

Predictive Modeling for Wellness and Chronic Conditions

Dr. Ravi S. Behra, Ph.D.

Department of IT & Operations Management
College Of Business
Florida Atlantic University
Boca Raton, FL, USA
rbehara@fau.edu

Dr. Ankur Agarwal, Ph.D.

Department of Computer & Electrical Engineering and
Computer Science
Florida Atlantic University
Boca Raton, FL, USA
aagarwa2@fau.edu

Pranitha Pulumati

Department of Computer & Electrical Engineering and
Computer Science
Florida Atlantic University
Boca Raton, FL, USA
ppulumati2013@fau.edu

Ritesh Jain

Department of Computer & Electrical Engineering and
Computer Science
Florida Atlantic University
Boca Raton, FL, USA
rjain2013@fau.edu

Dr. Vinaya Rao. M.D.

Methodist University Hospital Transplant Institute,
Memphis, TN, USA
vr3@uthsc.edu

Abstract — There is a significant increase in attention being paid to personal wellness as a preventative strategy in healthcare. At the same time, chronic diseases are the major cause of mortality, accounting for 7 out of 10 deaths in the United States. Healthcare costs involved in managing chronic diseases are also very high. So there is a need to help better maintain individual wellness, as well as better manage chronic conditions. Predictive analytics based clinical decision support systems need to be developed to help individuals and healthcare providers to better manage wellness or chronic conditions. In this paper, we investigate two different classifiers to predict the wellness outcome and the occurrence of a chronic condition (diabetes). The models were evaluated on the basis of overall accuracy, root mean squared error and Area under ROC. National CDC-NHANES data that is based on the health and nutritional status of individuals in the United States is used to develop the models.

Keywords— *Wellness, Diabetes, Bayes Network, Multilayer Perceptron*

I. INTRODUCTION

Broadly defined, wellness is the proactive process of being aware and making choices to attain a healthy and fulfilling life. It is dependent on many factors including physical, emotional, social, environmental and occupational wellness. Physical wellness specifically is related to efforts involved in maintaining a healthy body through efforts including exercise, diet, as well as preventative care. It is this aspect of clinically-oriented wellness that is the focus of the modeling effort in this study. In addition, we also

investigate the state of being unwell in the chronic sense of the term by focusing on Diabetes, a chronic condition that is caused by an imbalance in the secretion of insulin resulting in a disturbance in the sugar levels of the blood. The high blood glucose (blood sugar) levels are either because of inadequate insulin production, or because the body's cells do not respond properly to insulin. According to the American Diabetes Association, 29.1 million people in the United States (9.3% of the population), had diabetes in 2012. Among this population, there were 8.1 million people who were undiagnosed. The incidence of diabetes keeps increasing, with 1.7 million new diagnoses made in 2012. In addition, 86 million individuals over 20 years old had pre-diabetes, and could be considered to be in a state between wellness and chronic illness. Diabetes remains the 7th leading cause of death in the United States in 2010, with 69,071 deaths linked directly to it, and many more (a total of 234,05) having it as the underlying or contributing cause. So the effective management of wellness, pre-diabetes, and diabetes are important efforts.

An approach to effectively managing wellness and chronic conditions (diabetes in this study) requires us to better understand the drivers, and develop methods that can help guide individuals to develop strategies to maintain or improve their physical health. To achieve this, we develop different models in this study which may be used to develop individual/clinical decision support solutions. We focus on two machine learning approaches of Multilayer Perceptron and Bayes Network. We apply them to predict the wellness

outcome and the occurrence of diabetes in an individual when a certain set of parameters are provided. The models are developed using WEKA.

II. LITERATURE REVIEW

There have been many recent efforts at applying appropriate technologies to improve healthcare and its delivery [1, 2] using predictive modeling in healthcare based on artificial neural networks [3]. A Multilayer Perceptron (MLP) network was used for the diagnosis of low back pain and sciatica that resulted in the MLP network model producing the correct classification more often than any of the three groups of doctors or the fuzzy logic system [4]. A neural network based Structural Health Monitoring System has been proposed [5]. This system uses wireless sensor network where thousands of sensor nodes perform distributed sensing and collaborative computing for structural health analysis. It uses several algorithms to predict a particular disease.

A Collaborative Assessment and Recommendation Engine (CARE) has been proposed to predict future disease risk based on patient’s medical history and other similar patients history [6]. It uses collaborative methods to select the most significant risk factors and generates prediction based on it. Genetic Algorithm based models to predict asthma risks in children have also been developed [7], while a prognostic engine to predict physiological health status of a patient has been modeled with the help of Artificial Neural Network based on historical clinical data ranging from Respiratory System to Circulatory System [8]. While most models are based on structured data, unstructured data from physician notes have also been used towards the development of prognostic models [9, 10]. Researchers have begun exploring the use of Bayesian Networks [11] in developing predictive models.

Beyond modeling methods, critical factors in determining health and well-being are considered in this research. Age, gender and ethnicity are important factors in assessing health. BMI and Blood Pressure are the important factors in maintaining the health [12, 13]. National Cholesterol Education Program tells us that cholesterol and triglycerides levels are important in considering health [13]. Physical activity is also important in reducing the risk of diseases [14]. There is also evidence that indicates that there is no risk-free level of exposure to tobacco smoke [15, 16]. It shows that very heavy alcohol consumption is an independent risk for Chronic Pancreatitis [16]. It is shown that sleep is a critical factor of health and well-being [17].

Based on these streams of literature, we identify appropriate clinically relevant measures to develop predictive models for wellness and diabetes based on suitable machine learning methods.

III. METHODOLOGY AND MODELING

In order to predict the outcome for wellness and diabetes, we have implemented two classifiers: Multilayer Perceptron

(MLP) and Bayesian Networks (BN). Bayesian Networks are the directed acyclic graphs which facilitate the representation of joint probability distributions over a set of random variables [11]. Multilayer Perceptron is a feed forward neural network with one or more hidden layers between input and output. Feed Forward means data flows from one direction to another i.e. from input nodes towards the output node. This network is trained with a back propagation learning algorithm. MLP helps in distinguish between the data that is not linearly separable. Except input nodes all nodes consist of a non-linear activation function [18]. Input layer consist of a set of input parameters based on which prediction has to be made, hidden layer consist of a set of hidden nodes which helps in solving the non-linear data problem, these nodes helps in converting input data into the form which can be used by the output node and lastly output layer consist of a output node with a non-linear activation function to make the prediction. The tool which we have used to classify and analyze the results is WEKA.

Input feature values included a variety of variables based on individual demographics, physical examination, clinical laboratory tests self- assessment health-status questionnaires. The networks were trained on a set of variables for each individual, where the outcomes of the wellness and diabetes are known. When a set of input feature values taken from Demographics, Examination, Laboratory tests and Questionnaires are presented to the trained network, an output is generated that represents the classification (Excellent, Very good, Good, Fair and Poor for wellness and Yes, No and Borderline for Diabetes) of the prediction. The outcome is based on the trained knowledge which the network has learned through the training.

IV. DATA ANALYSIS AND RESULTS

The data used in this study is extracted from the National Health Nutrition Examination Survey (NHANES) which is a program of studies designed to assess the health and nutritional status of adults and children in the United States. NHANES provides health statistics to the Nation. NHANES is an important program of National Center for Health Statistics (NCHS) which is a part of the Centers for Disease Control (CDC). NHANES conducts a survey every year of around 5000 individuals and the data includes demographics, socioeconomic, dietary and health-related questions [19]. Table I shows some of the NHANES variables used in the study.

Table I: Sample Variables from NHANES 2011-12 Data

NHANES Variables	
Demographics	
Variable Name	Variable Definition
RIAGENDR	Gender
RIDAGEYR	Age in years at screening
RIDRETH3	Race/Hispanic origin w/ NH Asian

Examination	
BPXSY1	Systolic: Blood Pres (1st rdg) mm Hg
BPXDI1	Diastolic: Blood Pres (1st rdg) mm Hg
BPXSY2	Systolic: Blood Pres (2nd rdg) mm Hg
BPXDI2	Diastolic: Blood Pres (2nd rdg) mm Hg
BPXSY3	Systolic: Blood Pres (3rd rdg) mm Hg
BPXDI3	Diastolic: Blood Pres (3rd rdg) mm Hg
BPXSY4	Systolic: Blood Pres (4th rdg) mm Hg
BPXDI4	Diastolic: Blood Pres (4th rdg) mm Hg
BMXBMI	Body Mass Index (kg/m**2)
Laboratory	
LBXTC	Total Cholesterol (mg/dL)
LBXTR	Triglyceride (mg/dL)
Questionnaire	
ALQ101	Had at least 12 alcohol drinks.1 yr?
PAQ605	Vigorous work activity
PAQ620	Moderate work activity
SLD010H	How much sleep do you get (hours)?
SMQ020	Smoked at least 100 cigarettes in life

A. Data and Results – Wellness

For the development of the prediction model for wellness, a study dataset of 824 individuals was obtained from the NHANES 2011-2012 dataset. This included an equal number of individuals in each of the classes (those who self-assessed their health as excellent, very good, good, fair, or poor). The study dataset consisted of 341 females and 483 males, between 19 to 80 years old. Two-thirds or 544 records were used to train the model and 280 records were used to test the model. A total of 29 attributes (NHANES variables) were used as inputs to generate one of the five predict wellness classes. These attributes include age, gender blood pressure, ethnicity, and BMI among others as shown in Table I. A total of 155 instances out of 280 instances are correctly classified with MLP with 10 hidden layers and a total of 154 instances out of 280 instances are correctly classified with BN.

Table II shows the results of both the classifiers and it is found that even though the accuracy of MLP is marginally better than BN, it can be seen that the weighted average of ROC curve plotted for False Positive Rate on x-axis and True Positive Rate on y-axis is better for BN than the MLP. Also, it can be seen that, Root Mean Squared Error is better for BN than MLP.

Tables III and IV shows the confusion matrix for the two models considered. In the confusion matrix, each column represents the instances in a predicted class and each row represents the instances in an actual class.

Table II: Accuracy, RMSE and ROC Area from BN and MLP for Wellness Analysis

Classifier	Accuracy	RMSE	ROC Area
MLP	55.3571%	0.3925	0.809
BN	55%	0.3456	0.845

Table III: Confusion Matrix for Wellness MLP Model

Actual Class	Predicted Class				
	Excellent	Very good	Good	Fair	Poor
Excellent	35	13	2	6	1
Very good	5	37	9	2	1
Good	4	12	24	12	1
Fair	9	9	13	26	9
Poor	7	0	0	10	33

Table IV: Confusion Matrix for Wellness BN Model

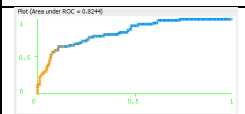
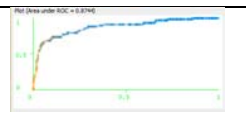
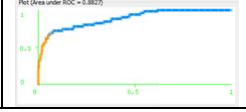
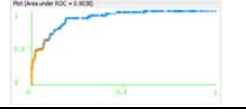
Actual Class	Predicted Class				
	Excellent	Very good	Good	Fair	Poor
Excellent	41	8	2	3	3
Very good	10	36	3	4	1
Good	1	12	23	15	2
Fair	7	11	10	23	15
Poor	5	0	0	14	31

Table V shows the Area under ROC for each of the five classes for both the MLP and BN models. To plot an ROC curve for multiclass classifier, one class for which the ROC curve needs to be plotted is considered as one class and all the remaining classes are considered as another class. In this way ROC curve for multiclass classification is plotted in a way similar to that of ROC curve for binary class classification to illustrate the performance of the model in terms of true positive rate against false positive rate. Table VI shows samples of the ROCs.

Table V: Area under ROC for Different Classes of Wellness

Actual Class	Area under ROC	
	MLP	BN
Excellent	0.8244	0.8744
Very Good	0.8523	0.8897
Good	0.8068	0.8710
Fair	0.7067	0.7192
Poor	0.8827	0.9038

Table VI: Sample Area under ROC Graphs for Wellness

Health Condition Assessment	Area under ROC	
	MLP	BN
Excellent		
Poor		

B. Data and Results – Diabetes

For the development of prediction model for diabetes, a study dataset of 1825 individuals was obtained from NHANES 2011-2012 dataset including 119 borderline diabetic patients, 706 diabetic patients and 1000 non-diabetic individuals. It consisted of 1051 males and 774 females between 18 & 80 years of age. Two-thirds or 1205 records were used to train the model and 620 records were used to test the model. A total of 43 attributes (NHANES variables) were used to generate the patterns to predict diabetes such as blood pressure, alcohol consumption, gender, and fasting glucose. A total of 556 instances out of 620 instances were correctly classified with MLP with 23 hidden layers and a total of 449 instances out of 620 instances were correctly classified with BN.

Performance of Multilayer Perceptron (MLP) model with 23 hidden layers was found to be better than that of Bayesian Network (BN). Results of both the models are shown in Table VII. It gives the measures in accuracy, root mean squared error, and ROC area.

Table VII: Accuracy, RMSE and ROC Area from BN and MLP for occurrence of Diabetes

Classifier	Accuracy	RMSE	ROC Area
MLP	89.68%	0.2392	0.969
BN	72.42%	0.3668	0.826

Table VIII and IX shows the confusion matrix for the two models considered. In confusion matrix, each column represents the instances in a predicted class and each row represents the instances in an actual class.

Table VIII: Confusion Matrix for Diabetes MLP Model

Actual Class	Predicted Class		
	Yes	No	Borderline
Yes	217	0	36
No	0	327	0
Borderline	28	0	12

Table IX: Confusion Matrix for Diabetes BN Model

Actual Class	Predicted Class		
	Yes	No	Borderline
Yes	188	64	1
No	261	66	0
Borderline	24	16	0

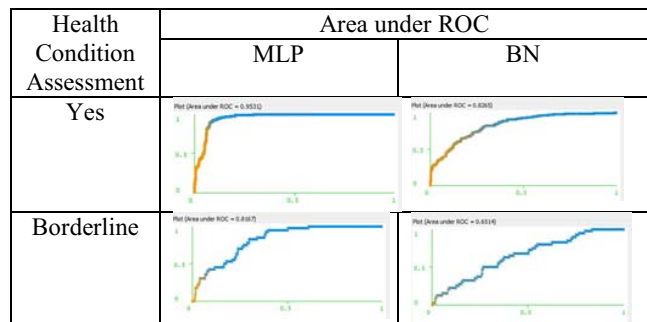
Table X shows the Area under ROCs for both the MLP and BN models. As before, to plot an ROC curve for multiclass classifier, one class for which the ROC curve needs to be plotted is considered as one class and all the remaining classes are considered as another class. That is, when considering the ROC for borderline, the remaining classes of yes/no are combined to form the other class. In this way ROC curve for multiclass classification is plotted

in a way similar to that of ROC curve for binary class classification to illustrate the performance of the model in terms of true positive rate against false positive rate. ROC area of MLP was found to be greater than that of BN for each of the classes. Table XI shows sample ROCs.

Table X: Area under ROC for Different Classes of Diabetes

Health Condition Assessment	Area under ROC	
	MLP	BN
Yes	0.9531	0.8265
No	1.0	0.8477
Borderline	0.8167	0.6514

Table XI: Confusion Matrix for Diabetes



V. CONCLUSION AND FUTURE ENHANCEMENTS

This paper describes two models, MLP and BN, to predict the wellness and the occurrence of diabetes in an individual. It is been found that in both the predictions with the accuracy of 55.36% in wellness prediction and with the accuracy of 89.68% prediction of occurrence of diabetes in an individual, the MLP model performed marginally better in wellness, and significantly better in the diabetes model, than the BN model. But at the class level, the BN resulted in better Area under ROC in the wellness model, while the MLP performed better in the diabetes model. Appropriate predictive models like those developed in this study may be used to develop individual/clinical decision support solutions, to help maintain wellness or improve the management of chronic diseases (such as diabetes). We expect to continue this research to improve the accuracy of the models by varying the dataset and the variables, exploring other models, as well as develop models for other chronic conditions such as asthma and coronary heart disease.

VI. REFERENCES

- [1] B. Furht and A. Agarwal (Editors), Handbook of Medical and Healthcare Technologies, Springer Publication, October 2013, ISBN: 978-1-4614-8494-3.
- [2] T. J. T. Cyr, A. Agarwal and B. Furht, Brief Overview of Various Healthcare Tools, Methods, Framework and

- Standards, "Handbook of Medical and Healthcare Technologies", Springer, 2014.
- [3] C. E. Floyd, J. Y. Lo, A. J. Yun, D. C. Sullivan and P. J. Kornguth, "Prediction of breast cancer malignancy using an artificial neural network," *Cancer*, vol. 74, no. 11, pp. 2944-2948, 1994.
- [4] D. Bounds, P. Lloyd, B. Mathew and G. Waddell, "A multilayer perceptron network for the diagnosis of low back pain," in *Neural Networks*, San Diego, 1988.
- [5] X. Xie, J. Guo, H. Zhang, T. Jiang, R. Bie and Y. Sun, "Neural-network based structural health monitoring with wireless sensor networks," in *Natural Computation (ICNC)*, 2013 Ninth International Conference, Shenyang, 2013.
- [6] D. Darcy A, C. Nitesh V, C. Nicholas A and A.-L. Barabási, "Time to CARE: a collaborative engine for practical disease prediction," *Data Mining and Knowledge Discovery*, vol. 20, no. 3, pp. 388-415, 2010.
- [7] E. Chatzimichail, E. Paraskakis and A. Rigas, "An Evolutionary Two-Objective Genetic Algorithm for Asthma Prediction," in *Computer Modeling and Simulation (UKSim)*, Cambridge, 2013.
- [8] P. Ghavami and K. Kapur, "Artificial neural network-enabled prognostics for patient health management," in *Prognostics and Health Management (PHM)*, Denver, 2012.
- [9] R. S. Behara, A. Agarwal, F. Fatteh and B. Furht, "Predicting Hospital Readmission Risk for COPD Using EHR Information", in *Handbook of Medical and Healthcare Technologies*, Springer, 2014.
- [10] R..S. Behara, F. Fatteh, P. Jain, P. Laxmi, A. Agarwal, "A Prognostic COPD Healthcare Management System", *IEEE International Conference on Systems Engineering*, 2013, Orlando, Florida.
- [11] F. Nir, G. Dan and G. Moises, "Bayesian Network Classifiers," *Machine Learning*, vol. 29, no. 2-3, pp. 131-163, 1997.
- [12] "Centers for Disease Control and Prevention," 11 July 2014. [Online]. Available: http://www.cdc.gov/healthyweight/assessing/bmi/adult_bmi/index.html#Consequences. [Accessed 2014].
- [13] J. Cleeman, "ATP III Guidelines At-A-Glance Quick Desk Reference," [Online]. Available: <http://www.nhlbi.nih.gov/files/docs/guidelines/atglance.pdf>.
- [14] L. Kushi, C. Doyle, M. McCullough, C. Rock, W. Demark-Wahnefried, E. Bandera, S. Gapstur, A. Patel, K. Andrews and T. Gansler, "American Cancer Society Guidelines on nutrition and physical activity for cancer prevention: reducing the risk of cancer with healthy food choices and physical activity," *CA Cancer J Clin*, vol. 62, no. 1, pp. 30-67, 2012.
- [15] "How Tobacco Smoke Causes Disease: The Biology and Behavioral Basis for Smoking-Attributable Disease: A Report of the Surgeon General," *Centers for Disease Control and Prevention*, 2010. [Online]. Available: <http://www.ncbi.nlm.nih.gov/books/NBK53017/>.
- [16] D. Yadav, R. Hawes, R. Brand and e. al, "Alcohol Consumption, Cigarette Smoking, and the Risk of Recurrent Acute and Chronic Pancreatitis," *Arch Intern Med*, vol. 169, pp. 1035-1045, 2009.
- [17] "Sleep Disorders and Sleep Deprivation: An Unmet Public Health Problem," *National Academies Press (US)*, 2006. [Online]. Available: <http://www.ncbi.nlm.nih.gov/books/NBK19944/>.
- [18] S. Pal and S. Mitra, "Multilayer perceptron, fuzzy sets, and classification," in *Neural Networks*, 1992.
- [19] "Centers for Disease Control and Prevention," [Online]. Available: http://www.cdc.gov/nchs/nhanes/about_nhanes.htm. [Accessed 03 February 2014].