boyce, Carine Vereecken, Caroline Mulvihill, Chris Roberts, Candace Currie, and William Pickett. Comparison of overweight and obesity prevalence in school-aged youth from 34 countries and their relationships with physical activity and dietary patterns. *Obesity reviews*, 6(2):123–132, 2005.

[11] Nancy F Krebs, John H Himes, Dawn Jacobson, Theresa A Nicklas, Patricia Guilday, and Dennis Styne. Assessment of child and adolescent overweight and obesity. *Pediatrics*, 120(Supplement 4):S193–S228, 2007.

[12] Stephenie C Lemon, Jason Roy, Melissa A Clark, Peter D Friedmann, and William Rakowski. Classification and regression tree analysis in public health: methodological review and comparison with logistic regression. *Annals of behavioral medicine*, 26(3):172–181, 2003.

[13] Tim Lobstein, Louise Baur, and Ricardo Uauy. Obesity in children and young people: a crisis in public health. *Obesity reviews*, 5(s1):4–85, 2004.

[14] Peter Lucas. Bayesian analysis, pattern analysis, and data mining in health care. *Current opinion in critical care*, 10(5):399–403, 2004.

[15] Cynthia L Ogden, Katherine M Flegal, Margaret D Carroll, and Clifford L Johnson. Prevalence and trends in overweight among us children and adolescents, 1999-2000. *Jama*, 288(14):1728–1732, 2002.

[16] Youfa Wang and May A Beydoun. The obesity epidemic in the united statesgender, age, socioeconomic, racial/ethnic, and geographic characteristics: a systematic review and meta-regression analysis. *Epidemiologic reviews*, 29(1):6–28, 2007.

[17] Youfa Wang and Tim Lobstein. Worldwide trends in childhood overweight and obesity. *International Journal of Pediatric Obesity*, 1(1):11–25, 2006.

[18] Shaoyan Zhang, Christos Tjortjis, Xiaojun Zeng, Hong Qiao, Iain Buchan, and John Keane. Comparing data mining methods with logistic regression in childhood obesity prediction. *Information Systems Frontiers*, 11(4):449–460, 2009.