Using Machine Learning to Predict Obesity in High School Students

Zeyu Zheng
Maternal & Child Health Epidemiology
Texas Department of State Health Services
Austin, Texas, US
Zeyu.Zheng@dshs.texas.gov

Karen Ruggiero
Maternal & Child Health Epidemiology
Texas Department of State Health Services
Austin, Texas, US
Karen.Ruggiero@dshs.texas.gov

***Abstract*— Four enhanced machine learning models were used to predict obesity in high school students by focusing on both risk and protective factors: binary logistic regression; improved decision tree (IDT); weighted k-nearest neighbor (KNN); and artificial neural network (ANN). Nine health-related behaviors from the 2015 Youth Risk Behavior Surveillance System (YRBSS) for the state of Tennessee were used as model inputs. Results show that, compared to the logistic regression model that achieved 56.02% accuracy and 54.77% specificity, IDT, weighted KNN, and ANN each performed significantly better. The IDT model achieved 80.23% accuracy and 90.74% specificity, while the weighted KNN model achieved 88.82% accuracy and 93.44% specificity. The ANN model achieved 84.22% accuracy and 99.46% specificity. Implications and suggestions for slowing the increase in adolescent obesity are discussed.**

Keywords- obesity; improved decision tree; weighted KNN; artificial neural networks

I. INTRODUCTION

According to the Youth Risk Behavior Surveillance System (YRBSS) [1][2], the rate of obesity in high school students has risen dramatically from 1999 to 2015 in the United States (U.S.). Moreover, obesity is significantly more prevalent among adolescents in southern states than it is in northern states. Among southern states, Tennessee is the only state that has been listed as one of the top five states in both 2013 and 2015, with an alarming increase in the percentage of obese high school students from 16.9% in 2013 to 18.6% in 2015.

Adolescent obesity is the result of an imbalance between the calories they consume from food/beverage vs. the calories they expend to support normal growth and physical activity. Many previous studies have shown that such an imbalance can be greatly affected by health behaviors, such as energy intake, physical activity, and sedentary behavior [3][4]. In terms of energy intake, consuming large portions of energy-dense food/snacks, and drinking sugar-sweeten beverages at school, all increase the risk for obesity. Sedentary behavior, such as watching TV and/or internet browsing for extended time periods, also compound the risk for obesity. In contrast, moderate physical exercise on a routine basis is necessary for maintaining normal body weight, blood pressure, and bone strength among high school students. Childhood and adolescent obesity has serious health consequences, including high cholesterol and elevated blood pressure, in addition to increased risk for cardiovascular disease, asthma, and Type 2 diabetes. Therefore, it is critical to determine effective and efficient means by which to predict and prevent obesity in high school students.

Machine learning and structural equation modeling techniques are now being used to examine large amounts of data to identify patterns and relationships that would otherwise go undetected [5-10]. Some studies have applied these methods to build predictive models to understand the obesity problem. Particularly noteworthy is an early childhood obesity prediction model based on data recorded at birth, 6 weeks, 8 months, and 2 years, respectively [7]. The inputs included children's physical body measurements, such as height and weight gain. However, no health behaviors related to the obesity problem were considered. Other research investigated the association of sleep behavior with adolescent obesity using YRBSS 2007 and 2009 data [8]. However, this study only applied a logistic regression model, and no consistent pattern of association between sleep and obesity was detected for all adolescents. Further research is needed to better understand the factors underlying this association. Some studies have, in fact, applied more than one machine learning technique to predict obesity based on validated health behavior questionnaire data [9-10]. However, the sample size in these studies was relatively small (several hundred), thus calling for further research using larger sample sizes to produce more accurate prediction results.

Given the shortcomings of previous research, the present study applied multiple machine learning models to predict obesity among U.S. high school students; both risk and protective factors obtained from a large YRBSS sample were used. Specifically, we applied four enhanced machine learning classification models for data analysis: binary logistic regression, improved decision tree, weighted k-nearest neighbor, and artificial neural network models. The aim was to determine which model best predicted obesity in high school students. Findings could then be used to inform needed obesity prevention strategies.

978-1-5090-3050-7/17/$31.00 ©2017 IEEE

II. DATA MATERIALS

In this study, we used 2015 survey data from the biennial YRBSS for the state of Tennessee [1][2], obtained from a three-stage cluster sample design. The result is a representative sample of students in grades 9-12 attending public and private high schools in Tennessee. These data are publically available through the Centers for Disease Control and Prevention (CDC). Survey participation is voluntary with parental permission, and is anonymous. Because the study was based on de-identified, aggregated data from U.S. government public-use data sets, the study was exempt from requiring institutional review board approval.

A. Subjects

As reported previously [11], Tennessee is the only state that has been listed as one of the top five states for adolescent obesity in both 2013 and 2015, with an alarming increasing rate from 16.9% (+/-0.9%) in 2013 to 18.6% (+/-1.0%) in 2015. The total sample size is 5127 high school students in grades 9-12 from Tennessee, with complete data on the variables of interest. The sample consists of 2707(52.8%) female, 2420 (47.2%) male; 3239 (63.2%) White, 1107 (21.6%) Black or African American, 390 (7.6%) Hispanic/Latino, 391 (7.6%) other races; 1321 (25.8%) 9th grade, 1410 (27.5%) 10th grade, 1409 (27.5%) 11th grade, and 987 (19.3%) 12th grade. Subject demographic characteristics are shown in Table I.

TABLE I. SUBJECT DEMOGRAPHIC CHARACTERISTICS

Characteristic		Number	Percentage (Confidence Interval)
Gender	Female	2707	52.8% (+/-1.3%)
	Male	2420	47.2% (+/-1.3%)
Race	White	3239	63.2% (+/-1.3%)
	Black/African American	1107	21.6% (+/-1.1%)
	Hispanic/Latino	390	7.60% (+/-0.7%)
	All Other Races	391	7.60% (+/-0/7%)
Grade	9th	1321	25.8% (+/-1.2%)
	10th	1410	27.5% (+/-1.2%)
	11th	1409	27.5% (+/-1.2%)
	12th	987	19.3% (+/-1.1%)
Obesity	Yes	860	16.8% (+/-1.0%)
	No	4267	83.2% (+/-1.0%)

B. Obesity

Body mass index (BMI) is body weight in kilograms divided by height squared expressed in meters (kg/m^2). BMI is used to measure obesity. In the YRBSS, students are asked for their height and weight without shoes on, which are then used to calculate their BMI. Per the CDC, students with a BMI greater than or equal to the 95th percentile for their gender and age met the definition of obesity [8]. As presented in Table I, the number of obese students was 860 or 16.8%.

C. Health-Related Behaviors

As shown in Table II, we classified nine health-related behaviors into one of three categories: energy intake *(En)*, physical activity *(Pa)*, or sedentary behavior *(Se)*. *En* behavior includes eating fruits/vegetables/breakfast, and drinking soda or soft drinks in the past week. *Pa* includes engaging in physical activity for at least 60 minutes, and taking part in physical education classes in the past week. *Se* includes time spent on watching TV, using a computer for a non-school-related purpose or playing video games, and hours of sleep on school days. Using the recommendations of the American Academy of Pediatrics [8], we then dichotomized student responses as either *1 (yes)* or *0 (no)*.

III. CLASSIFICAITON MODELS

We applied four enhanced classification models: binary logistic regression, improved decision tree, weighted k-nearest neighbor, and artificial neural network models to predict obesity. The model input variables, also known as predictive variables, consisted of the nine health-related behaviors described in Table II. Student gender, race, and school grade were also served as predictive variables.

A. Training and Testing Dataset

For the training and testing datasets, we used 70% of our sample dataset as training dataset *S*, and the remaining 30% sample dataset as the testing dataset *T*. Note that the training and testing datasets were maintained consistently among the four models, so that the performance of each model was reliable and comparable.

In particular, suppose that we have a total of *n* training data points in the training dataset *S*. And for each data point denoted as d^i, we have a vector of predictive variables denoted as $x^i = (x_1^i, x_2^i, x_3^i, \ldots, x_Q^i)$ and a class label denoted as y^i ($y^i = 1$ if the subject met the definition of obesity; otherwise, $y^i = 0$).

TABLE II. HEALTH-RELATED BEHAVIOR DESCRIPTIONS AND CATEGORIES

Category	Description
Energy Intake 1 *En1*	Whether the student ate fruit or drank 100% fruit juices 3+ times/day during the past 7 days. *En1 = 1 (yes) or En1 = 0 (no)*
Energy Intake 2 *En2*	Whether the student ate vegetables 3+times/day during the past 7 days. *En2 = 1 (yes) or En2 = 0 (no)*
Energy Intake 3 *En3*	Whether the student drank soda 1+ times/day during past 7 days. *En3 = 1 (yes) or En3 = 0 (no)*
Energy Intake 4 *En4*	Whether the student ate breakfast on all 7 days during the past 7 days. *En4 = 1 (yes) or En4 = 0 (no)*
Physical Activity 1 *Pa1*	Whether the student were physically active doing any kind of physical activity that increased their heart rate and made them breathe hard some of the time for at least 60 minutes per day on all 7 past days. *Pa1 = 1 (yes) or Pa1 = 0 (no)*
Physical Activity 2 *Pa2*	Whether the students attended physical education classes on all 5 school days. *Pa2 = 1 (yes) or Pa2 = 0 (no)*
Sedentary Behavior 1 *Se1*	Whether the students watched TV for 3+ hours per day on an average school day. *Se1 = 1 (yes) or Se1 = 0 (no)*
Sedentary Behavior 2 *Se2*	Whether the students used a computer for something that was not school work or played video games for 3+ hours per day on an average school day. *Se2 = 1 (yes) or Se2 = 0 (no)*
Sedentary Behavior 3 *Se3*	Whether the student got 8+ hours sleep on an average school night. *Se3 = 1 (yes) or Se3 = 0 (no)*

B. Binary Logistic Regression Model

Logistic regression is a generalized linear regression model for predicting a categorical dependent variable [9]. We applied the logistic model formula to compute the probability of obesity y as a function of the predictive variables. If the student was obese, the conditional probability is denoted as $p(y = 1|x) = p(x)$, and the logistic model formula took the form shown as:

$\log[p(x)/1-p(x)]=\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_Q x_Q$. Note that $\beta= (\beta_0, \beta_1, \beta_2, \ldots, \beta_Q)$ represents the parameters of the logistic model formula. Specifically, we denote $\beta'= (\beta_1, \beta_2, \ldots, \beta_Q)$ for the sake of presentation simplicity. The likelihood will be denoted as:

$$L(\beta) = \prod_{i=1}^{n} p(x^i)^{y^i}(1 - p(x^i))^{1-y^i}, \quad (1)$$

and the log-likelihood will be:

$$\begin{aligned} L(\beta) &= \sum_{i=1}^{n} \log\left(1 - \mathrm{p}(x^i)\right) + \sum_{i=1}^{n} y^i \log\left(\frac{\mathrm{p}(x^i)}{1 - \mathrm{p}(x^i)}\right) \\ &= -\sum_{i=1}^{n} \log\left(1 + e^{\beta_0 + x^i \cdot \beta'}\right) \\ &\quad + \sum_{i=1}^{n} y^i (\beta_0 + x^i \cdot \beta'). \end{aligned} \quad (2)$$

We then determined the optimal parameters to maximize the log-likelihood so that the model could best fit the data points. We can differentiate the log-likelihood with respect to the parameters, and set the derivatives to zero, and then obtain the solution. Without the loss of generality, the derivative with respect to one component of the parameters, say β_q, will take the form of $\frac{\partial l}{\partial \beta_q} = \sum_{i=1}^{n}(y^i - p(x^i))\, x_q^i$, which has no closed-form solution when setting it to zero. Hence, we applied *Newton-Raphson's Method* for numerical optimization. Basically, *Newton-Raphson's Method* will start with a certain initial parameter setting, and then iteratively updates the parameters until it resembles the optimal solution. Typically, the update step is:

$$\begin{aligned} &\beta^{(m+1)} = \beta^{(m)} - H^{-1}\left(\beta^{(m)}\right)\nabla l\left(\beta^{(m)}\right), \\ &with\ \nabla l = \left(\frac{\partial l}{\partial \beta_0}, \frac{\partial l}{\partial \beta_1}, \ldots, \frac{\partial l}{\partial \beta_Q}\right), H_{ij} = \frac{\partial^2 l}{\partial \beta_i \beta_j} \end{aligned} \quad (3)$$

where m is the iteration step of the update process, ∇l and H are the Gradient vector and the Hessian matrix of the log-likelihood function respectively.

C. Improved Decision Tree Model

Note that the traditional decision tree model developed by Quinlan [10] has two main shortcomings. First, it often incurs bias towards predictive variables (also known as attribute) with multiple values, which could lead to inappropriate classifications. Second, traditional decision tree models involve a large amount of logarithm operations, which comes at a computational cost, especially when the training data set is large, as it is in this study. Therefore, we proposed an improved decision tree (IDT) model that can overcome the shortcomings of the traditional decision tree model, and operate in an effective and efficient way.

Given the training dataset S, our IDT model aims to create a decision tree by firstly calculating the *specific gain ratio* denoted as Gr for each attribute, and then splitting the

dataset *S* into subsets based on the attribute that has the maximal *Gr*, and generates a decision tree node containing this attribute. Our IDT model then recursively applies the same method to the generated node to build the tree. Note that the *specific gain ratio* of an attribute represents the measurement of uncertainty reduction of *S* after splitting *S* based on this attribute, with the adjustment of bias towards attributes that have more values.

In particular, the s*pecific gain ratio* of attribute x_q is denoted as $Gr(x_q)$. Then we will have:

$$Gr(x_q) = \frac{H(S) - \sum_{S'} p(S')H(S')}{\sum_{S'} -p(S')log(p(S'))}, \quad (4)$$

where *S'* is the subset of *S* when splitting *S* by the attribute x_q. And *p(S')* is the proportion of the number of data points in *S'* to the number of data points in *S*. And $H(S)$ is the entropy with $H(S) = -\sum_y p(y)\log(p(y))$, where y belongs to the class labels of *S*, and p(y) is the proportion of the number of data points with class label y to the number of data points in *S*. Moreover, the logarithm operations in (4) can be further simplified based on the McLaughlin formula in [12].

The pseudo-code of the improved decision tree model is shown as follows:

IDT(Dataset *S*, Attributes: {x_1,x_2, ... ,x_Q} , Class labels: {y}**)**

1. Create a node *t* for tree.
2. If all samples in *S* have class label "1", return *t* as a leaf node, with class label "1"; If all samples in *S* have class label "0", return *t* as a leaf node, with class label "0".
3. If Attributes is empty, return *t* as a leaf node with the most common value of class labels in *S*.
4. Otherwise:
   a) Let x_{q^*} be the attribute that best classifies *S*, with maximal *Gr* obtained by (4);
   b) Assign attribute x_{q^*} to node *t*;
   c) For each possible attribute value "*v*" in x_{q^*} do:
      - Add a new tree branch below t, corresponding to x_{q^*} = "*v*".
      - Let *Sa* be subset of *S* that has value "*v*" for x_{q^*}.
        If *Sa* is empty:
            Add leaf node with label of most common value of class labels in *S*.
        Else: Add sub-tree **IDT(***Sa*, Attributes\{x_{q^*} }, Class labels: {y}**)** .

D. Weighted K-Nearest Neighbor Model

K-nearest neighbor (KNN) model is an instance-based classifier, in which the classification of unknown instances can be implemented by relating the unknown to the known based on certain distance or similarity functions [10][13]. Note that KNN is a non-parametric method because it does not involve the estimation of parameters for an assumed function as in logistic regression. Generally, for a new data point $u = (u_1, u_2, \dots, u_Q)$ with an unknown class label, the KNN model will first determine the k nearest neighbors of this data point, and then assign the class label associated with the majority of the k nearest neighbors as the label. Usually, the neighbors will be found based on distance functions. There are two distance functions that are widely applied: Euclidean distance and Manhattan distance. In our KNN model, we applied weighted Euclidean distance based on the importance of the predictive variables, as shown in (5). Note that the weight associated with each predictive variable is proportional to the value indicated by the adjusted odds ratio of that variable related to obesity using logistic regression.

$$\sqrt{w_1(x_1^i - u_1)^2 + w_2(x_2^i - u_2)^2 + \cdots + w_Q(x_Q^i - u_Q)^2}. \quad (5)$$

The pseudo code of the weighted KNN model is shown as follows:

Input: *k* as the desired number of nearest neighbors; *S* as the set of training data points with known class labels $\{(x^1, y^1), (x^2, y^2), \dots, (x^n, y^n)\}$; *u* as new data point with unknown class label.

1. For each data point in *S*, compute the distance between that data point and *u* based on (5);
2. Sort all data points in *S* in descending order based on the distances obtained from the above step;
3. Select the first *k* data points in the sorting list, and mark them as *k* nearest neighbors of *u;*
4. Find out the majority class label of k selected data points, and assign such class label to *u*.

E. Artificial Neural Network Model

An artificial neural network (ANN) is a computational model that emulates the biological neural system [9] to conduct comprehensive data analysis. In this study, we applied a Multi-layer Perceptron Neural Networks model, which maps a set of input data onto a set of appropriate output data through three layers of neurons. The three layers are the input layer, hidden layer, and output layer. Further, there are multiple connections between adjacent layers with a weight assigned to each connection. Specifically, the input layer consists of neurons corresponding to predictive variables (x_1,x_2, ... ,x_Q), which are connected to neurons in the hidden layer. Each neuron in the hidden layer will sum the data received from the input layer through weighted connections, and then modify the sum by a non-linear transfer function before passing the sum to the output layer.

To appropriately train the ANN model, we used the back-propagation algorithm, and applied the hyperbolic-tangent sigmoid transfer function. Note that for the model training process, it is essential to appropriately set the learning rate

and the momentum parameters so that the model can be well trained within a proper amount of time. In our study, we set the learning rate and the momentum as 0.1 and 0.7 respectively [9].

IV. DISCUSSION

A. Observation

Table III presents the prevalence of health-related behaviors according to subject demographic characteristics. Statistical Analysis System (SAS) was used to examine differences across demographic group, with statistical significance set at 0.05.

In terms of the impact of gender, male students were more likely than female students to engage in energy intake *(En)* behavior. In particular, male (16.4%) students were significantly more likely than female (12.5%) students to eat fruit or drink 100% fruit juices 3+ times/day during the past 7 days ($z=-4.04$, $P<0.05$). Also, the prevalence of eating vegetables 3+times/day during the past 7 days was significantly greater among male (11.1%) than female (8.7%) students ($z=-2.88$, $P<0.005$). The prevalence of drinking soda 1+ times/day during past 7 days was significantly greater among male (31.5%) than female (26.0%) students ($z=-4.37$, $P<0.05$). Similarly, there were more male (39.4%) than female (31.4%) students that ate breakfast on all 7 days during the past week ($z=-6.03$, $P<0.05$).

As for physical activity *(Pa)*, significantly more male (34.8%) than female (17.1%) students were physically active engaging in exercise that increased their heart rate and made them breathe hard some of the time for at least 60 minutes per day on all 7 past days ($z=-14.55$, $P<0.05$). Also significantly more male (25.3%) than female (19.6%) students attended physical education classes on all 5 school days ($z=-4.87$, $P<0.05$).

In terms of sedentary behavior *(Se)*, a high percentage of both male and female students watched TV (30.1%) for 3+ hours per day on an average school day. Both male and female students also spent a large amount of time using a computer for something that was not school work and playing video games for 3+ hours per day on an average school day. Perhaps this is due, in part, to increased Internet and Wi-Fi accessibility on school campuses and in family households. Moreover, only 28.6% of male students and 27.2% of female students got 8+ hours sleep on an average school night, which may due to heavy school homework/study time and/or excessive TV watching and computer usage.

Table III also presents the prevalence of these health-related behaviors according to whether students were defined as obese vs. non-obese. Significantly more obese (31.4%) than non-obese (28.0%) students drank soda 1+ times/day during past 7 days ($z=2.0074$, $P<0.05$). Additionally, significantly more obese (46.5%) than non-obese (40%) students used a computer for something other than school work or played video games for 3+ hours per day on an average school day ($z=3.5512$, $P<0.005$).

B. Performance Evaluation

To evaluate the accuracy and specificity of each of the four machine learning models in predicting obesity, we applied a 10-fold cross-validation technique to perform an unbiased performance evaluation using Rapid-Miner software [10]. The model performance was adjusted by gender, age, and grade. For the logistic regression model, the accuracy rate was 56.02%, and specificity was 54.77%. For the IDT model, the accuracy rate was 80.23%, and specificity was 90.74%. Whereas the weighted KNN model had an accuracy rate of 88.82% and specificity at 93.44%, the ANN model had an accuracy rate of 84.22%, with 99.46% specificity. Furthermore, our models indicated that engaging in frequent physical activity and having breakfast everyday were protective factors that significantly reduce the risk for obesity. And frequent consumption of sugar-sweeten beverages and excessive computer use significantly increase the risk for obesity among high school students.

V. CONCLUSION

The prediction of health conditions and diseases using machine learning can be challenging, but it can significantly increase the analytical accuracy and specificity. Applying effective and efficient machine learning techniques for data analysis can greatly reduce the cost and time constraints involved. In this study, we compared four machine learning models to predict obesity in high school students in Tennessee using their health-related risk and protective factors as reported on the YRBSS. The results show that, compared to logistic regression, IDT, weighted KNN, and ANN models yielded better performance in classifying and predicting the obesity.

Although this study examined adolescent obesity in Tennessee, the machine learning models that have been built and presented in this paper can be applied and used to further our understanding of obesity in other southern states, where the rate of obesity among high school students has also increased. The findings of the present study suggest that healthy lifestyle habits, including healthy eating and regular physical activity, can lower the risk of obesity and related health conditions among high school students. To this end, students need to have access to healthy foods at school by implementing higher nutrition standards, and providing incentives to businesses located within their school districts to limit the sale of sugar-sweeten beverages to students. To slow the increase in adolescent obesity, schools must also set a higher standard for physical activity participation among

students, while also setting limits on their non-school-related computer usage at school.

In the future, we will expand our research by integrating into our prediction models additional demographic and social factors that may affect the prevalence of obesity. For instance, family demographics (e.g., parental occupation, education, marital status) may play an important role in a student's physical and mental health, which, in turn, might affect that student's health behaviors related to obesity. Moreover, advertisements displaying fast foods and unhealthy snacks may encourage students to eat more high-calorie foods, which may increase the risk for adolescent obesity. By examining these additional factors, we may be able to better understand and predict obesity among high school students in future studies.

REFERENCES

[1] D.K. Eaton, L. Kann, S. Kinchen, S. Shanklin, J. Ross, J. Hawkins, W.A. Harris, R. Lowry, T. McManus, D. Chyen, C. Lim, N.D. Brener, H. Wechsler, "Yourth Risk Behavior Surveillance – United States 2007", in Morbidity and Mortality Weekly Report, Surveillance Summaries 2002.

[2] https://www.cdc.gov/healthyyouth/data/yrbs/data.htm

[3] N.S. Marshall, N. Glozier, R.R. Grunstein, "Is Sleep Duration Related to Obesity? A Critical Review of the Epidemiological Evidence", in Sleep Medicine Reviews 2008, vol. 12, no. 4, pp. 289-298.

[4] American Academy of Pediatrics, Committee on Public Education, "American Academy of Pediatrics: Children, adolescents, and television", Pediatrics, vol. 107, no. 2, pp. 423-426, 2001.

[5] J. Yang, H. Gu, X. Jiang, Q. Huang, X. Hu, X. Shen, "Walking in the PPI Network to Identify the Risky SNP of Osteoporosis with Decision Tree Algorithm", in IEEE International Conference on Bioinformatiocs and Biomedicine (BIBM 2016), pp. 1283-1287.

[6] J. Zhao, T. He, X. Hu, Y. Wang, X. Shen, M. Fang, J. Yuan, "A Novel Disease Gene Prediction Method Based on PPI Network", in IEEE International Conference on Bioinformatiocs and Biomedicine (IEEE BIBM 2014), pp. 311-314.

[7] S. Zhang, C. Tjortjis, X. Zeng, H. Qiao, I. Buchan, J. Keane, "Comparing Data Mining Methods with Logistic Regression in Childhood Obesity Prediction", in Information Systems Frontiers 2009, pp. 449-460.

[8] R. Lowry, D. K. Eaton, K. Foti, L. McKnight-Eily, G. Perry, and D. A. Galuska, "Association of Sleep Duration with Obesity among US High School Students", in Journal of Obesity, Vol. 2012.

[9] S. T. Heydari, S. M. Ayatollahi, N. Zare, "Comparison of Artificial Neural Networks with Logistic Regression for Detection of Obesity", in Journal of Medical System 2012.

[10] T. Sivaranjani, "Comparative Study on Obesity Based on ID3 and KNN", in International Journal of Advance Research in Computer Science and Management Studies, Vol. 2, Issue 9, September 2014, pp. 389-396.

[11] http://stateofobesity.org/high-school-obesity/

[12] W. Gao, Y. Dong, K. Li, "The Reseach and Application of Improved Decision Tree Algorithm in University Performance Analysis", in Proceedings of the 2nd International Conference on Computer Science and Electronics Enginnering (ICCSEE 2013), pp. 97-100.

[13] M. Kuhkan, "A Method to Improve the Accuracy of K-Nearest Neighbor Algorithm", in International Journal of Computer Engineering and Information Technology, Vol. 8, No. 6, June 2016, pp. 90-95.

TABLE III. PREVALENCE OF HEALTH-RELATED BEHAVIORS BY DEMOGRAPHIC CHARACTERISTICS

Demographic Characteristic			**Health-Related Behaviors**								
			En1	***En2***	***En3***	***En4***	***Pa1***	***Pa2***	***Se1***	***Se2***	***Se3***
Gender	Female	n	337	235	703	849	462	530	816	1080	737
		%	12.5% (+/- 1.3%)*	8.7% (+/- 1.1%)*	26.0% (+/- 1.6%)*	31.4% (+/- 1.7%)*	17.1% (+/- 1.4%)*	19.6% (+/- 1.5%)*	30.1% (+/- 1.7%)	40.0% (+/- 1.8%)	27.2% (+/- 1.7%)
	Male	n	397	268	762	954	842	611	728	1026	691
		%	16.4% (+/- 1.5%)*	11.1% (+/- 1.3%)*	31.5% (+/- 1.8%)*	39.4% (+/- 1.9%)*	34.8% (+/- 1.9%)*	25.3% (+/- 1.7%)*	30.1% (+/- 1.8%)	42.4% (+/- 1.9%)	28.6% (+/- 1.8%)
Race	White	n	357	294	1023	1214	855	729	849	1292	917
		%	11.0% (+/-1.1%)	9.1% (+/- 1.0%)	31.6% (+/- 1.6%)	37.5% (+/- 1.6%)	26.4% (+/- 1.5%)	22.5% (+/- 1.4%)	26.2% (+/- 1.5%)	39.9% (+/- 1.7%)	28.3% (+/- 1.5%)
	Black/ African American	n	233	111	271	314	268	234	464	454	303
		%	21.1% (+/-2.3%)	10.0% (+/- 1.9%)	24.5% (+/- 2.6%)	28.4% (+/- 2.7%)	24.2% (+/- 2.6%)	21.1% (+/- 2.5%)	41.9% (+/- 2.9%)	41.0% (+/- 2.9%)	27.4% (+/- 2.7%)
	Hispanic/ Latino	n	72	40	82	125	77	84	119	170	110
		%	18.5% (+/-4.1%)	10.26 (+/- 3.4%)	21.0% (+/- 4.3%)	32.1% (+/- 4.7%)	19.7% (+/- 4.2%)	21.5% (+/- 4.3%)	30.5% (+/- 4.7%)	43.6% (+/- 4.9%)	28.2 (+/- 4.6%)
	Other Races	n	72	58	89	150	104	94	112	190	98
		%	18.4% (+/-4.1%)	14.8% (+/- 3.8%)	22.8% (+/- 4.4%)	38.4% (+/- 4.9%)	26.6% (+/- 4.5%)	24.0% (+/- 4.4%)	28.6% (+/- 4.6%)	48.6% (+/- 4.9%)	25.1% (+/- 4.5%)
Grade	9th	n	202	126	395	478	365	523	432	562	463
		%	15.3% (+/-2.0%)	9.5% (+/- 1.7%)	29.9% (+/- 2.5%)	36.2% (+/- 2.6%)	27.6% (+/- 2.4%)	39.6% (+/- 2.6%)	32.7% (+/- 2.5%)	42.5 (+/- 2.6%)	35.1% (+/- 2.6%)
	10th	n	202	129	375	542	384	295	402	617	411
		%	14.3% (+/-1.9%)	9.2% (+/- 1.6%)	26.6% (+/- 2.3%)	38.4% (+/- 2.5%)	27.2% (+/- 2.3%)	20.9% (+/- 2.2%)	28.5% (+/- 2.4%)	43.8% (+/- 2.6%)	29.2% (+/- 2.4%)
	11th	n	202	143	400	489	338	183	412	536	368
		%	14.3% (+/-1.9%)	10.2% (+/- 1.6%)	28.4% (+/- 2.4%)	34.7% (+/- 2.5%)	24.0% (+/- 2.3%)	13.0% (+/- 1.8%)	29.2% (+/- 2.4%)	38.0% (+/- 2.5%)	26.1% (+/- 2.3%)
	12th	n	128	105	295	294	217	140	298	391	186
		%	13.0% (+/-2.2%)	10.6% (+/- 2.0%)	29.9% (+/- 2.9%)	29.8% (+/- 2.9%)	22.0% (+/- 2.6%)	14.2% (+/- 2.3%)	30.2% (+/- 2.9%)	39.6% (+/- 3.0%)	18.8% (+/- 2.5%)
Obesity	Yes	n	136	110	270	296	210	200	282	400	232
		%	15.8% (+/-2.5%)	12.8% (+/- 2.4%)*	31.4% (+/- 3.1%)*	34.4% (+/- 3.2%)	24.4% (+/- 2.9%)	23.3% (+/- 2.9%)	32.8% (+/- 3.2%)	46.5% (+/- 3.3%)*	27.0% (+/- 3.0%)
	No	n	598	393	1195	1507	1094	941	1262	1706	1196
		%	14.0% (+/-1.0%)	9.2% (+/- 0.9%)*	28.0% (+/- 1.3%)*	35.3% (+/- 1.4%)	25.6% (+/- 1.3%)	22.1% (+/- 1.2%)	29.6% (+/- 1.3%)	40.0% (+/- 1.4%)*	28.0% (+/- 1.3%)

*P<0.05