EXERCISE AND QUALITY OF LIFE Research article
Volume 3, No. 1, 2011, 11-21
UDC 616-057.875:796.035]:004.9

AN AUTOMATED IDENTIFICATION OF INDIVIDUALS AT HEALTH RISK BASED ON DEMOGRAPHIC CHARACTERISTICS AND SELF-REPORTED PERCEPTIONS

Dejan Magoc*
Department of Health Studies
Eastern Illinois University

Tanja Magoc, Ph.D.
Center for Bioinformatics and Computational Biology
University of Maryland

Joe Tomaka, Ph.D.
Department of Public Health
University of Texas at El Paso

Abstract

The risks of developing diabetes, high blood pressure, and cardiovascular disease could be reduced by increasing the number of individuals receiving adequate levels of physical activity (PA). Centers for Diseases Control and Prevention (CDC) has reported that about 30% of Americans do not engage in any PA and about 40% engage in some levels of PA, but still not meeting the recommended levels defined by the American College of Sports Medicine (ACSM). Studies have shown that the greatest declines in PA occur during the transitions from high school to college and beyond. Thus, it is important to identify students at young age that are at health risk due to lack of PA, so that specific steps could be taken toward helping these individuals develop a healthier lifestyle. We used data on 100 college students to develop a preliminary computer program (using a backpropagation multilayer neural network approach) to automatically identify individuals at risk of being not sufficiently physically active. Besides various types of demographic variables, data included information on the association between students' self-reported levels of PA and Social Cognitive Theory (SCT) constructs (e.g., self-efficacy, self-regulation, social support, expectations), as predictors of participation in PA. The results of this study indicated that the backpropagation multilayer neural network identified and classified individuals at risk of being not sufficiently physically active into right categories (at-risk individuals or not at-risk individuals) 77% of the time. Collecting additional data points that contain more at-risk individuals will improve the neural network's prediction of at-risk individuals.

Keywords: automated identification, physical activity, college students

* Corresponding author. Department of Health Studies, Eastern Illinois University, Lantz Charleston, IL 61920, e-mail: dejanmagoc@yahoo.com

Introduction

Lack of physical activity (PA) in general population has become a major public health concern (Petosa, Suminski, & Hortz, 2003). Even though a relatively large number of people report participating in some PA, majority of population is not sufficiently physically active to prevent diseases such as hypertension, diabetes, and cardiovascular disease. According to the American Heart Association (AHA) and the American College of Sports Medicine (ACSM), at least 30 minutes of moderate PA (e.g., walking) five days per week or 20 minutes of vigorous PA (e.g., running) three days per week is required to keep a healthy living style (Haskell et al., 2007). However, studies show that only about 30% of Americans satisfy the minimum exercise requirements and another 30% does not exercise at all (CDC, 2005).

Research has also shown that levels of PA dramatrically decrease from high school to college years and beyond (Rovniak, Eileen, & Winett, 2002). Thus, it is of high importance to motivate college students to engage in regular physical activity. Unfortunately, many PA events and promotions on college campuses, such as intramural leagues and sport-specific clubs, tend to attract students who are already physically active. In order to attract physically inactive or not sufficiently active students, it is important to identify these individuals quickly and efficiently. After identifying students at health risk due to physical inactivity, such students could be approached by targeted interventions.

Studies have shown consistently correlations between PA levels and demographic characteristics, such as race and gender (Pratt, Macera, & Blanton, 1999; Mouton, Calmbach, & Dhanda, 2000; Dunn & Wang, 2003). Research has also shown that other factors, such as self-motivation , previous physical activity engagement, perception of importance to exercise, perception of current physical and psychological health and support from friends and family to exercise are highly correlated to the amount of PA an individual performs (Pratt et al., 1999; Magoc, Tomaka, & Thompson, 2010). These factors could be used to identify individuals at risk of being physically inactive at early stage, which could prevent students from becoming sedentary.

Although such data could be assessed relatively quickly, the review and evaluation process for each individual would take considerable time and effort to perform. Instead of manually examining all responses, we propose a system of computerized questionnaires with automatic and practically instantaneous analysis and presentation of results to immediately identify students at risk of not being insufficiently physically active to prevent negative health outcomes.

This paper presents the preliminary results regarding the application of one such tool. Specfically, using data collected from 100 college students about their weekly physical activities as well as their demographic characteristics, the machine learning algorithm (see the section on machine learning) was used to build a model that predict whether an individual is likely to be at risk of being physically inactive. The predicted output for a particular individual was compared to the weekly physical activity information entered by that individual to assess the accuracy of the prediction.

Machine Learning

Machine learning (see Russell & Norvig, 2010; Tan, Steinbach, & Kumar, 2006) is a branch of computer science that aims at developing computer programs that simulate human reasoning and can therefore “replace” humans in numerous tasks including data analysis and

decision making. A major focus of machine learning is to automatically learn to recognize patterns and infer relationships among different variables. Based on the learned relationships, these computer programs are able to make decisions that are equivalent to decisions that would be made by humans in a given situation.

A machine learning algorithm consists of two phases: the training phase and the application phase. In the training phase, the algorithm learns patterns and relationships in a given data set, while in the application phase, a decision for a new instance (e.g., a new individual) is made.

The training phase could be performed in a supervised or an unsupervised mode. In either option, a data set is provided for training. A data set consists of numerous (tens to thousands) data points. Each data point corresponds to one instance and contains a value for each variable used to make the final decision. In the supervised learning, the final decision is also provided in the training sample. Thus, for supervised learning, we need to know the correct decision in all training data points. The known decisions are used to reduce the error in the machine learning algorithm by aiding the algorithm to learn patterns and relationships that yield particular decision. On the other hand, the unsupervised learning does not require knowledge of correct decisions.

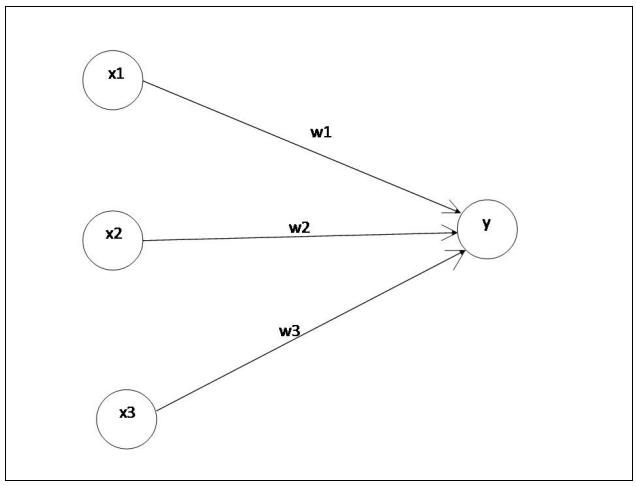
One of the main applications of machine learning is in classification problems. Given an instance, a machine learning program classifies this instance in one of several possible groups. An instance that is being classified consists of a value for each variable used to make the decision, but the correct class is unknown. The classification is based on the patterns and relationships previously observed in the training phase.

To test the accuracy of a machine learning algorithm, the algorithm is usually trained on a data set and tested on a different set of data, both of which have known classes for each data point, to avoid bias that would result in the algorithm performing well only on the data set used for training. One common method to split available data points in the training and testing data sets is to randomly split all data points into five groups, and perform 5-fold cross validation. This method uses all possible combinations of four out of five created sets for training and one set for testing. Thus, five tests are performed, one for each of five sets to be the testing set, and the results of the five tests are combined to determine the accuracy of the algorithm.

Neural Network

A neural network (NN) is a type of a machine learning algorithm that is designed to imitate the actions of the human neural system (see Russell & Norvig, 2010; Tan et al., 2006). A NN is represented as a directed weighted graph where nodes simulate human neural cells (neurons), and directed edges simulate the links between neurons (axons). The strength of the signal transferred between neurons determines the action of a human. This signal strength is simulated by the weights on the edges in an NN (Figure 1). The basic task of a NN is to learn these weights in order to yield accurate results when applied to real life data.

Figure1. A simple Neural Network with Three Input Nodes (x1,x2,x3) and One Output Node (y)

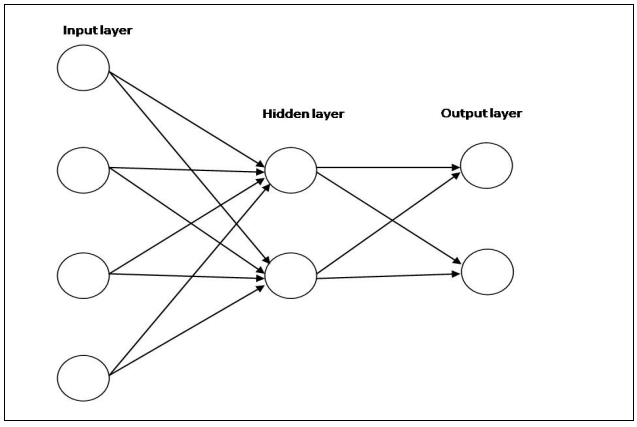


The simplest NN is called the single-layer NN and consists of only input and output layers of nodes. The inputs nodes take the values of variables used to make a decision, and the output node contains the decision. The decision is made by combining the values of input nodes and the weights on the edges.

Each data point from the training data set is processed by the NN, one by one, and after each data point is processed, the error is calculated and the necessary adjustments to weights are made. When all data points are processed, the same process is repeated over and over until a NN, with a satisfying (i.e., very low) error in classification, is built or until a predefined number of iterations is reached (usually several hundred iterations).

Single-layer neural networks are good classifiers in simple cases. However, more complex multilayer networks are much more powerful than neural networks that contain only input and output layers. Multilayer neural networks contain one or more hidden layers (Figure 2).

Figure 2. A Fully Connected Multilayer Neural Network



Similarly to a single-layer NN, a multi-layer NN takes values of variables of a data point as inputs (i.e., the input layer), aggregates these values by calculating the weighted sum, and applies a function, such as *sign* or *sigmoid* function, to produce the values of the next level of nodes (i.e., the hidden layer). Furthermore, the hidden layer of nodes acts as the input layer for the next level of nodes (i.e., a next hidden layer or the output layer) until the values of the output nodes are calculated. In a multi-layer network, the links between nodes can go either from a lower layer to a higher layer (input being the lowest layer and output the highest layer), which is the case in feed-forward networks, or can be directed from a node to a node at a higher, same, or lower level, which is the case in recurrent networks.

Similarly to a single-layer NN, a multi-layer NN learning algorithm works by minimizing the error. The error could be contributed equally to each node or a backpropagation algorithm could be used to more precisely determine the impact of each node to misclassification and therefore allow different levels of weights adjustment on edges.

Neural Network to Identify Individuals at Health Risk

Currently available data obtained in numerous studies about PA and the factors that influence the level of an individual's PA are characterized by still unknown probability distributions, not precisely known dependencies among different data variables, and unclear level to which each variable impacts an individual's readiness and commitment to exercise. Due to all uncertainties about the relationships between PA and features highly correlated with the level of PA, machine learning is well suited approach to capture important relationships among features present in the existing data, which are then used to identify individuals at health risk using only a few demographics and self-reported characteristics of individuals.

Method

Participants and Setting

The participants in this study were 100 part- or full-time currently enrolled male and female students from a large southwestern university in the U.S. with a large Hispanic enrollment. All participants were recruited through classroom settings, and all completed the cross-sectional survey.

Measures

Demographic variables - included self-reported gender, race/ethnicity, class, height and weight. In addition, participants self-reported perception of their current physical and psychological health.

International Physical Activity Questionnaire (IPAQ; Booth, 2000) - a self-reporting measurement of the level of PA the individual performed during the last seven days. It asked participants to record the number of sessions and the average duration of an exercise session for vigorous and moderate activities.

Self-Efficacy for Exercise Behavior Scale (Sallis, Pinski, Grossman, Patterson, & Nader,1988) - consisted of 12 questions measuring the individual's readiness to overcome obstacles (e.g., tiredness, large amount of work, not accomplishing set physical activity goals) in order to exercise. Moreover, the participants reported the *importance of setting aside time for exercise* in their schedules and following the set goals.

The Family and Friend Support for Exercise Habits Scales (Sallis, Grossman, Pinski, Patterson, & Nader, 1987) - measured the support and motivation to exercise that participants received from family and friends, including exercising together with another individual, receiving reminders from an individual to exercise, or having a discussion about exercising.

Procedures

Participants for this study were largely recruited through regular classroom meetings and activities, with most receiving extra course credit for participation. All participants completed informed consent forms prior to completing the questionnaires. It took approximately 20 minutes for participants to complete the questionnaire.

Results

Descriptive Analysis

A slightly more females (59%) participated in the study than males. The majority of participants were Hispanics (82%). Because of the high proportion of participants being Hispanic, for the purpose of the study, all participants were classified either as Hispanic or non-Hispanic, therefore, not making a distinction among non-Hispanic participants, which included Caucasians, African Americans, Native Hawaiians, American Indians, and Asians.

Most participants recorded their major being health or sports related studies (73%). The majority of participants self-reported their physical health to be good or fair (48% and 31%, respectively) and only 13% self-reported their physical health to be excellent. Majority of the participants rated their psychological health as good or excellent (54% and 23%, respectively).

Majority of participants recorded at least a medium level for readiness to overcome obstacles in order to exercise as well as for motivation and support received from friends and family members. In addition, 21% of the sample self-reported the low importance to exercise.

Almost half of the participants (41%) failed to meet the recommended levels of PA, with a higher percentage of females being not sufficiently physically active than males (30% and 11%, respectively). In addition, 56% of the sample was overweight, including 26% of those being classified as obese.

The collected information was used to build an automated predictor of students at risk of being not sufficiently physically active by applying a machine learning algorithm.

Primary Analysis

Using the information collected from 100 students, we trained a backpropagation NN with eight input variables, one hidden layer with 13 nodes, and the output layer with two nodes. The input variables included the following:

- Gender: this variable could take two values {male, female}.
- Hispanic: this variable could take two values {yes, no} describing whether the individual is Hispanic or not, respectively. This distinction was made since studies have shown that Hispanic population exhibits different attitudes towards physical activities from those exhibited by, for example, Caucasian people (Pratt, 1999). Since not enough data were available to train the NN on different ethnicities among non-Hispanics, further distinction among ethnicities was not included.

- Major: this variable could take two values {sport related, not sport related}. This distinction was made because it is expected that students majoring in sport or health related studies are more likely to be aware of PA importance and therefore exercise more than their peers who major in other disciplines.
- Physical health: this variable was a self-reported individual's perception, and could take any of the five values {excellent, good, fair, poor, very poor}.
- Psychological health: this variable was a self-reported individual's perception, and could take any of the five values {excellent, good, fair, poor, very poor}.
- Self-efficacy: this variable is a summary of an individual's answers to the 12 questions on the self-efficacy scale assessment. Since each question allowed participants to express the level of self-efficacy in the range 1-5 (1 meaning 'low' and 5 meaning 'high'), the values were averaged, and the score above 4.00 was reported as "really high", the score between 3.00 and 4.00 as "high", etc. The self-efficacy variable could take one of five values {very high, high, medium, low, very low}.
- Importance of exercise: this variable represented how important it was for an individual to make time for exercise in his/her schedule and to accomplish the scheduled PA goals. The variable could take one of five values {very high, high, medium, low, very low}.
- Support: this variable is a summary of the individual's answers to the exercise habits scale assessment. It was created similarly to the self-efficacy variable and could take one of the five values {very high, high, medium, low, very low}.

The output layer of the NN consisted of two nodes {risk, no risk}. Only one of these two nodes is *on* as a result of applying the NN to data collected from a new individual. Depending on which node is *on*, the person is classified to be or not to be at health risk based on the self-reported characteristics.

A neural network was trained using free software package *weka* (Hall, Frank, Holmes, Pfahringer, Reutemann, & Witten, 2009). Since all collected data was represented as non-numeric data to *weka*, each variable that could take more than two values was represented by multiple input nodes, one for each value the variable could take. For example, the variable *importance of exercise* could take one of five values (very high, high, medium, low, and very low), thus five input nodes were designed for this particular variable. However, even though some variables could take one of five values, not all five values have shown up in the collected data sample, and thus, less nodes were used to represent such a variable. For example, no one reported his/her physical health to be *very poor*, thus the *physical health* variable used only four nodes in the developed NN. Variables with multiple nodes, such as physical health and importance of exercise, would have only one of their nodes set *on* in each training or testing sample.

Furthermore, if a variable could take exactly two values, this variable was represented by only one node. This node was either *on* or *off*, representing two different values that the variable could take. With this representation, the input layer of the NN developed for the collected data sample consists of 25 nodes.

The hidden layer was automatically generated by *weka*. By default, *weka* generates (number of attributes + number of classes)/2 nodes in the hidden layer, which resulted in (25 input nodes + 2 output nodes)/2=13.5 for our data sample, so 13 nodes were generated in the hidden layer.

A fully connected NN (i.e., a NN where each node from a previous layer is connected to each node of the next layer (see Figure 2) was trained using a backpropagation algorithm.

The learning rate in the NN was initially set at 0.2 and decreased throughout training cycles. Total of 500 iterations were performed to train this NN, which took 0.47 seconds for the given data set.

To validate the accuracy of the developed NN, we performed 5-fold cross-validation with 80 data points used for training and 20 data points for testing. Combining results of all five runs in cross-validation, the NN classified correctly 77% of test data. Table 1 shows the classification of individuals into "at risk" and "not at risk" categories. Out of 41 individuals who should have been classified "at risk", 18 individuals were classified correctly. Similarly, out of 59 individuals who should have been classified "not at risk", only one individual was classified incorrectly.

Table 1
Classification of Individuals "at risk" and "not at risk" Categories

	Individuals classified "at risk"	Individuals classified "NOT at risk"
Individuals "at risk"	18	23
Individuals "NOT at risk"	1	58

Discussion

Even though there are two possible types for misclassification (i.e., an at-risk individual classified as not at-risk, and a not at-risk individual classified as at-risk individual), most misclassified data points corresponded to individuals that should have been classified as individuals at risk, but were classified as not being at risk. This is encouraging for two reasons. First, data in this study were collected from somewhat not traditional group of individuals (more individuals in this group met the minimum physical activity requirements, which is usually not the case). Thus, this particular data sample might not contain enough data points to train the NN to correctly identify all at-risk individuals. We suspect that some training samples in the 5-fold cross validation did not contain enough at-risk individuals to train the NN. We expect that collecting additional data points that contain more at-risk individuals will improve the NN's prediction of at-risk individuals.

Second, the current NN is able to identify almost perfectly the individuals not at risk. Recall that the main reason for developing this automatic identification of at-risk individuals is to target at-risk individuals by physical activities on college campuses. While the current NN does not identify all at-risk individuals and will not contribute to targeting all at-risk individuals, it will ensure that physically active individuals are not targeted and therefore, no resources will be "wasted" on not at-risk individuals. Here, by resources, we consider items such as gift certificates for participations in PA studies, individual attention by personal trainers conducting a research, time spent on providing health and PA seminars, promotional offers through gym memberships for inactive students, etc., which are often taken by individuals who do not clearly satisfy the requirements of participation such studies, but rather take the advantage of promotional offers to continue doing what they have already been doing. Thus, even though not all inactive students would be reached by the currently developed NN, it is a starting point to increase PA participation of students at health risk.

Expected Improvements in Neural Network Predictions

Even though the current NN predictions are acceptable and will allow college campuses to target a large amount of students at health risk due to physical inactivity, further improvements are possible. We are currently collecting more data from students at different collegiate institutions. The following factors are expected to contribute towards better NN training, and therefore an improved prediction rate.

First, a larger amount of data allows easier detection of patterns and relationships among variables in a NN. Thus, collecting more data will allow for better training of the NN.

Second, a majority of the current data sample is of Hispanic origin. All the other ethnicities are classified as non-Hispanics even though there are differences in PA attitudes among, for example, Caucasian and African American individuals. However, currently these two ethnic groups are considered as the same group. The newly collected data will add to wider variety of demographic characteristics, which would aid predictions in the subgroups that are not adequately represented in our current sample.

Third, we believe that students who do not major in health or sport related studies are more likely to fall into the at-risk category than students majoring in health and sports studies. Since majority of the current sample contained students that are not expected to be at health risk, adding data points from students that are at risk will help train NN. The newly collected data will contain more information from students at risk, and will allow a better training of NN in at risk cases.

Fourth, a more objective assessment of the physical and psychological health of each individual is preferable. The current sample relies on self-evaluation of one's health. However, since not every person has same goals, the reported physical and psychological health might not be consistent. For example, if a person weighted 220lbs a year ago, and currently weights 200lbs, this person might feel good about his/her physical health even though this person might still be overweight. While reports of this type are not expected to happen often and should not drastically harm machine learning when a large amount of data is available, this type of report might be harmful in our relatively small sample. A more objective assessment of the PA could be obtained by combining facts such as the height, weight, and body mass index with the individual's subjective perception.

Finally, a more objective evaluation of physical activities is needed in order to correctly determine whether each person satisfies the minimum PA requirements as set by AHA and ACSM. The current classification is obtained based on an individual's PA within the last seven days and his/her subjective perception on the intensity of the activity (i.e., whether an exercise is moderate or vigorous intensity). For example, while running at 6 mph might be considered as vigorous PA by a person who does not exercise often, it would not be considered vigorous intensity by an individual who satisfies the minimum PA requirements. Thus, collecting facts (e.g., the length, time, and incline level of a run) along with a person's perception of the intensity would provide a more objective evaluation of physical intensity of an activity. Moreover, collecting the information over at least a few weeks would show the consistency of the exercise rather than relying on how physically active a person was in only seven days since inactive persons generally greatly fluctuate in the weekly amount of exercise.

Recommendations for Future Research

The majority of general population is prone to health diseases that could be easily prevented by regular PA. While most people are aware that PA is important for their health, many individuals are not aware of how much PA is necessary to keep their bodies in good health. People live with the idea that some exercise is better than none (and are therefore

satisfied by finding time for little exercise), but only a small percent of individuals is aware that they do not exercise enough to reduce the chance of diseases, such as heart attack and high blood pressure.

The amount of PA drastically drops from high school to college and beyond. Therefore, it is of high importance to target the first year college students in promoting PA and spreading awareness of its importance. Since physically active students are the ones who usually respond to PA promotions on collegiate campuses, it is important to identify students who are under the risk of inactivity and target these particular individuals in PA studies and promotion programs.

With larger amount of data, we plan to enhance the classifier to allow three output classes (sufficiently physically active, not sufficiently physically active, and not physically active at all) rather than only two classes as presented in this study. The first class would include individuals that are clearly physically active on regular bases and are reaching the minimum PA requirements. The second class would include individuals that are physically active to certain extend, but not enough to meet the minimum PA requirements, and are therefore under the health risk. The third class would include individuals that are not physically active at all, and are therefore under a great health risk. Even though the last two groups both contain individuals under health risk, the risks are at very different levels, and the individuals belonging to these two classes certainly need different care and motivation to increase their physical activities.

Once more data are collected and more diverse sample of population is reached, we will train a final NN and develop a web-based tool to administer the questionnaire and immediately identify individuals at risk of being not physically active enough. This program will be easily accessible to everyone in order to improve well being of general population.

The developed method predicted accurately 77% of time. Even though the results are not 100% perfect, they show a great potential for quick identification of individuals at health risk due to physical inactivity. Collecting a larger amount of data to use in the machine learning approach as well as collecting data from wider variety of collegiate population will improve already promising results of the proposed approach.

References

Booth, M. L. (2000). Assessment of physical activity: An international perspective. *Research Quarterly for Exercise and Sport, 71*(2), s114-120.

CDC. (2005). Trends in leisure-time physical inactivity by age, sex, and race/ethnicity: United States, 1994-2004. *Morbidity and Mortality Weekly Report, 54*(39), 991-994.

Dunn, M. S., & Wang, M. Q. (2003). Effects of physical activity on substance use among college students. *American Journal of Health Studies, 18*(2/3), 126-132.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA data mining software: an update. *SIGKDD Explorations*, 11(1).

Haskel, W., Lee, I. M., Pate, R. R., Powell, K. E., Blair, S. N., Franklin, B. A., Macera, C. A., Heath, G. W., Thompson, P. D., & Bauman, A. (2007). Physical activity and public health: Updated recommendations for adults from the American College of Sports Medicine and the American Heart Association. *Journal of American Heart Association, 116*, 1081-1093.

Magoc, D., Tomaka, J., & Thompson, S. (2010). Overweight, obesity and strong attitudes: Predicting participation in physical activity in a predominantly Hispanic college population. *Health Education Journal, 69*(4), 427-438.

Mounton, C. P., Calmbach, W. L., and Dhanda, R. (2000). Barriers and benefits to leisure-time physical activity among older Mexican Americans. *Archives of Family Medicine, 9*, 892-897.

Petosa, R. L., Suminski, R. R., & Hortz, B. (2003). Predicting vigorous physical activity using social congnitive theory. *American Journal of Health Behavior, 27*(4), 301-310.

Pratt, M., Macera, C. A., & Blanton, C. (1999). Levels of physical activity and inactivity in children and adults in the United States: Current evidence and research issues. *Medicine and Science in Sports and Exercise, 31*(Nov), S526-533.

Rovniak, L. S., Eileen, S. A., & Winett, R. A. (2002). Social cognitive determinants of physical activity in young adults: A prospective structural equation analysis. *Annals of Behavioral Medicine, 24*(2), 149-156.

Rusell, S. & Norvig, P. (2010). Artificial intelligence, a modern approach. *Prentice Hall.*

Sallis, J. F., Grossman, R. M., Pinski, R. B., Patterson, T. L., & Nader, P. R. (1987). The development of scales to measure social support for diet and exercise behaviors. *Preventive Medicine, 16*, 825-836.

Sallis, J. F., Pinski, R. B., Grossman, R. M., Patterson, T. L., & Nader, P. R. (1988). The development of self-efficacy scales for health-related diet and exercise behavior. *Health Education Research, 3*, 283-292.

Tan, P., Steinbach, M., & Kumar, V. (2006). Introduction to data mining. *Addison Wesley.*

Submitted 22 April, 2011
Accepted 15 June, 2011