طراحی سیستم هوشمند برای تشخیص بیماری دیابت با استفاده از رویکرد داده‌کاوی: *گزارش کوتاه*

روح‌اله کلهر[۱]، اصغر مرتضی‌قلی[۲]
فاطمه ناجی[۳]، سعید شهسواری[٤]
محمد زکریا کیایی[۵*]

۱- مرکز تحقیقات عوامل اجتماعی موثر برسلامت، دانشگاه علوم پزشکی قزوین، قزوین، ایران.
۲- گروه هوش مصنوعی، رایانه و فناوری اطلاعات، دانشکده برق، دانشگاه آزاد اسلامی قزوین، قزوین، ایران.
۳- گروه اپیدمیولوژی، دانشگاه علوم پزشکی قزوین، قزوین، ایران.
٤- مربی آمار زیستی، مرکز تحقیقات ایمنی محصولات بهداشتی، دانشگاه علوم پزشکی قزوین، قزوین، ایران.
۵- مربی مدیریت خدمات بهداشتی و درمانی، دانشکده بهداشت، دانشگاه علوم پزشکی قزوین، قزوین، ایران.

*نویسنده مسئول: قزوین، بلوار شهید باهنر، دانشگاه علوم پزشکی قزوین، دانشکده بهداشت، گروه مدیریت خدمات بهداشتی و درمانی.
کد پستی: ۳٤۱۹۷۵۹۸۱۱
تلفن: ۳۳۳٦۹۵۸۱ -۰۲۸
E-mail: kiaei_mzsa@yahoo.com

دریافت: ۱۳۹۷/۰٦/۲٦ ویرایش: ۱۳۹۷/۰۷/۰۳ پذیرش: ۱۳۹۷/۱۲/۰۳ آنلاین: ۱۳۹۷/۱۲/۱۰

چکیده

زمینه و هدف: بیماری دیابت عوارض متعددی دارد، تشخیص دیر هنگام دیابت در افراد منجر به گسترش عوارض بیماری می‌شود. مطالعه حاضر با هدف بررسی امکان پیش‌بینی دیابت با استفاده از فنون داده‌کاوی انجام شد.

روش بررسی: این پژوهش از نوع توصیفی- تحلیلی بود که به‌صورت مقطعی انجام شد. جامعه پژوهش شامل افراد مراجعه‌کننده به مراکز بهداشتی شهرستان محمدیه در استان قزوین جهت انجام غربالگری دیابت بودند. داده‌های مورد مطالعه مربوط به فروردین تا خرداد ۱۳۹٤ بود. داده‌ها در نهایت با استفاده از سه روش نزدیک‌ترین همسایگی (k-nearest neighbors algorithm, k-NN)، درخت تصمیم‌گیری (Decision tree, DT) و ماشین‌های بردار پشتیبان (Support vector machine, SVM) تحلیل و مورد مقایسه قرار گرفتند. جهت تحلیل داده‌ها از MATLAB® software, version 8.2 (Mathworks Inc., Natick, MA, USA) استفاده شد.

یافته‌ها: در تمامی معیارها، بهترین نتایج توسط درخت تصمیم‌گیری با صحت (۰/۹٦) به‌دست آمد. پس از آن روش‌های نزدیک‌ترین همسایگی با صحت (۰/۹٦) و ماشین‌های بردار پشتیبان با صحت (۰/۹٤) قرار داشتند.

نتیجه‌گیری: براساس نتایج ارایه شده، درخت تصمیم‌گیری بهترین نتایج را در کلاس‌بندی نمونه‌های تست نشان داد. این مدل می‌تواند به‌عنوان مدلی مناسب در پیش‌بینی دیابت با استفاده از داده‌های ریسک فاکتور توصیه شود.

کلمات کلیدی: هوش مصنوعی، داده‌کاوی، دیابت ملیتوس نوع دو، تشخیص زودهنگام، یادگیری ماشینی، عوامل خطر.

مقدمه

دیابت ملیتوس بیماری مزمن است که بروز و شیوع آن به‌طور چشمگیری در دهه‌های اخیر به‌دلیل تغییر در سبک زندگی، رواج چاقی و افزایش طول عمر بیشتر شده است تشخیص دیرهنگام یا عدم تشخیص دیابت در افراد منجر به گسترش عوارض مختلف عروقی مزمن می‌شود.[۱و۲] بیماری دیابت عوارض متعددی دارد که معمولا برگشت‌ناپذیر هستند.[۳] دیابت نوع دو در بیشتر مواقع تشخیص داده نمی‌شود تا اینکه عوارض ظاهر شود.[٤] تشخیص و پیشگیری به‌موقع باعث کاهش مرگ‌ومیر و همچنین جلوگیری و کاهش عوارض دیابت و بهبود کیفیت زندگی می‌شود.[۵] روش‌های داده‌کاوی در سال‌های اخیر در حوزه پزشکی و مراقبت‌های بهداشتی در زمینه تشخیص و پیشگیری بیماری و انتخاب روش درمان و پیش‌بینی مرگ‌ومیر و پیش‌بینی هزینه‌های درمانی به‌طور گسترده‌ای مورد استفاده قرار گرفته است.[٦] داده‌کاوی و کشف دانش می‌تواند برای خودکارسازی کار تشخیص در پزشکی مورد استفاده قرار گیرد.[۷]

سیستم خبره در مراقبت سلامت برای بهبود کیفیت، امنیت و کارآیی سیستم داده‌های سلامت در مراقبت از بیماران به‌کار گرفته می‌شود.[۸] مطالعات گوناگونی در حوزه تشخیص بیماری دیابت با استفاده از روش داده‌کاوی با رویکردهای متنوع صورت گرفته است.[۹و۱۰] مطالعه حاضر با هدف بررسی امکان پیش‌بینی دیابت با استفاده از فنون داده‌کاوی و ویژگی‌های مربوط به ریسک فاکتورهای انجام شد.

روش بررسی

این پژوهش از نوع کاربردی و به روش توصیفی- تحلیلی گذشته‌نگر بر اساس پنج گام مدل CRISP انجام شد. این پنج گام شامل شناسایی سیستم، شناخت و آماده‌سازی داده‌ها، مدل‌سازی، ارزیابی و توسعه می‌باشد. جامعه پژوهش شامل افراد مراجعه‌کننده به مراکز بهداشتی شهرستان محمدیه جهت انجام غربالگری دیابت بودند. در این مطالعه تعداد ۱۰۵۵ از افراد که دارای اطلاعات کامل بودند به‌عنوان نمونه وارد مطالعه شدند. از این تعداد از نظر تشخیص دیابت و یا سالم بودن، ۱۵۹ نفر سالم و ۸۹٦ نفر دیابتی تشخیص داده شده بودند. داده‌های مورد مطالعه مربوط به فروردین تا خرداد ماه ۱۳۹٤ بوده است.

ابزار گردآوری داده‌ها، چک لیست پژوهشگر ساخته بود که با استفاده از آن ویژگی‌های سن، جنس، فشارخون سیستولیک، فشارخون دیاستولیک، سابقه خانوادگی دیابت، شاخص توده بدنی، قد، وزن، دور کمر، دور باسن و تشخیص به‌عنوان متغیرهای مطالعه گردآوری شدند. برای حل مشکل عدم توازن کلاس‌ها از روش‌های شناخته شده باز نمونه‌گیری و روش مبتنی بر الگوریتم استفاده شد. نتایج به‌دست‌آمده در این مطالعه بر اساس روش اعتبارسنجی 10-fold cross-validation به‌دست آمده است. در این روش ابتدا کل پایگاه داده به دو مجموعه آموزشی و تست تقسیم شده، سپس مجموعه آموزشی به ۱۰ بخش تقسیم می‌شود و در هر بار تکرار، یک بخش از ۱۰ بخش به‌عنوان مجموعه اعتبارسنجی و مابقی ۹ بخش به‌عنوان مجموعه آموزشی انتخاب می‌شود.

لازم به یادآوری است ۷۰٪ از کل نمونه‌های پایگاه داده برای مجموعه آموزشی و ۳۰٪ باقی‌مانده برای مجموعه تست تعیین شده‌اند. داده‌ها در نهایت با استفاده از سه روش نزدیک‌ترین همسایگی (k-nearest neighbors algorithm, k-NN)، درخت تصمیم‌گیری (Decision tree, DT) و ماشین‌های بردار پشتیبان (Support vector machine, SVM) تحلیل و مورد مقایسه قرار گرفتند. جهت تحلیل داده‌ها از MATLAB® software, version 8.2 (Mathworks Inc., Natick, MA, USA) استفاده شد. لازم به یادآوری است که این پژوهش با حمایت مالی دانشگاه علوم پزشکی قزوین با شماره طرح ب/٦٤ و شناسه اختصاصی کمیته اخلاق (۲۸/۲۰/۱۰۰٤۲) انجام شد.

یافته‌ها

در جدول ۱ نتایج هر سه روش کلاس‌بندی براساس چهار معیار ارزیابی ذکرشده مورد مقایسه قرار گرفته است. براساس این جدول در تمامی معیارها، بهترین نتایج توسط درخت تصمیم‌گیری با صحت (۰/۹٦) و دقت (۰/۸۹) به‌دست آمد. پس از آن روش‌های نزدیک‌ترین همسایگی با صحت (Accuracy) (۰/۹٦) و دقت (۰/۸۳) و ماشین‌های بردار پشتیبان با صحت (۰/۹٤) و دقت (۰/۸۵) قرار داشتند.

جدول ۱: بررسی نتایج روش پیشنهادی به‌منظور کلاس‌بندی بیماران دیابتی براساس چهار معیار ارزیابی

فنون داده‌کاوی	معیارهای ارزیابی			
	صحت (درصد)	نمره F (درصد)	فراخوانی (درصد)	دقت (درصد)
نزدیکترین همسایگی	۰/۹٦	۰/۸٤	۰/۸٦	۰/۸۳
درخت تصمیم‌گیری	۰/۹٦	۰/۸۷	۰/۸٦	۰/۸۹
ماشین‌های‌بردار پشتیبان	۰/۹٤	۰/۸۲	۰/۸٦	۰/۸۵

در دومین مرحله از ارزیابی، نتایج تمامی روش‌های کلاس‌بندی بر اساس معیار ماتریس اغتشاش مورد مقایسه قرار گرفته‌اند. در این مرحله تعداد کل نمونه‌های تست ۳۱۶ نمونه بود که از بین آن‌ها تعداد ۲۶۱ نمونه مربوط به کلاس دیابت و ۵۵ نمونه مربوط به کلاس سالم بودند.

در این پژوهش درخت تصمیم‌گیری بیشترین میزان دقت کلاس‌بندی را برای هر دو کلاس دیابت و سالم به‌دست آورده بود، ماتریس اغتشاش در الگوریتم درخت تصمیم‌گیری تعداد افراد سالم درست تشخیص داده شده ۴۷ مورد و تعداد افراد دیابتی درست تشخیص داده شده ۲۵۲ مورد نشان داد. در الگوریتم نزدیکترین همسایه تعداد افراد سالم درست تشخیص داده شده ۴۱ مورد و تعداد افراد دیابتی درست تشخیص داده شده ۲۴۴ بود. در نهایت ماتریس اغتشاش در الگوریتم ماشین‌بردار پشتیبان تعداد افراد سالم درست تشخیص داده شده ۴۵ مورد و تعداد افراد دیابتی درست تشخیص داده شده ۲۵۰ مورد بود.

بحث

تحلیل یافته‌های این مطالعه نشان داد که بهترین نتایج در هر چهار معیار مورد مطالعه توسط درخت تصمیم‌گیری به‌دست آمده است. پس از آن روش‌های نزدیک‌ترین همسایگی و ماشین‌های بردار پشتیبان قرار داشتند. یافته‌های مطالعه Habibi و Huang و همکاران یافته‌های این مطالعه را تایید می‌نمایند و نتایج این مطالعات نشان می‌دهد که درخت تصمیم‌گیری عملکرد بهتری نسبت به سایر مدل‌های مورد مقایسه داشته است.[۱۱و۱۲] با این‌حال پژوهش‌هایی نیز بر توانایی بالاتر سایر فنون نسبت به درخت تصمیم تاکید کردند از جمله آن‌ها Aruna و همکاران در مطالعه مقایسه‌ای فنون داده‌کاوی بالاترین دقت و حساسیت را مربوط به ماشین‌بردار پشتیبان ذکر کرد، پس از آن بیز ساده و سپس درخت تصمیم با الگوریتم J48 قرار گرفت.[۱۳] همچنین Jeatrakul و همکاران در مطالعه خود دریافتند که یافته‌های مدل شبکه عصبی نسبت به سایر مدل‌ها درستی بهتری نشان می‌دهد.[۱۴] نتایج مطالعه Tama و همکار در مقایسه عملکرد بین فنون دسته‌بندی درخت تصمیم با الگوریتم J48، ماشین‌بردار پشتیبان و بیز ساده نشان داد که تفاوت چندانی در صحت دسته‌بندی این سه مدل وجود ندارد.[۱۵] به‌نظر می‌رسد این اختلاف در نتایج مطالعات مختلف شاید ناشی از تفاوت در داده‌های مورد استفاده هر مطالعه باشد و همین داده‌های متنوع موجب عملکرد متفاوت مدل‌های مختلف شده است. بنابراین هم داده‌های مورد استفاده و هم ویژگی‌های استفاده شده در مدل در نتیجه نهایی مطالعه موثر است.

براساس نتایج ارایه شده، مدل درخت تصمیم‌گیری با صحت (۰/۹۶) و دقت (۰/۸۹) بهترین نتایج را در کلاس‌بندی نمونه‌های تست نشان می‌دهد.

References

1. Tuomilehto J, Lindström J, Eriksson JG, Valle TT, Hämäläinen H, Ilanne-Parikka P, et al. Prevention of type 2 diabetes mellitus by changes in lifestyle among subjects with impaired glucose tolerance. *N Engl J Med* 2001;344(18):1343-50.
2. Heydari I, Radi V, Razmjou S, Amiri A. Chronic complications of diabetes mellitus in newly diagnosed patients. *Int J Diabetes Mellit* 2010;2(1):61-3.
3. Luijks H, Schermer T, Bor H, van Weel C, Lagro-Janssen T, Biermans M, et al. Prevalence and incidence density rates of chronic comorbidity in type 2 diabetes patients: an exploratory cohort study. *BMC Med* 2012;10(1):128.
4. Beagley J, Guariguata L, Weil C, Motala AA. Global estimates of undiagnosed diabetes in adults. *Diabetes Res Clin Pract* 2014;103(2):150-60.
5. Zhuo X, Zhang P, Hoerger TJ. Lifetime direct medical costs of treating type 2 diabetes and diabetic complications. *Am J Prev Med* 2013;45(3):253-61.
6. Rafeh R, Arbabi M. Data mining techniques to diagnose diabetes using blood lipids. *J Ilam Univ Med Sci* 2015;23(4):239-47.
7. Soni J, Ansari U, Sharma D, Soni S. Predictive data mining for medical diagnosis: An overview of heart disease prediction. *Int J Comput Appl* 2011;17(8):43-8.
8. Rajalakshmi K, Chandra Mohan S, Dhinesh BS. Decision Support System in Healthcare Industry. *Int J Comput Appl* 2011;26(9):42-4.
9. Temurtas H, Yumusak N, Temurtas F. A comparative study on diabetes disease diagnosis using neural networks. Expert Syst Appl 2009;36(4):8610-5.
10. Habibi S, Ahmadi M, Alizadeh S. Type 2 diabetes mellitus screening and risk factors using decision tree: results of data mining. *Glob J Health Sci* 2015;7(5):304-10.
11. Habibi S. A study on diabetes type II predictive models applying data mining techniques in expert systems development [Dissertation]. Tehran: Iran University of Medical Sciences; School of Health Management and Information Science: 2015. [Persian]
12. Huang Y, McCullagh P, Black N, Harper R. Feature selection and classification model construction on type 2 diabetic patients' data. *Artif Intell Med* 2007;41(3):251-62.

13. Aruna S, Rajagopalan S, Nandakishore L. An empirical comparasion of supervised learning algorithms in disease detection. Int J Inf Technol Converg Serv 2011;1:81-92.
14. Jeatrakul P, Wong KW, Fung CC. Classification of imbalanced data by combining the complementary neural network and SMOTE algorithm. In: Neural Information Processing Models and Applications. Springer; 2010. P. 152-9.
15. Tama BA, Rodiyatul F, Hermansyah H. An early detection method of type-2 diabetes mellitus in public hospital. *TELKOMNIKA* 2011;9(2):287-94.

TUMJ Tehran University Medical Journal, March 2019; Vol. 76, No. 12: 827-831 Brief Report

Designing an intelligent system for diagnosing type 2 diabetes using the data mining approach: *brief report*

Rohollah Kalhor Ph.D.[1]
Asghar Mortezagholi M.Sc.[2]
Fatemeh Naji M.Sc., M.P.H.[3]
Saeed Shahsavari M.Sc.[4]
Mohammad Zakaria Kiaei M.Sc.[5*]

1- Social Determinants of Health Research Center, Qazvin University of Medical Sciences, Qazvin, Iran.
2- Department of Artificial Intelligence, Computer and Information Technology (IT) Engineering, Faculty of Electrical, Qazvin Islamic Azad University, Qazvin, Iran.
3- Department of Epidemiology, Qazvin University of Medical Sciences, Qazvin, Iran.
4- Instructor of Biostatistics, Health Products Safety Research Center, Qazvin University of Medical Sciences, Qazvin, Iran.
5- Instructor in Health Services Management, School of Health, Qazvin University of Medical Sciences, Qazvin, Iran.

[*]Corresponding author: Department of Health Services Management, School of Health, Qazvin University of Medical Sciences, Shahid Bahonar Blvd., Qazvin, Iran.
Post code: 3419759811
Tel: +98- 28- 33369581
E-mail: kiaei_mzsa@yahoo.com

Abstract

Received: 17 Sep. 2018 Revised: 25 Sep. 2018 Accepted: 22 Feb. 2019 Available online: 01 Mar. 2019

Background: Diabetes mellitus has several complications. The Late diagnosis of diabetes in people leads to the spread of complications. Therefore, this study has been done to determine the possibility of predicting diabetes type 2 by using data mining techniques.

Methods: This is a descriptive-analytic study that was conducted as a cross-sectional study. The study population included people referring to health centers in Mohammadieh City in Qazvin Province, Iran, from April to June 2015 for screening for diabetes. The 5-step CRISP method was used to implement this study. Data were collected from March 2015 to June 2015. In this study, 1055 persons with complete information were included in the study. Of these, 159 were healthy and 896 were diabetic. A total of 11 characteristics and risk factors were examined, including the age, sex, systolic and diastolic blood pressure, family history of diabetes, BMI, height, weight, waistline, hip circumference and diagnosis. The results obtained by support vector machine (SVM), decision tree (DT) and the k-nearest neighbors algorithm (k-NN) were compared with each other. Data was analyzed using MATLAB® software, version 3.2 (Mathworks Inc., Natick, MA, USA).

Results: Data analysis showed that in all criteria, the best results were obtained by decision tree with accuracy (0.96) and precision (0.89). The k-NN methods were followed by accuracy (0.96) and precision (0.83) and support vector machine with accuracy (0.94) and precision (0.85). Also, in this study, decision tree model obtained the highest degree of class accuracy for both diabetes classes and healthy in the analysis of confusion matrix.

Conclusion: Based on the results, the decision tree represents the best results in the class of test samples which can be recommended as a model for predicting diabetes type 2 using risk factor data.

Keywords: artificial intelligence, data mining, diabetes mellitus type 2, early diagnosis, machine learning, risk factors.