A HEURISTIC COMPARISON OF PREDICTIVE MODELS IN ANALYZING ADULT OBESITY IN THE UNITED STATES

Ajaya K. Swain, Greehey School of Business, St. Mary's University, One Camino Santa Maria, San Antonio, TX 78228, 210-436-3619, aswain@stmarytx.edu
Monica J. Parzinger, Greehey School of Business, St. Mary's University, One Camino Santa Maria, San Antonio, TX 78228, 210-431-2026, mparzinger@stmarytx.edu
Orion J. Welch, Greehey School of Business, St. Mary's University, One Camino Santa Maria, San Antonio, TX 78228, 210-431-2031, owelch@stmarytx.edu

ABSTRACT

Considering the ever-rising cost of healthcare in the United States and the significant impact of obesity on the country's healthcare expenditure, this study attempts to identify key demographic and lifestyle behavior variables associated with adult obesity in the United States. Specifically, the study aims to create a profile of adult population who are at risk of being obese using predictive modeling techniques. Using the Centers for Disease Control and Prevention data, two predictive models, decision tree and logistic regression, are developed and compared to examine the important predictors of adult obesity.

Keywords: Predictive modeling, Obesity, Decision trees, Regression, and Healthcare

INTRODUCTION

Obesity has grown at such an alarming rate over the past decades that it has become a serious challenge in America today [15] [19] [21]. Fighting this epidemic has been part of the intense government [21] and not-for-profit efforts in the U.S. for years. In fact, many agencies and sources have been trying to assess the extent of obesity by collecting relevant information and creating an extensive related database. For instance, since 1984 the CDC-BRFSS (Centers for Disease Control and Prevention-Behavioral Risk Factor Surveillance System) has been collecting uniform data on preventive health behaviors, and risk behaviors that are associated with obesity.

Trend data from the BRFSS shows increased prevalence of obesity in the United States, regardless of sex, age, race, or educational level. For example, in 1991, only 4 of the 50 states reported obesity prevalence. These were in the 15%-19% range. At that time, no state reported higher than 19%. Seventeen years later, in 2008, the situation was aggravated; only one state had a prevalence of obesity less than 20%. Thirty-two of the fifty states had obesity prevalence rates equal to or greater than 25%, including six states with obesity rates higher than 30%. As recent as 2014, data indicates that more than one-third of U.S. adults are obese [22]. The percentages suggest that these peoples' health is compromised and their well-being is negatively affected. The healthcare issue extends into the financial realm. As the proportion of the U.S. population becomes more obese, related costs also increase. One study [11] estimated the potential increase in the country's annual obesity-related expenditures to be about $139 billion. Job absenteeism related to obesity costs $45 billion annually [5]. Hence, a closer look at the obesity situation by investigating the profile of individuals at risk of becoming obese is warranted.

Originating from the field of statistics and machine learning, data mining has advanced to be one of the top emerging technologies for exploring, detecting and discovering new knowledge from big data [27] [32]. The successful application of data mining techniques in analyzing healthcare data can provide novel

medical and healthcare knowledge to support clinical and administrative decision making [17]. The purpose of this research is to illustrate how data mining technologies can be used for analyzing rich, huge healthcare data. More specifically, this research explores, extracts, and analyzes the implicit, unknown and potentially useful information and meaningful patterns from the BRFSS survey data with a focus on adult obesity. The objective is to use predictive models (decision tree and regression) to examine the association of various demographic, socio-economic, and lifestyle behavioral factors with adult obesity. While comparing the performance and effectiveness of these models, the study generates profiles for at-risk obese adults.

LITERATURE REVIEW

Having emerged at the intersection of various methodologies, including statistics, databases, pattern recognition, artificial intelligence and parallel computing, data mining has been an extremely powerful approach to extract meaningful information from large multi-dimensional databases [27]. Data mining can be defined as: "the analysis of (often large) observational datasets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner" [16]. While statistics takes a mathematical approach and deals with numeric data only, data mining can analyze a variety of data including: numeric, categorical, images, text, and audio data [28]. The two approaches presented in this study are decision tree and logistic regression.

Decision trees (DT) are rule induction type models that can segment data by applying a set of rules. Search heuristics use recursive partitioning algorithms to split a large collection of observations into smaller homogeneous groups with respect to a particular target variable. The algorithms have to find the optimum number of splits and determine where to partition the data to maximize the information gain. The target variable is usually categorical and the decision tree model calculates the probability of a given record belonging to each of the target categories, or to classify the record by assigning it to the most likely category. Because of their fast construction process, visual representation, and easily interpretable rules, DTs are one of the most preferred among the available predictive models and have been widely applied in healthcare analytics [31]. Researchers in the field of healthcare [6] [8] [18] have successfully used decision tree models in diagnosis and prediction of several diseases to suggest appropriate prevention strategies.

Multivariate regression analysis is a valuable technique to quantify the relationship between two or more predictors and an outcome variable. Logistic regression belongs to a larger class of generalized linear models that predicts the expected value of the dichotomous target variable by a logit link function using the linear predictors. It is a widely used predictive modeling technique in which a binary outcome is linked to a set of potential predictor variables [30]. Step-wise selection methods are widely applied to identify a limited number of predictors for inclusion in regression models, especially in prediction problems or in situations with a large number of predictor variables. Although this method has been criticized for its selection approach [3], it is still a commonly used method.

The use of data mining techniques is not uncommon in the existing body of literature for obesity prediction. Most of the studies, however, have limited their predictions to childhood obesity. For example, Adnan & Husain compared three data mining methods that can be useful in predicting childhood obesity: Artificial Neural Networks, naïve Bayes, and the decision tree [1]. A study utilizing the medical history of 34 Hungarian children was conducted by Ferenci, et al using hierarchical cluster analysis, a technique often employed in data mining [9]. The results revealed the presence of two distinct clusters. Data mining

suggested a connection between systematic inflammation and obesity. Other studies utilizing data mining techniques have segmented the population by adolescence and sex. For example, Boone-Heinonen et al], in a cluster analysis, examined obesity-related behavior pattern and identified high risk adolescent groups using the National Longitudinal Study of Adolescent Health dataset consisting of 9251 observations [2]. Their study identified seven differing clusters in males and six differing clusters in females. Clusters represented characteristics such as School Clubs, Sports, and Dieting. Socio-demographic factors varied across clusters. Compared to School Clubs and Sports clusters, adjusted odds of prevalent and incident obesity were higher for most clusters in females but not males. A study focusing on midlife women in the U.S. used segmentation analyses in identifying five distinct subgroups based on their prevailing attitudes toward food and its preparation and consumption [26]. It was discovered that attitude segments are a significant predictor of obesity indicators. Several studies have reported that obesity prevalence varies by sex, age, and race [12] [22], and on socioeconomic and cigarette smoking status [14] [23]. A meta-analysis using BMI to determine overweight and obesity status performed with data from 97 major epidemiologic studies found that the entire population of obese individuals had a significantly higher mortality rate than did those with a normal BMI [13]. Finally, one study reported several economic factors (such as price of a meal in a restaurant, price of alcohol and cigarette) responsible for increased numbers of obese adults in the United States [7].

All these studies more or less have demonstrated the application and usefulness of data mining techniques on studying obesity but based on a limited population. Our study applies similar techniques in studying a wider range of population and using a much larger data set that encompasses survey respondents who are 18 years or older in households from all the 50 states in the US and its territories.

DATA AND METHODS

This paper utilizes the rich collection of health data from Centers for Disease Control and Prevention (CDC)-Behavioral Risk Factor Surveillance System (BRFSS). BRFSS is a state-based system of health surveys that collects information on health risk behaviors, preventive health practices, and health care access primarily related to chronic disease and injury. Data are annually collected through cross-sectional telephone surveys targeting adults 18 years or older in households from all 50 states, the District of Columbia, Puerto Rico, the U.S. Virgin Islands, and Guam. This is the world's largest continuously conducted telephone health surveillance system with more than 400,000 interviews conducted every year. (https://www.cdc.gov/brfss/ December 23, 2017). For this study, a year data from the BRFSS survey database was considered for analysis. The dataset has 292 variables and 449,526 observations.

As the focus of this study is to investigate the association between the demographic, socio-economic and lifestyle risk behavior, and obesity, several unrelated variables such as disease related variables were not considered for analysis. Furthermore, a few categorical variables were created in lieu of variables that were measured on Likert scale such as education. The final dataset contained 109 variables. The target variable was BMI (body mass index). Variables for this study are: General Health, Gender, Exercise in past 30 days, Education level, Employment status, Frequency of emotional support, Health care coverage, Hispanic, Height, and Average alcoholic drinks per day in past 30 days. BMI = [(Weight in lbs) ÷ (Height in inches)^2]*703. A level between 25.0 and 29.9 are considered overweight and a level of 30.0 and above are considered obese.

SAS Enterprise Miner (SAS EM) 14.2 was used as the data mining software. Logistic regression and decision trees were selected because of their ease in interpretation. Some of the characteristics of the sample data are noted. About 65% of observations have BMI levels of 30 and above suggesting that a notable portion of our sample is obese. Sixty percent of the sample data are female and about 34% are white non-Hispanics. Approximately 45% have income more than $50,000 and about 25% of the sample have income less than $25,000. Looking at the age, 8% are between 18 and 29 and 26% are 65 and above.

Prior to building a decision tree and logistic regression model, data partition was used as a strategy for honest assessment of the model performance. Unlike decision tree which has its own method for handling missing values, logistic regression in SAS EM uses a complete-case analysis approach. In this approach, the variables with missing values are not considered as inputs to the model. Rejecting all incomplete observations may ignore useful or important information which is still contained in the non-missing variables. To deal with this situation, an impute node was added before the regression node (Figure 1). By default, the impute node replaces the missing values with the mean of the non-missing values for interval variables, and mode of the non-missing values for categorical variables.

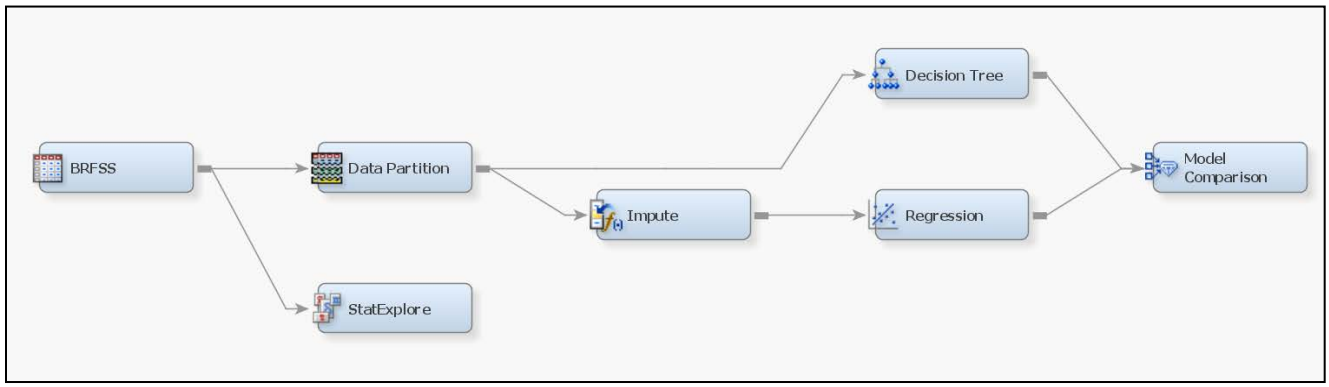

Figure 1. The process flow diagram

A logistic regression model (Figure 1) was built with stepwise selection method for variable selection. The stepwise stopping criteria was set at the default value of 0.05. Only the main effects were included. The default settings for optimization and convergence were used. The model with the smallest misclassification rate was chosen. The model comparison node produces a comparison of competing models using various benchmarking criteria.

A decision tree was constructed to identify homogeneous, high-risk subgroups and to assess the relative importance of potential risk factors in predicting overweight/obesity among the respondents. First, a large and complex tree is grown with data from all study variables by recursively partitioning the sample space into binary subsamples that lead to the formation nodes that can be split further and terminal nodes. The partitioning technique seeks variable splits that produce homogenous subgroups. The splitting criterion for the tree is the p-value of the Pearson Chi-square statistic for the target versus the branch node. The significance level that specifies the maximum acceptable p-value for a variable was 0.2. Maximum branch and depth were 2 and 6, respectively. The leaf size that specifies the minimum number of training observations allowed in a leaf node was 5. The split size that specifies the smallest number of training observations that a node must have before it is eligible to be split was set to 2. Up to 5 splitting rules were saved in each node. All other default settings in the tree node were retained. A pruning technique was applied to reduce the size of the original tree. The optimal tree was selected based on the lowest misclassification rate.

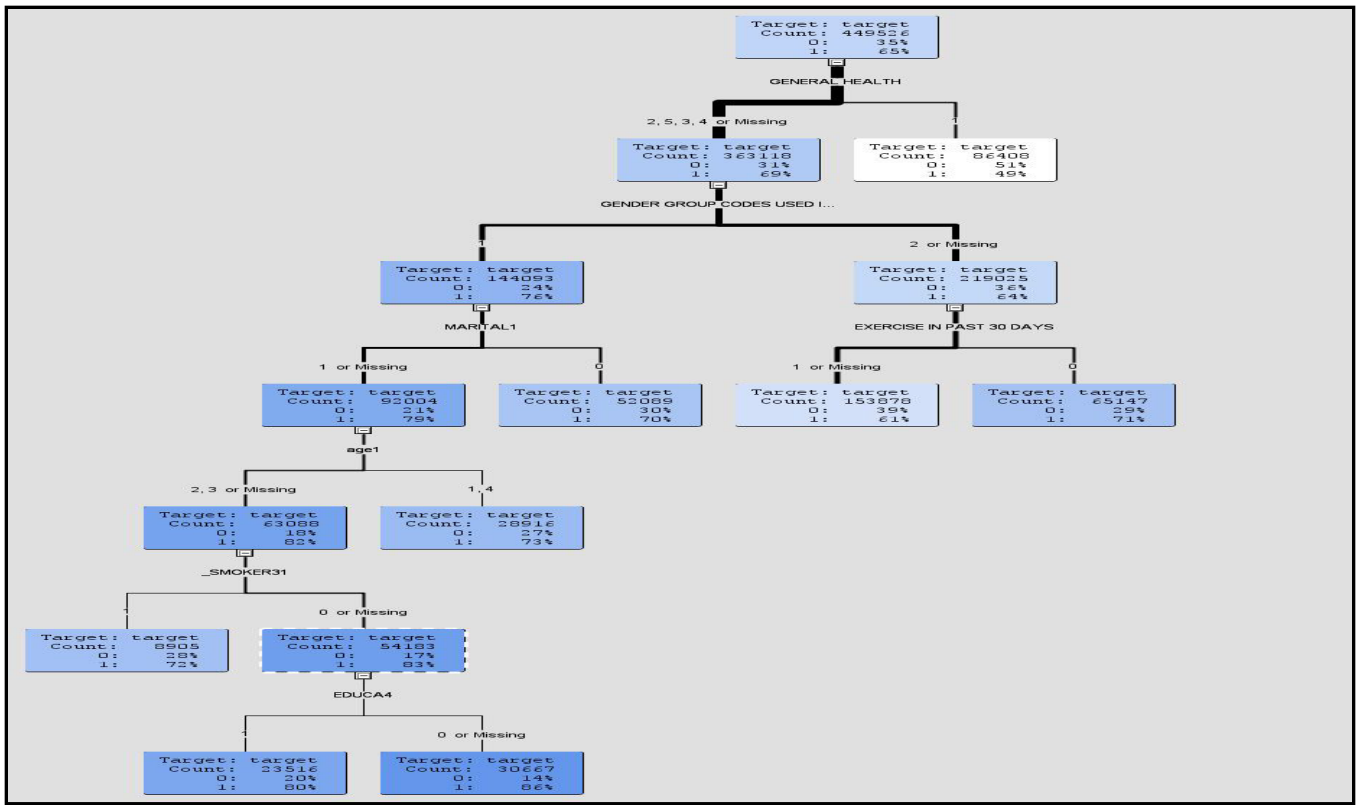

Figure 2. The decision tree

RESULTS, DISCUSSION, AND CONCLUSIONS

The decision tree as shown in Figure 2 has a validation sample root node with 65% of the population being obese (Target 1). The percentage of obese who reported being in good health is 49, whereas 69% of the people that are obese reported being in poor health or did not respond. Seventy one percent of males who have not done any exercise in the last 30 days are obese. The proportion of obese married females (who are in poor health) is 79% which is more than their unmarried counterparts. Again, married females that are in poor health aged between 18-29 and 65 and older are more likely to be obese than the other female age groups in this category. Non-smoking females in the age group 18-29 and 65 and older who do not have a college degree are more likely to be obese than their counter parts having a college degree. From the decision tree results, the profile of a person highly likely to be identified as obese is: a married non-smoking female, reportedly not in good health, aged either between 18-29 or over 65 who does not have a college degree.

The models were compared on multiple performance characteristics. The decision tree has higher validation misclassification rate (0.34) than regression (0.32). This means that the prediction accuracy of the regression model is better than the decision tree as it misclassifies fewer numbers of observations in to the wrong class. Accuracy measures are one of the most popular measures of performance of predictive models. However, to choose the best model, experts (e.g. Provost et al., 1998) suggest using the receiver operating characteristic (ROC) chart that graphically displays “sensitivity” (percentage of obese adults predicted correctly as obese) versus “1-specificity” (percentage of non-obese adults wrongly classified as obese). The different curves in the chart exhibit various degrees of concavity. The higher the degree of concavity, the better the model is expected to be. In the chart, the regression model appears to be the better

model. This finding is in agreement with the earlier comparison on misclassification rate.

To demonstrate an application of data mining in analyzing big data, two predictive models, decision tree and logistic regression, were developed. Accuracy of the prediction of these models were assessed and compared by misclassification rate as well as the ROC chart. This provided an opportunity to explore the models from different perspectives leading to identification of potential risk factors contributing to obesity. Past studies have shown that the proportion of overweight and obese adults in the United States has been consistently on the rise with almost one in every three adults currently either overweight or obese [4]. Our analyses confirmed these findings with 65% of the population in the sample found to be overweight or obese.

The logistics regression analysis indicated that obesity is strongly associated with multiple demographic and lifestyle risk behaviors such as: educational background, employment status, emotional support in the family, general health condition, race, access to healthcare coverage, and drinking habits of individuals. This is in line with several previous findings [10] [20] [24] highlighting the need for public health initiatives aimed at modifying and improving those behaviors. Gender differences were apparent for obese adults with physical inactivity. Among females, obesity was found to be associated with lower education and smoking habits. This finding reinforces an earlier report on obesity and smoking habits [25]. Age differences were obvious among obese adults females with females aged 29-65 years more likely to be obese than the females in other age groups.

Consistent with findings that social support from family impacts health parameters [29], this study finds that family-level factors such as greater emotional support is negatively associated with obesity. Individual demography such as race was found to be associated with obesity. Hispanics were found less likely to be obese than other racial groups in the data. Previous studies have reported inconclusive associations between exercise and smoking habits. This study could not detect any significant relationship between exercise and smoking habits among either genders within different age groups. Among the individual health behaviors, the amount of exercise per month was found to be positively associated with a reduced likelihood of obesity.

The prevalence of obesity in the U.S. has increased rapidly in the past several decades and constitutes a serious public health problem. This study uses a data mining approach and a nationally-representative sample to identify risk profiles for adult individuals in the United States. The findings confirm the current knowledge about the associations that demographic variables, marital status, gender, education, age, and behavioral factors such as smoking and exercise have with overweight/obesity besides identifying several new associations among these variables. Most often, analyzing big data in healthcare is a huge challenge. This research demonstrates how healthcare professionals and administrators in the United States can use predictive analytics to draw important insights from raw healthcare data that can help them establish appropriate management strategies.

REFERENCES

References are available upon request from the authors.

REFERENCES

[1] Adnan, M.H.M. and Husain, W. Hybrid approaches using decision tree, naïve bayes, means and euclidean distances for childhood obesity prediction. *International Journal of Software Engineering and Its Applications*, 2012, *6* (3), 99-106.

[2] Boone-Heinonen, J., Gordon-Larsen, P. and Adair, L.S. Obesogenic clusters: multidimensional adolescent obesity-related behaviors in the US. *Annals of Behavioral Medicine*, 2008, *36* (3), 217. doi:10.1007/s12160-008-9074-3

[3] Buckland, S.T., Burnham, K.P. and Augustin, N.H. Model selection: an integral part of inference. *Biometrics*, 1997, 603-618.

[4] Buxton, O.M. and Marcelli, E. Short and long sleep are positively associated with obesity, diabetes, hypertension, and cardiovascular disease among adults in the United States. *Social science & medicine*, 2010, *71* (5), 1027-1036. doi:10.1016/j.socscimed.2010.05.041

[5] Cawley, J. and Meyerhoefer, C. The medical care costs of obesity: an instrumental variables approach. *Journal of health economics*, 2012, *31* (1), 219-230. doi:10.1016/j.jhealeco.2011.10.003

[6] Chang, C.L. and Chen, C.H. Applying decision tree and neural network to increase quality of dermatologic diagnosis. *Expert Systems with Applications*, 2009, *36* (2), 4035-4041.

[7] Chou, S.Y., Grossman, M. and Saffer, H. An economic analysis of adult obesity: results from the Behavioral Risk Factor Surveillance System. *Journal of health economics*, 2004, *23* (3), 565-587.

[8] Eom, J.H., Kim, S.C. and Zhang, B.T. AptaCDSS-E: A classifier ensemble-based clinical decision support system for cardiovascular disease level prediction. *Expert Systems with Applications*, 2008, *34* (4), 2465-2479.

[9] Ferenci, T., Almássy, Z., Merkei, Z.O., Kovács, A. and Kovács, L. Cluster analysis of obesity-related parameters of Hungarian children. In *Proc. of BUDAMED'08 Conference, Budapest, Hungary, November 2008*, 33-37.

[10] Fine, L.J., Philogene, G.S., Gramling, R., Coups, E.J. and Sinha, S. Prevalence of multiple chronic disease risk factors: 2001 National Health Interview Survey. *American journal of preventive medicine*, 2004, *27* (2), 18-24.

[11] Finkelstein, E.A., Ruhm, C.J. and Kosa, K.M. Economic causes and consequences of obesity. *Annu. Rev. Public Health*, 2005, *26*, 239-257. doi:10.1146/annurev.publhealth.26.021304.144628

[12] Flegal, K.M., Carroll, M.D., Kit, B.K. and Ogden, C.L. Prevalence of obesity and trends in the distribution of body mass index among US adults, 1999-2010. *Jama*, 2012, *307* (5), 491-497. doi:10.1001/jama.2012.39

[13] Flegal, K.M., Kit, B.K., Orpana, H. and Graubard, B.I. Association of all-cause mortality with overweight and obesity using standard body mass index categories: a systematic review and meta-analysis. *Jama*, 2013, *309* (1), 71-82. doi:10.1001/jama.2012.113905

[14] Flegal, K.M., Kruszon-Moran, D., Carroll, M.D., Fryar, C.D. and Ogden, C.L. Trends in obesity among adults in the United States, 2005 to 2014. *Jama*, 2016, *315* (21), 2284-2291.

[15] Fryar, C.D., Carroll, M.D. and Ogden, C.L. Prevalence of overweight, obesity, and extreme obesity among adults: United States, trends 1960–1962 through 2009–2010. *Hyattsville, 2012, MD: National Center for Health Statistics*.

[16] Hand, D.J., Mannila, H. and Smyth, P. *Principles of data mining*. 2001, MIT press.

[17] Harper, P.R. A review and comparison of classification algorithms for medical decision making. *Health Policy*, 2005, *71* (3), 315-331. doi:10.1016/j.healthpol.2004.05.002

[18] Kukar, M., Kononenko, I., Grošelj, C., Kralj, K. and Fettich, J. Analysing and improving the diagnosis of ischaemic heart disease with machine learning. *Artificial intelligence in medicine*, 1999, *16* (1), 25-50.

[19] Kung, H.C., Hoyert, D.L., Xu, J. and Murphy, S.L. Deaths: final data for 2005. *Natl Vital Stat*

Rep, 2008, *56* (10), 1-120.
[20] Laaksonen, M., PRÁTTALÄ, R. and Karisto, A. Patterns of unhealthy behaviour in Finland. *The European Journal of Public Health*, 2001, *11* (3), 294-300. doi:10.1093/eurpub/11.3.294
[21] Levi, J., Segal, L.M., St Laurent, R., Lang, A. and Rayburn, J. F as in fat: how obesity threatens America's future 2012, 2012.
[22] Ogden, C.L., Carroll, M.D., Kit, B.K. and Flegal, K.M. Prevalence of childhood and adult obesity in the United States, 2011-2012. *Jama*, 2014, *311* (8), 806-814.
[23] Ogden, C.L., Yanovski, S.Z., Carroll, M.D. and Flegal, K.M. The epidemiology of obesity. *Gastroenterology*, 2007, *132* (6), 2087-2102.
[24] Poortinga, W. The prevalence and clustering of four major lifestyle risk factors in an English adult population. *Preventive medicine*, 2007, *44* (2), 124-128. doi:10.1016/j.ypmed.2006.10.006
[25] Strine, T.W. and Chapman, D.P. Associations of frequent sleep insufficiency with health-related quality of life and health behaviors. *Sleep medicine*, 2005, *6* (1), 23-27. doi:10.1016/j.sleep.2004.06.003
[26] Sudo, N., Degeneffe, D., Vue, H., Merkle, E., Kinsey, J., Ghosh, K. and Reicks, M. Relationship between attitudes and indicators of obesity for midlife women. *Health Education & Behavior*, 2009, *36* (6), 1082-1094. doi:10.1177/1090198109335653
[27] Swain, A.K., 2016a. Big data analytics: An expert interview with Bipin Chadha, data scientist for United Services Automobile Association (USAA). *Journal of Information Technology Case and Application Research*, 2016, *18* (3), 181-185. DOI: 10.1080/15228053.2016.1223497
[28] Swain, A.K., 2016b. Mining big data to support decision making in healthcare. *Journal of Information Technology Case and Application Research*, 2016, *18* (3), 141-154. DOI: 10.1080/15228053.2016.1245522
[29] Uchino, B.N., Cacioppo, J.T. and Kiecolt-Glaser, J.K. The relationship between social support and physiological processes: a review with emphasis on underlying mechanisms and implications for health. *Psychological bulletin*, 1996, *119* (3), 488. doi:10.1037/0033-2909.119.3.488
[30] Yap, B.W., Ong, S.H. and Husain, N.H.M. Using data mining to improve assessment of credit worthiness via credit scoring models. *Expert Systems with Applications*, 2011, *38* (10), 13274-13283. doi:10.1016/j.eswa.2011.04.147
[31] Yeh, J.Y., Wu, T.H. and Tsao, C.W. Using data mining techniques to predict hospitalization of hemodialysis patients. *Decision Support Systems*, 2011, *50* (2), 439-448.
[32] Yoo, I., Alafaireet, P., Marinov, M., Pena-Hernandez, K., Gopidi, R., Chang, J.F. and Hua, L. Data mining in healthcare and biomedicine: a survey of the literature. *Journal of medical systems*, 2012, *36* (4), 2431-2448. doi:10.1007/s10916-011-9710-5