Global Journal of Health Science; Vol. 7, No. 5; 2015
ISSN 1916-9736 E-ISSN 1916-9744
Published by Canadian Center of Science and Education

Type 2 Diabetes Mellitus Screening and Risk Factors Using Decision Tree: Results of Data Mining

Shafi Habibi[1], Maryam Ahmadi[1,2] & Somayeh Alizadeh[3]

[1] Department of Health Information Management, School of Health Management and Information Sciences, Iran University of Medical Sciences, Tehran, Iran

[2] Health Management and Economics Research Center, School of Health Management and Information Sciences, Iran University of Medical Sciences, Tehran, Iran

[3] School of Industrial Engineering, Khajeh Nasir Toosi University of Technology, Tehran, Iran

Correspondence: Maryam Ahmadi, Professor, School of Health Management and Information Sciences, Iran University of Medical Sciences, No. 6, Rashid Yasemi st., Tehran, Iran. Tel: 98-218-866-5052; Fax: 98-218-879-3805. E-mail: ahmadi.m@iums.ac.ir

Received: December 23, 2014 Accepted: January 19, 2015 Online Published: March 16, 2015

doi:10.5539/gjhs.v7n5p304 URL: http://dx.doi.org/10.5539/gjhs.v7n5p304

Abstract

Objectives: The aim of this study was to examine a predictive model using features related to the diabetes type 2 risk factors.

Methods: The data were obtained from a database in a diabetes control system in Tabriz, Iran. The data included all people referred for diabetes screening between 2009 and 2011. The features considered as "Inputs" were: age, sex, systolic and diastolic blood pressure, family history of diabetes, and body mass index (BMI). Moreover, we used diagnosis as "Class". We applied the "Decision Tree" technique and "J48" algorithm in the WEKA (3.6.10 version) software to develop the model.

Results: After data preprocessing and preparation, we used 22,398 records for data mining. The model precision to identify patients was 0.717. The age factor was placed in the root node of the tree as a result of higher information gain. The ROC curve indicates the model function in identification of patients and those individuals who are healthy. The curve indicates high capability of the model, especially in identification of the healthy persons.

Conclusions: We developed a model using the decision tree for screening T2DM which did not require laboratory tests for T2DM diagnosis.

Keywords: data mining, decision trees, Diabetes Mellitus Type 2, early diagnosis, risk factors

1. Introduction

Type 2 diabetes is a chronic disease and one of the most common endocrine diseases including 90 to 95 percent of diabetic patients (American Diabetes Association, 2013) with different degrees of prevalence in various societies (King, Aubert, & Herman, 1998). It was recognized by an asymptomatic phase between the real onset of diabetic hyperglycemia and clinical diagnosis which lasts at least for 4-7 years (Brown, Critchley, Bogowicz, Mayige, & Unwin, 2012). Late or lack of diabetes diagnosis causes the increase of various chronic vascular complications (Heydari, Radi, Razmjou, & Amiri, 2010). Moreover, early diagnosis and prevention of diabetes reduces the high expenses associated with disease control and complication treatments and prevents hospital admissions due to its severe complications (Karter et al., 2003).

It is well known that about 30 to 80 percent of type 2 diabetic cases remain undiagnosed (Brown et al., 2012). Therefore, considering the prevention principle and in order to fight against the current widespread prevalence of diabetes, there is great emphasis on the significance of screening and recognition of those who might have diabetes or its higher probability without any symptoms. Timely diagnosis and prevention result in the decrease in mortality and prevention and decrease in the diabetes complications and improvement of quality of life (Gregg et al., 2001). The main challenge in diabetes screening, however, is the need to study and take blood samples of several people, which is expensive in terms of financial and personnel resources, and is beyond the capabilities

of the health system, especially in the developing countries. Using data mining and knowledge discovery capabilities which identify latent patterns associated with diagnostic decisions from within the data could help in the prediction and diagnosis of the disease. In order to perform data mining and knowledge discovery, one needs large amounts of data related to the subject saved in the databases. Today, the importance of data saving and providing electronic records for the patients and those who refer to health centers is considered as a tool to improve the level of health of the people in society, and discussions about establishing and developing databases of Electronic Health Records (EHR) would pave the way for new studies in the health issues in society, even in the developing countries (DeVoe et al., 2011).

Data mining is a practical branch of artificial intelligence which discovers latent patterns through looking for relationships between features in large databases. The pattern discovered should be significant and provide advantages including economic ones (Witten & Frank, 2005). Classification is one of data mining methods used in the various studies in the health sector in order to build prediction models (Bellazzi, Ferrazzi, & Sacchi, 2011). The capability of these classification methods has been confirmed in discovering the relationships and patterns among the features in the databases and using the results and patterns thus discovered for diagnosis and prediction (Upadhyaya, Farahmand, & Baker-Demaray, 2013). In the screening of type 2 diabetes mellitus (T2DM), however, the capabilities of the classification techniques have not yet been demonstrated. The decision tree is one powerful and highly used method in classification which has found use in various medical arenas, in studies associated with the prediction and diagnosis and its application for prediction increased significantly (Liao, Chu, & Hsiao, 2012). As screening is performed at a wide level of society and requires great expense, facilities and time, further research in this field could help and improve health level of the society. Therefore, the aim of this study was to examine a predictive model using the features related to diabetes type 2 risk factors in order to help in the screening of diabetes using the decision tree.

2. Method

The data were obtained from the database of a web-based health center diabetes control system in Tabriz (center of East Azerbaijan province in Iran) designed for recording the data of those who had referred for diabetes screening from 2009 to 2011. According to the screening model communicated by the Iran Health Ministry to the health centers in the provinces and cities, those who had at least one risk factor of obesity or overweight, history of diabetes in first-degree relatives, hypertension, pregnancy, history of gestational diabetes, history of abortion, stillbirth, and birth of a child more than 4 kg, and past history of diabetes entered the system, and the tests required to diagnose diabetes were performed after recording the anthropometric data, blood pressure, and family history. Among 60,010 extracted records, 32,044 ones lacked value in the diagnosis field or each of the six fields used for entering the decision tree, which were all omitted. Furthermore, pregnant women (418 records) and those with a past history of diabetes (5,150 records) were excluded, as type 2 diabetes alone was considered in this study. After excluding the records mentioned above with the missed value, a total number of 22,398 records were used as diabetic, pre-diabetic and healthy categories, among which 924 persons (4.1%) had diabetes, 3,062 individuals (13.7%) were pre-diabetic, and 18,412 ones (82.2%) were healthy. The number of women and men were 15,019 and 7,379, respectively.

Features including age, sex, systolic and diastolic blood pressure, family history of diabetes, and body mass index (BMI) were used as inputs for the decision tree and the diagnosis feature were used as class. Unnecessary features such as glucose and triglyceride levels, etc. were omitted. The age variable was calculated and added using the date of birth and date of reference to the health center, and the same was performed for the BMI variables, using height and weight.

The technique of decision tree and J48 algorithm, which is the most important algorithm used for developing the decision tree in WEKA (3.6.10 version), was applied to develop the prediction model. J48 is Weka's implementation of Quinlan's C4.5 for building the decision tree (Witten & Frank, 2005). Knowledge flow environment is one of the WEKA environments for performing the algorithm mentioned. Due to the imbalance in the data class distribution, the boosting effective method (Witten & Frank, 2005) which is AdaboostM1 in Weka was used to identify the diabetic patients. The 10-fold Cross Validation method was used to validate the model, and Equations 1, 2 and 3 were used to calculate the Precision, Recall and Accuracy of the model, respectively.

$$\text{Precision} = TP/(TP + FP) \quad (1)$$

$$\text{Recall or Sensitivity} = TP/(TP+FN) \quad (2)$$

and

$$\text{Accuracy} = (TP + TN)/(TP + TN + FP + FN) \quad (3)$$

3. Results

After data preprocessing and preparation, a total of 22,398 records were used for data mining and developing the decision tree. The following results demonstrate the evaluation of the model developed in detail, according to which the precision of the model to identify patients was 0.717 and identifying those who were healthy was better and even more precise than identifying the patients. The Precision of patient identification was, however, just acceptable. The FP rate was low due to the higher identification of healthy individuals (Table 1).

Table 1. The results of the decision tree model evaluation

Evaluation measures	ROC Area	F-Measure	Recall	Precision	Accuracy	FP Rate
results	0.875	0.705	0.694	0.717	0.976	0.012

Note. ROC Area= Receiver Operating Characteristics Area, FP Rate= False Positives Rate.

The results indicated that the area below the ROC curve reached 0.875. The Kappa statistical value was about 0.69. According to the method mentioned, the confusion matrix is as follows:

Table 2. Confusion matrix of the decision tree model

Classes	Diabetic	Healthy
Healthy	253	21221
Diabetic	641	283

As the confusion matrix reveals, the model could separate 98.8 percent (21,221) of healthy individuals from the other patients, while the number of patients identified was 69.4 percent (641 persons), which is less than that of the healthy ones. The whole total of those who belonged in the other class in this incorrect model was 2.4 percent (536 persons) (Table 2). According to the equations mentioned in the Methods Section, the precision and accuracy of the model was 71.7 and 97.6 percent, respectively.

The ROC curve indicates the model function in identification of patients and those individuals who are healthy (Figure 1). The curve indicates high capability of the model, especially in identification of the healthy persons.

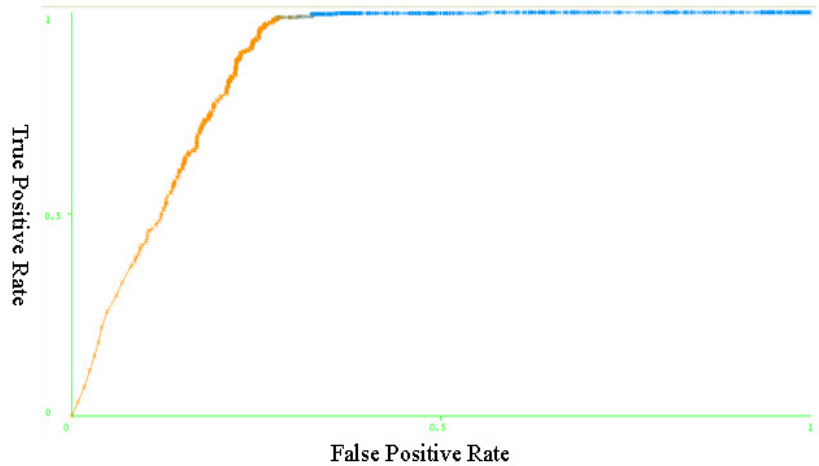

Figure 1. ROC curve-Performance of the J48

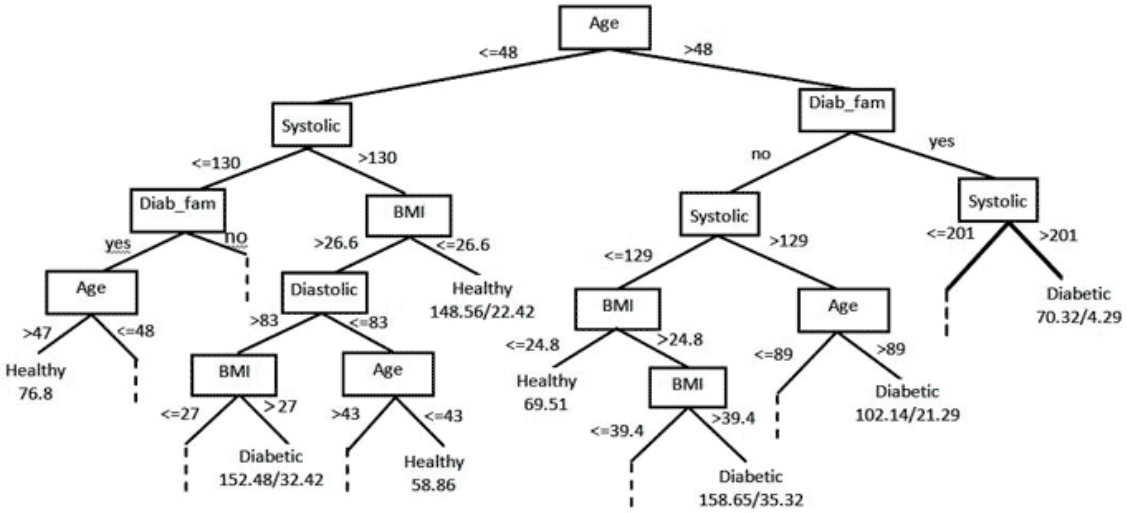

Figure 2. The resulting decision tree. Only the leaf nodes at the higher levels are displayed while the rest are indicated by dashes. The node "Family history of diabetes" was abbreviated to "Diab_fam". The leaf nodes are assigned by one or two numbers where the former and latter indicate the number of correctly and incorrectly classified samples, respectively

The last iteration (10th) of the pruned tree included 289 leaves and the size of the tree was 577. Age feature was placed at the root node of the tree due to higher information gain. The tree was split into two branches: =<48 years old and >48 years old. Those with a family history of diabetes and systolic blood pressure were placed at the next level of the tree (Figure 2).

4. Discussion

T2DM is frequently undiagnosed until complications appear (American Diabetes Association, 2004). According to the Centers for Disease Control and Prevention (CDC), from among the 29.1 millions of US citizens who have diabetes, 8.1 million (27.8%) were undiagnosed in 2012 (Centers for Disease Control Prevention, 2014). Nearly one-third of the patients with T2DM who were newly diagnosed were observed to have complications with microvascular nephropathy, neuropathy and retinopathy (Raman et al., 2009). The results and models of studies similar to this project can facilitate early detection of T2DM and the prevention of potential complications associated with late diagnosis. The results of this study is consistent with findings of Meng et al., who used the features of risk factors such as eating habits, physical activity, and so on for T2DM prediction (Meng, Huang, Rao, Zhang, & Liu, 2013). However, most studies reported higher amounts of evaluation measures of accuracy, precision and sensitivity (Kahramanli & Allahverdi, 2008; Kayaer & Yıldırım, 2003; Lekkas & Mikhailov, 2010; K. Polat, Gnnes, & Arslan, 2008; Kemal Polat & Güneş, 2007; Prati, Batista, & Monard, 2009; Temurtas, Yumusak, & Temurtas, 2009). This was probably because of the features used in T2DM prediction. The features used in this study for T2DM prediction were indeed the same as those used in diabetes screening and recommended by most well-known authorities (American Diabetes Association, 2004; FRCSC, 2008; National Diabetes Information Clearinghouse (NDIC), 2014). Prior studies also have confirmed that these features are important predictor variables (Dong et al., 2011; Glumer et al., 2004; Li, Bergmann, Reimann, Bornstein, & Schwarz, 2009; Robinson, Agarwal, & Nerenberg, 2011). These studies, which aimed to identify and score the main variables affecting the development of diabetes, identified age and BMI as the main predictor variables and blood pressure measure, family history of DM and sex as the highest risk factor scores for the detection of undiagnosed diabetes (Brown et al., 2012).

The Pima Indians dataset has been used widely for data mining on diabetes mellitus. Out of the nine features two include plasma glucose and serum insulin. Findings from the past studies based on the Pima Indians or other datasets reported prediction models with high accuracy levels (Huang, McCullagh, Black, & Harper, 2007; Kahramanli & Allahverdi, 2008; Tama, Rodiyatul, & Hermansyah, 2013; Temurtas et al., 2009). This was due to the inclusion of plasma glucose and serum insulin. Fauci believes that the diagnosis of endocrine diseases is not yet established using the symptoms instead, and diagnosis is made by measuring the hormone levels secreted or their target (Fauci, 2008). The main diagnosis of diabetes mellitus is based on several techniques for measuring

the plasma glucose or serum insulin level, confirming that the selection and application of the predictor features requires further attention.

Therefore, the prior researches, especially those which had employed those features related to the main diagnosis of diabetes, have compared the capability of the data mining techniques and algorithms and it was found that they were not to be considered in diabetes prediction or diagnosis. The current study tested the decision tree using real data and the features used in diabetes screening communicated by the Health Ministry of Iran to the health centers of the provinces and cities as defined diabetes risk factors so that the primary screening was performed by evaluating such features in those individuals who referred. The study demonstrated that the decision tree could be used in the screening and that it would help in patient screening by automating the screenings in the electronic systems.

5. Conclusion

We developed a model using the decision tree for the screening of T2DM that does not require laboratory tests for T2DM diagnosis. We used the J48 algorithm and the model proposed is different from the previous models for three reasons: 1) We used real dataset. 2) We used the features applied to primary screening, excluding those such as plasma glucose for the main diagnosis of T2DM. 3) Capability of the decision trees for T2DM screening. Although the exclusion of diabetes laboratory diagnostic tests features lowered the sensitivity and precision of the model proposed compared with models suggested in the literatures, this study is a step forward for the early diagnosis of diabetes without using diagnostic laboratory tests.

Acknowledgments

This study was a part of PhD thesis supported and funded by Iran University of Medical Sciences (Grant No. 233). We express our deep gratitude to Dr. Zeinalzadeh, Dr. Sherbaf and other staff members of the Center for Non-communicable Diseases Control of the Tabriz University of Medical Sciences for permitting us to use the data and their assistance in retrieving the patients' data.

References

American Diabetes Association. (2004). Screening for type 2 diabetes. *Diabetes Care, 27*, S11. http://dx.doi.org/10.2337/diacare.27.2007.S11

American Diabetes Association. (2013). Diagnosis and classification of diabetes mellitus. *Diabetes Care, 36*(Supplement 1), S67-S74. http://dx.doi.org/10.2337/dc13-S067

Bellazzi, R., Ferrazzi, F., & Sacchi, L. (2011). Predictive data mining in clinical medicine: a focus on selected methods and applications. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 1*(5), 416-430.

Brown, N., Critchley, J., Bogowicz, P., Mayige, M., & Unwin, N. (2012). Risk scores based on self-reported or available clinical data to detect undiagnosed Type 2 Diabetes: A systematic review. *Diabetes research and clinical practice, 98*(3), 369-385. http://dx.doi.org/10.1016/j.diabres.2012.09.005

Centers for Disease Control Prevention. (2014). *National diabetes statistics report: estimates of diabetes and its burden in the United States, 2014*. Atlanta, GA: US Department of Health and Human Services, 8.

DeVoe, J. E., Gold, R., McIntire, P., Puro, J., Chauvie, S., & Gallia, C. A. (2011). Electronic health records vs Medicaid claims: completeness of diabetes preventive care data in community health centers. *The Annals of Family Medicine, 9*(4), 351-358. http://dx.doi.org/10.1370/afm.1279

Dong, J.-j., Lou, N.-j., Zhao, J.-j., Zhang, Z.-w., Qiu, L.-l., Zhou, Y. et al. (2011). Evaluation of a risk factor scoring model in screening for undiagnosed diabetes in China population. *Journal of Zhejiang University SCIENCE B, 12*(10), 846-852. http://dx.doi.org/10.1631/jzus.B1000390

Fauci, A. S. (2008). *Harrison's principles of internal medicine* (Vol. 2). New York: McGraw-Hill Medical.

Frcsc, F. A. M. (2008). Canadian Diabetes Association Clinical Practice Guidelines Expert Committee. *Canadian Journal of Diabetes, 32*, 134.

Glumer, C., Carstensen, B., Sandbaek, A., Lauritzen, T., Jorgensen, T., & Borch-Johnsen, K. (2004). A Danish diabetes risk score for targeted screening: The Inter99 study. *Diabetes Care, 27*(3), 727-733. http://dx.doi.org/10.2337/diacare.27.3.727

Gregg, E. W., Geiss, L. S., Saaddine, J., Fagot-Campagna, A., Beckles, G., Parker, C. et al. (2001). Use of diabetes preventive care and complications risk in two African-American communities. *American journal of preventive medicine, 21*(3), 197-202. http://dx.doi.org/10.1016/S0749-3797(01)00351-8

Heydari, I., Radi, V., Razmjou, S., & Amiri, A. (2010). Chronic complications of diabetes mellitus in newly diagnosed patients. *International Journal of Diabetes Mellitus, 2*(1), 61-63. http://dx.doi.org/10.1016/j.ijdm.2009.08.001

Huang, Y., McCullagh, P., Black, N., & Harper, R. (2007). Feature selection and classification model construction on type 2 diabetic patients' data. *Artificial Intelligence in Medicine, 41*(3), 251-262. http://dx.doi.org/10.1016/j.artmed.2007.07.002

Kahramanli, H., & Allahverdi, N. (2008). Design of a hybrid system for the diabetes and heart diseases. *Expert Systems with Applications, 35*(1), 82-89. http://dx.doi.org/10.1016/j.eswa.2007.06.004

Karter, A. J., Stevens, M. R., Herman, W. H., Ettner, S., Marrero, D. G., Safford, M. M. et al. (2003). Out-of-Pocket Costs and Diabetes Preventive Services The Translating Research Into Action for Diabetes (TRIAD) study. *Diabetes Care, 26*(8), 2294-2299. http://dx.doi.org/10.2337/diacare.26.8.2294

Kayaer, K., & Yıldırım, T. (2003). *Medical diagnosis on Pima Indian diabetes using general regression neural networks.* Paper presented at the Proceedings of the international conference on artificial neural networks and neural information processing (ICANN/ICONIP).

King, H., Aubert, R. E., & Herman, W. H. (1998). Global burden of diabetes, 1995-2025: Prevalence, numerical estimates, and projections. *Diabetes Care, 21*(9), 1414-1431. http://dx.doi.org/10.2337/diacare.21.9.1414

Lekkas, S., & Mikhailov, L. (2010). Evolving fuzzy medical diagnosis of Pima Indians diabetes and of dermatological diseases. *Artificial Intelligence in Medicine, 50*(2), 117-126.

Li, J., Bergmann, A., Reimann, M., Bornstein, S. R., & Schwarz, P. E. (2009). A more simplified Finnish diabetes risk score for opportunistic screening of undiagnosed type 2 diabetes in a German population with a family history of the metabolic syndrome. *Horm Metab Res, 41*(2), 98-103. http://dx.doi.org/10.1055/s-0028-1087191

Liao, S.-H., Chu, P.-H., & Hsiao, P.-Y. (2012). Data mining techniques and applications–A decade review from 2000 to 2011. *Expert Systems with Applications, 39*(12), 11303-11311. http://dx.doi.org/10.1016/j.eswa.2012.02.063

Meng, X.-H., Huang, Y.-X., Rao, D.-P., Zhang, Q., & Liu, Q. (2013). Comparison of three data mining models for predicting diabetes or prediabetes by risk factors. *The Kaohsiung Journal of Medical Sciences, 29*(2), 93-99. http://dx.doi.org/10.1016/j.kjms.2012.08.016

National Diabetes Information Clearinghouse (NDIC) (Producer). (2014) *Am I at risk for type 2 diabetes*? Retrieved from http://www.diabetes.niddk.nih.gov/dm/pubs/riskfortype2/index.aspx

Polat, K., Gnnes, S., & Arslan, A. (2008). A cascade learning system for classification of diabetes disease: Generalized Discriminant Analysis and Least Square Support Vector Machine. *Expert Systems with Applications, 34*(1), 482-487. http://dx.doi.org/10.1016/j.eswa.2006.09.012

Polat, K., & Güneş, S. (2007). An expert system approach based on principal component analysis and adaptive neuro-fuzzy inference system to diagnosis of diabetes disease. *Digital Signal Processing, 17*(4), 702-710. http://dx.doi.org/10.1016/j.dsp.2006.09.005

Prati, R. C., Batista, G. E., & Monard, M. C. (2009). *Data mining with imbalanced class distributions: Concepts and methods.* Paper presented at the IICAI.

Raman, R., Rani, P. K., Reddi Rachepalle, S., Gnanamoorthy, P., Uthra, S., Kumaramanickavel, G. et al. (2009). Prevalence of diabetic retinopathy in India: Sankara Nethralaya Diabetic Retinopathy Epidemiology and Molecular Genetics Study report 2. *Ophthalmology, 116*(2), 311-318. http://dx.doi.org/10.1016/j.ophtha.2008.09.010

Robinson, C. A., Agarwal, G., & Nerenberg, K. (2011). Validating the CANRISK prognostic model for assessing diabetes risk in Canada's multi-ethnic population. *Chronic Dis Inj Can, 32*(1), 19-31.

Tama, B. A., Rodiyatul, F., & Hermansyah, H. (2013). An Early Detection Method of Type-2 Diabetes Mellitus in Public Hospital. *TELKOMNIKA (Telecommunication Computing Electronics and Control), 9*(2), 287-294.

Temurtas, H., Yumusak, N., & Temurtas, F. (2009). A comparative study on diabetes disease diagnosis using neural networks. *Expert Systems with Applications, 36*(4), 8610-8615. http://dx.doi.org/10.1016/j.eswa.2008.10.032

Upadhyaya, S., Farahmand, K., & Baker-Demaray, T. (2013). Comparison of NN and LR classifiers in the context of screening native American elders with diabetes. *Expert Systems with Applications, 40*(15), 5830-5838. http://dx.doi.org/10.1016/j.eswa.2013.05.012

Witten, I. H., & Frank, E. (2005). *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann.