مجله دانشکده پزشکی، دانشگاه علوم پزشکی تهران، مرداد ۱۳۹۸، دوره ۷۷، شماره ۵، صفحه‌های ۳۰۱ تا ۳۰۷

مقاله اصیل

مقایسه کارایی روش‌های داده‌کاوی در پیش‌بینی ابتلا به بیماری دیابت نوع ۲

حسین تیره[۱]
محمدتقی شاکری[۲]
صادق رسولی‌نژاد[۱]
حبیب‌اله اسماعیلی[۲]
راضیه یوسفی[۱*]

۱- گروه اپیدمیولوژی و آمار زیستی، دانشکده بهداشت، دانشگاه علوم پزشکی مشهد، مشهد، ایران.
۲- مرکز تحقیقات عوامل اجتماعی موثر بر سلامت، دانشگاه علوم پزشکی مشهد، مشهد، ایران.

[*] نویسنده مسئول: مشهد، خیابان دانشگاه، دانشکده بهداشت، گروه اپیدمیولوژی و آمار زیستی.
کدپستی: ۵۱۳۷۶۷۳۱۱۹
تلفن: ۰۵۱-۳۱۸۹۲۷۰۰
E-mail: yousefir9@mums.ac.ir

چکیده

دریافت: ۱۳۹۷/۱۲/۱۱ ویرایش: ۱۳۹۷/۱۲/۱۸ پذیرش: ۱۳۹۸/۰۵/۲۱ آنلاین: ۱۳۹۸/۰۵/۳۱

زمینه و هدف: دیابت بیماری مزمنی است که شایع‌ترین بیماری اختلالات متابولیسم و یکی از عمده‌ترین مسایل بهداشتی-درمانی در سرتاسر دنیا است. امروزه تکنیک‌های داده‌کاوی (Data mining techniques) با توجه به توانمندی آن‌ها، در حوزه‌های مختلف به‌کار برده می‌شوند. بنابراین در این پژوهش به مقایسه کارایی روش‌های داده‌کاوی در پیش‌بینی ابتلا به بیماری دیابت نوع ۲ پرداخته شد.

روش بررسی: در این مطالعه تحلیلی-مقطعی، از مجموع ۷۰۰۰ نفر شرکت‌کننده در طرح غربالگری دیابت سال ۱۳۹۵ در منطقه ثامن مشهد که شامل بررسی میزان قندخون، چشم‌ها، آسیب کلیوی و سلامت پاها بود، تعداد ۵۴۰ نفر مبتلا به بیماری دیابت (درمان نشده) بودند و برای حفظ توازن ۶۰۰ فرد سالم به‌صورت نمونه‌گیری متناسب با حجم انتخاب شدند (حجم کل نمونه ۱۱۴۰ نفر). در این مطالعه مبتلایان به دیابت بالای ۳۰ سال وارد مطالعه گردیدند و افرادی که از پیش سابقه‌ی دیابت نوع ۲ داشتند و اکنون به‌سبب استفاده از دارو و یا رعایت سایر نکات مقدار قندخون طبیعی داشتند از مطالعه حذف گردیدند.

یافته‌ها: صحت هر سه مدل برابر بود (۸۶٪) اما از نظر سطح زیرمنحنی راک (Receiver operating characteristic, ROC)، مدل‌های رگرسیون لجستیک (Logistic regression model) و بیزین ساده (Simple Bayesian model) با ۹۰٪ سطح زیرمنحنی راک عملکرد بهتری داشتند. در مدل بیزین ساده و رگرسیون لجستیک به‌ترتیب متغیرهای شاخص توده بدنی و سن دارای بیشترین میزان اهمیت بودند، درحالی‌که در مدل ماشین‌بردار پشتیبان (Support vector machine, SVM)، متغیرهای شاخص توده بدنی و فشارخون دارای بیشترین اهمیت بودند.

نتیجه‌گیری: صحت هر سه مدل برابر بود. از نظر سطح زیرمنحنی راک، مدل‌های رگرسیون لجستیک و بیزین ساده مناسب‌تر بودند. در مجموع هر سه مدل عملکرد مشابهی داشتند. در هر سه مدل شاخص توده بدنی دارای بیشترین اهمیت بود.

کلمات کلیدی: داده‌کاوی، دیابت شیرین، بیماری‌های متابولیک، حساسیت و ویژگی.

مقدمه

بیماری دیابت شایع‌ترین بیماری ناشی از اختلالات متابولیسم و همچنین یکی از علت‌های مرگ‌ومیر در جوامع است. عوارض زیاد این بیماری هم از نظر هزینه‌های درمانی و هم از جهت از کارافتادگی که ایجاد می‌کند، موجب تبدیل آن به یکی از عمده‌ترین مسایل بهداشتی-درمانی شده است. سازمان بهداشت جهانی با توجه به آمار و روند روبه‌افزایش دیابت در سراسر جهان، این بیماری را به‌عنوان یک اپیدمی نهفته اعلام کرد و از سال ۱۹۹۳ تمام کشورهای جهان را به مقابله با این همه‌گیری فراخواند. میزان شیوع دیابت نوع ۲ در کل جمعیت بین ۳ تا ٤٪ و در افراد بالای ٤۱ سال بین ۵ تا ۱۰٪ گزارش شده است.[۱و۲] همچنین براساس سرشماری آماری که در سال ۲۰۱۳ صورت گرفته، بیش از ٦ میلیون نفر

ایرانی مبتلا به دیابت هستند.[۳] براساس گزارش سازمان بهداشت جهانی در سال ۲۰۱۶ شیوع دیابت در ایران ۱۰/۳٪ اعلام شده است.[۴]

تشخیص بیماری دیابت و یا آگاهی یافتن از احتمال بالای ابتلا به این بیماری به‌طور معمول سخت است، زیرا علایم این بیماری گاهی اوقات با سایر بیماری‌ها مشترک است. بنابراین نیاز به تکنیک‌هایی وجود دارد که بتواند این تشخیص‌ها را ممکن سازد و دقت این کار را بالاتر ببرد. سیستم‌های پشتیبان تصمیم‌گیری که در حل مسایل و تصمیم‌گیری‌های پیچیده به‌کمک انسان‌ها آمده‌اند به‌تازگی برای تشخیص بیماری‌ها مورد توجه تعداد زیادی از پژوهشگران قرار گرفته‌اند. یکی از روش‌های بسیار مهمی که با آن می‌توان الگوهای مفیدی را در میان داده‌های بیماران شناسایی کرد، تکنیک‌های داده‌کاوی (Data mining techniques) است.[۵] داده‌کاوی در پزشکی در پیشگیری و یا تشخیص نوع بیماری‌ها و انتخاب روش‌های درمان بیماری‌ها کاربرد دارد.[۶] در سالیان اخیر استفاده از تکنیک‌های داده‌کاوی در مطالعات روبه‌افزایش بوده همچنان که Liao و همکاران در سال ۲۰۱۲ به بررسی پژوهش‌های حوزه داده‌کاوی در طی یک دهه از سال ۲۰۰۰ تا ۲۰۱۱ پرداخته‌اند. در این مدت ۲۵۶ مقاله به تکنیک‌های داده‌کاوی و کاربردهای آن پرداخته بودند.[۷]

داده‌کاوی شیوه‌های مختلفی دارد که در حوزه‌های مختلف علوم به‌کار گرفته شده‌اند. از جمله این روش‌ها می‌توان به‌روش ماشین‌بردار پشتیبان (Support vector machine, SVM)، رگرسیون لجستیک (Logistic regression model)، درخت تصمیم (Decision tree)، جنگل تصادفی (Random forest) و بیزین ساده (Simple Bayesian model) اشاره کرد که هر کدام قابلیت‌ها و ویژگی‌های گوناگونی دارند. به‌عنوان نمونه، مدل‌های رگرسیون لجستیک از جمله نخستین تکنیک‌های کلاسیک داده‌کاوی به‌شمار می‌روند که در آن دسته مورد نظر طبقه‌بندی می‌شود.[۸] شبکه‌های عصبی مصنوعی مدل‌های جدید محاسباتی هستند که برای انجام بازشناسی الگوها از توجه به ساختار و عملکرد شبکه‌های عصبی زیستی الهام گرفته شده‌اند.[۹] مدل ماشین‌بردار پشتیبان به خانواده‌ای از مدل‌های خطی تعمیم‌یافته تعلق دارند. در این خانواده از مدل‌ها براساس مقدار مربوط به ترکیب خطی خصیصه‌ها تصمیماتی در مورد دسته‌بندی و رگرسیون گرفته می‌شود. این روش افزون‌بر دارا بودن یک شالوده ریاضیاتی منسجم در تئوری یادگیری آماری، عملکرد بسیار خوبی در کاربردهای عملی از خود به نمایش گذاشته است. ماشین‌بردارهای پشتیبان، مشابه با شبکه‌های عصبی، قادر هستند تا برای هر تابع چندمتغیره، تقریب‌هایی را با درجه دقت دلخواه به‌دست دهند، بنابراین در مدل‌سازی سیستم‌ها و فرآیندهای غیرخطی و بسیار پیچیده کاربرد دارند.[۱۰]

پژوهش کنونی با هدف مقایسه کارایی روش‌های داده‌کاوی در پیش‌بینی ابتلا به بیماری دیابت نوع ۲ انجام شد.

روش بررسی

این مطالعه به‌صورت مقطعی-تحلیلی بر روی ۱۱۴۰ فرد انجام شد. در این مطالعه از داده‌های موجود در پرونده‌ی افراد مراجعه‌کننده به مراکز بهداشتی-درمانی منطقه ثامن مشهد شامل کارکنان و خدام فعال آستان قدس رضوی، کسبه بازار و کارمندان منطقه ثامن در طرح غربالگری دیابت در سال ۱۳۹۵ که شامل بررسی میزان قندخون، چشم‌ها، آسیب کلیوی و سلامت پاها بود، استفاده شده است. با توجه به اینکه در مجموعه داده‌ی مورد استفاده از ۷۰۰۰ نفر تعداد ۵۴۰ نفر مبتلا به بیماری دیابت می‌باشند و از آنجایی‌که حفظ توازن در دسته‌ها در بررسی روش‌های داده‌کاوی به‌ویژه روش ماشین‌بردار پشتیبان از اهمیت خاصی برخوردار است، بنابراین از میان افراد سالم موجود در مجموعه داده تعداد ۶۰۰ نفر به‌صورت نمونه‌گیری متناسب با حجم به‌روش تصادفی ساده انتخاب گردیدند و حجم کل نمونه برای استفاده در این پژوهش ۱۱۴۰ نفر در نظر گرفته شده است.

در این طرح برای تمامی افراد بالای ۳۰ سال اندازه‌گیری‌ها شامل قد، وزن، فشارخون، دور کمر، دور باسن و نمایه توده بدنی انجام گرفته و سپس این داده‌ها در دفاتر ویژه طرح غربالگری دیابت ثبت و برای تحلیل و بررسی در رایانه وارد گردیده است. افرادی که پیشتر سابقه‌ی دیابت نوع ۲ داشته‌اند و اکنون به‌سبب استفاده از دارو و یا رعایت سایر نکات از جمله رعایت رژیم غذایی مقدار قندخون طبیعی دارند از مطالعه حذف گردیدند. ابتدا با استفاده از رگرسیون لجستیک، مدل بیزین ساده و ماشین‌بردار پشتیبان مدل‌های مختلف پیش‌بینی ابتلا به دیابت نوع ۲ تعیین شد و سپس مقادیر حساسیت، ویژگی، ارزش اخباری مثبت و منفی، سطح زیرمنحنی راک، صحت برای هر یک از موارد به‌دست آورده و در جهت به‌دست‌آوردن بهترین مدل به مقایسه مدل‌های مختلف با استفاده از شاخص‌های گفته‌شده پرداخته شد. در انتها نیز مهمترین متغیرها از نظر

هر یک از مدل‌ها نشان داده شد. برای ارزیابی دقیق‌تر کارایی هر یک از مدل‌ها، مجموعه داده به زیرمجموعه‌های آموزش و آزمایش تقسیم گردید. سپس مدل پیش‌بینی با مجموعه داده‌ی آموزش ایجاد و دقت آن با مجموعه داده‌ی آزمایش ارزیابی گردید. اینکه چه بخشی از مجموعه داده اولیه برای آموزش و چه کسری از آن برای آزمایش به‌کار می‌رود بستگی کامل به‌روش ارزیابی مورد استفاده دارد. در این مطالعه از دو نسبت، دوسوم برای آموزش و یک‌سوم برای آزمایش استفاده گردید. تمامی تجزیه و تحلیل‌ها با استفاده از R statistical software, version 3.4.3 (Vienna, Austria, https://www.r-project.org) و بسته‌های نرم‌افزاری، pROC، caret، kernlab، e1071 و rminer انجام شده است.

یافته‌ها

از ۱۱۴۰ فرد مورد مطالعه تعداد ۵۴۰ نفر (۴۷/۴٪) مبتلا به دیابت و ۶۰۰ نفر (۵۲/۶٪) نیز از افراد بدون بیماری بودند. سایر داده‌های جمعیت‌شناختی در جدول ۱ نشان داده شده است. نتایج سه مدل براساس شاخص‌های مقایسه‌ای حساسیت، ویژگی، ارزش اخباری مثبت و منفی، صحت و سطح زیرمنحنی راک نشان داده شده است (نمودار ۱ و جدول ۲). براساس نتایج، مدل ماشین‌بردار پشتیبان دارای حساسیت ۱ و پس از آن مدل رگرسیون لجستیک با حساسیت ۹۸٪ دارای بیشترین میزان حساسیت نسبت به مدل بیزین ساده با حساسیت ۹۵٪ می‌باشند. هر سه مدل دارای صحت ۰/۸۶ درصد هستند اما در سطح زیرمنحنی راک، ماشین‌بردار پشتیبان با ۸۸٪ دارای سطح کمتری نسبت به دو روش دیگر که دارای سطح زیرمنحنی ۹۰٪ است، می‌باشند.

در جدول ۳ ترتیب اهمیت متغیرهای مورد بررسی بر حسب سه مدل بیزین، لجستیک و ماشین‌بردار پشتیبان نشان داده شده است. در مدل بیزین ساده و رگرسیون لجستیک به‌ترتیب متغیرهای شاخص توده بدنی و سن دارای بیشترین میزان اهمیت و متغیرهای نسبت دور کمر به باسن و جنسیت دارای کمترین اهمیت بودند، درحالی‌که در مدل ماشین‌بردار پشتیبان متغیرهای شاخص توده بدنی و فشارخون دارای بیشترین اهمیت و متغیرهای جنسیت و سن دارای کمترین میزان اهمیت بودند.

جدول ۱: داده‌های جمعیت‌شناختی افراد تحت مطالعه

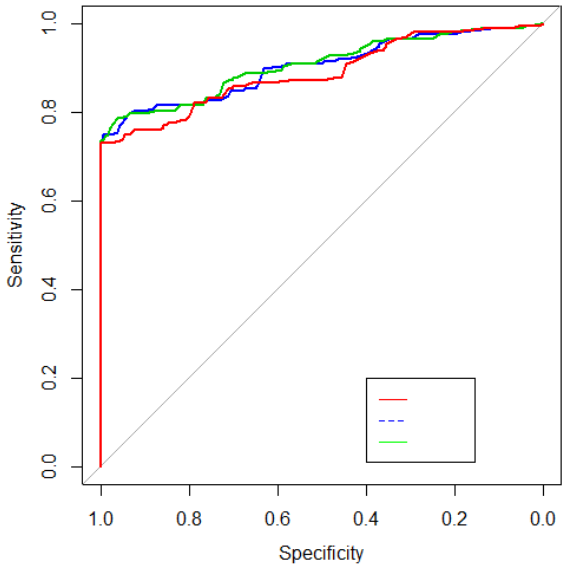
		تعداد	درصد
جنس	مرد	۸۱۳	۷۱/۳
	زن	۳۲۷	۲۸/۷
سن (سال)	۳۰-۴۰	۳۹۷	۳۴/۸
	۴۰-۵۰	۳۵۶	۳۱/۲
	بالای ۵۰	۳۸۷	۳۳/۹

جدول ۲: مقایسه سه مدل بر حسب شاخص‌های مقایسه‌ای مختلف

مدل	حساسیت	ویژگی	ارزش اخباری مثبت	ارزش اخباری منفی	صحت	سطح زیرمنحنی راک
بیزین ساده	۰/۹۵	۰/۷۷	۰/۷۹	۰/۹۴	۰/۸۶	۰/۹۰
رگرسیون لجستیک	۰/۹۸	۰/۷۴	۰/۷۸	۰/۹۸	۰/۸۶	۰/۹۰
ماشین‌بردار پشتیبان	۱	۰/۷۲	۰/۷۷	۱	۰/۸۶	۰/۸۸

جدول ۳: ترتیب اهمیت متغیرهای بر حسب مدل‌های مورد بررسی

مدل	سن	شاخص توده بدنی	فشارخون	جنسیت	نسبت دور کمر به قد	نسبت دور کمر به باسن
بیزین ساده	۲	۱	۴	۶	۳	۵
رگرسیون لجستیک	۲	۱	۴	۶	۳	۵
ماشین‌بردار پشتیبان	۶	۱	۲	۵	۳	۴



نمودار ۱: مقایسه منحنی‌های راک مدل رگرسیون لجستیک، بیزین ساده و ماشین‌بردار پشتیبان

بحث

براساس نتایج این مطالعه هر سه مدل دارای میزان صحت یکسان (۸٦٪) بودند ولی مدل ماشین‌بردار پشتیبان با میزان حساسیت ۱۰۰درصدی نشان داد که عملکرد بهتری در تشخیص افراد بیمار دارد. اما در بررسی سطح زیرمنحنی راک، مدل‌های بیزین ساده و رگرسیون لجستیک عملکرد بهتری نسبت به مدل ماشین‌بردار پشتیبان داشتند. در مطالعه Tapak و همکاران در مقایسه چندین روش داده‌کاوی در طبقه‌بندی افراد مبتلا به دیابت و غیردیابتی، مدل ماشین‌بردار پشتیبان به‌عنوان بهترین مدل ارایه شد.[۱۱] در برخی مطالعات دیگر نیز برتری ماشین‌بردار پشتیبان به نسبت مدل‌های دیگر بیان شده است.[۱۲-۱۵] اما در مطالعه Yu و همکاران در مقایسه دو روش ماشین‌بردار پشتیبان و رگرسیون لجستیک در طبقه‌بندی افراد مبتلا به دیابت و غیردیابتی هر دو روش عمکردی شبیه همدیگر داشتند،[۱٦] و در مطالعه Choi و همکاران دو روش ماشین‌بردار پشتیبان و شبکه عصبی عملکردی به‌طور کامل مشابه داشتند.[۱۲] در مطالعه متاآنالیزی که Dreiseitl و همکاران انجام دادند، در ۵۰٪ موارد روش‌های رگرسیونی و روش‌های جدید داده‌کاوی عملکردی شبیه هم دارند.[۱۷]

در مطالعات مشابهی که در ایران انجام شده است، Esmaily و همکاران به مقایسه سه روش شبکه عصبی مصنوعی، ماشین‌بردار پشتیبان و رگرسیون لجستیک چندگانه در شناسایی عوامل خطر بیماری دیابت نوع ۲ پرداختند. در این مطالعه شبکه عصبی مصنوعی با صحت ۷٦٪، حساسیت ٦۳٪، ویژگی ۸۱٪، به‌عنوان مدل بهتر انتخاب گردید. در این مطالعه صحت، حساسیت و ویژگی برای مدل ماشین‌بردار پشتیبان به‌ترتیب برابر ۷٦٪، ٦٤٪ و ۷۹٪ به‌دست آمد.[۱۸] در مطالعه کنونی صحت در روش ماشین‌بردار پشتیبان برابر ۸٦٪، و حساسیت و ویژگی به‌ترتیب برابر ۱۰۰٪ و ۷۲٪ به‌دست آمد.

در مطالعه دیگر که توسط Esmaily و همکاران دو روش درخت تصمیم و جنگل تصادفی در تعیین عوامل خطر دیابت نوع ۲ به‌کار گرفته شد، این دو روش با استفاده از مقادیر دقت، حساسیت و ویژگی و سطح زیرمنحنی راک مورد مقایسه قرار گرفتند. براساس نتایج به‌دست‌آمده، مدل جنگل تصادفی مناسب‌تر شناخته شد. مقادیر به‌دست‌آمده برای صحت مدل، حساسیت و ویژگی به‌ترتیب برابر ٦٤٪، ٦٤٪ و ٦٦٪ بود.[۱۹] در پژوهش کنونی سطح زیرمنحنی راک در روش ماشین‌بردار پشتیبان برابر ۸۸٪، در هر دو روش لجستیک و بیزین ساده ۹۰٪ به‌دست آمد. اما در مطالعه Esmaily و همکاران که به مقایسه سه روش پرداخته بودند، سطح زیرمنحنی راک در مدل ماشین‌بردار پشتیبان برابر ۷۳٪ محاسبه گردیده بود.[۱۸]

میزان موفقیت این مدل‌ها براساس شاخص‌های مختلف و بنا به اهداف مطالعه می‌توانند متفاوت باشند اما نکته قابل‌توجه این است که میزان موفقیت این الگوریتم‌ها، به فاکتورهای مختلفی بستگی دارد و نمی‌توان یک روش را به‌طور قطعی به‌عنوان بهترین روش برگزید. عواملی چون تعداد متغیرها، بزرگ‌تر بودن پایگاه داده، کم‌بودن تعداد داده‌های مفقوده، دسترسی به داده‌های مناسب و درست و همچنین در برخی روش‌ها مانند ماشین‌بردار پشتیبان انتخاب تابع کرنل درست، شانس موفقیت در داده‌کاوی را افزایش می‌دهند و نتایج الگوریتم‌ها را به موفقیت نزدیک‌تر می‌کنند. بنابراین تفاوت جزیی در نتایج مطالعات ممکن است از همین موارد ناشی شده باشد.

بحث دیگر در روش‌های داده‌کاوی پس از انتخاب مدل، انتخاب متغیرهاست. در مطالعات گوناگون شاخص توده بدنی و فشارخون به‌عنوان مهمترین متغیرهای پیش‌بین بیماری دیابت نوع ۲ ارایه گردیده‌اند.[۲۰-۲۳] در مطالعه Esmaily و همکاران که به مقایسه روش درخت تصمیم و جنگل

تصادفی پرداخته شده بود، براساس مدل جنگل تصادفی به‌ترتیب متغیرهای تری‌گلیسرید، hs-CRP، فشارخون سیستولیک و سابقه خانوادگی دیابت به‌عنوان مهمترین عوامل خطر شناخته شدند و در مدل درخت تصمیم به‌ترتیب متغیرهای سابقه خانوادگی دیابت، تری‌گلیسرید، سن و hs-CRP مهمترین عوامل خطر بودند.[۱۹]

در این مطالعه نیز در دو مدل رگرسیون لجستیک و شبکه بیزین ساده، متغیرهای شاخص توده بدنی و سن و در مدل ماشین‌بردار پشتیبان، شاخص توده بدنی و فشارخون به‌عنوان مهمترین متغیرهای پیش‌بین ارایه شدند که با نتایج به‌دست‌آمده از سایر مطالعات مشابه بود. بنابراین با تغییر سبک زندگی، تغذیه و افزایش فعالیت فیزیکی می‌توان این عوامل را کنترل نمود و از ابتلا به این بیماری جلوگیری کرد، چراکه پیشگیری از یک بیماری از لحاظ اجتماعی و اقتصادی بسیار بااهمیت می‌باشد.

با توجه به اینکه هر سه مدل با استفاده از داده‌های اولیه و غیرآزمایشگاهی دارای عملکرد مشابه و خوبی بودند و تنها دارای تفاوت‌های جزیی در برخی شاخص‌ها بودند به‌نظر می‌رسد که از این مدل‌ها می‌توان در سیستم بهداشتی و درمانی به‌عنوان یک ابزار کمکی ارزان و با دقت بالا استفاده کرد. گرچه این مطالعه نیز دارای محدودیت‌هایی است که لزوم در نظر گرفتن آن‌ها می‌تواند دید بهتری برای مطالعه کنونی و افق جامع‌تری را فراروی پژوهشگران آتی بنمایاند. از محدودیت‌های موجود می‌توان به عدم استفاده از داده‌های آزمایشگاهی و همچنین استفاده از یک جمعیت محدود در منطقه‌ای خاص اشاره کرد که پیشنهاد می‌گردد مطالعات در ابعاد گسترده‌تر و با استفاده از متغیرهای مختلف و یا استفاده از مدل‌های دیگر داده‌کاوی در کنار مدل‌های ارایه‌شده استفاده گردد.

سپاسگزاری: این مقاله حاصل بخشی از طرح تحقیقاتی تحت عنوان "مقایسه کارایی روش‌های داده‌کاوی در پیش‌بینی ابتلا به بیماری دیابت نوع ۲ در شهر مشهد" مصوب دانشگاه علوم پزشکی مشهد در سال ۱۳۹۵ به کد ۹۳۱۳۲۸ می‌باشد که با حمایت دانشگاه علوم پزشکی مشهد اجرا شده است.

References

1. Bakhtiari MR. A review on the latest criteria for laboratory diagnosis of diabetes mellitus, pre-diabetes and gestational diabetes. *J Lab Diagn* 2014;6(24):4-8.
2. Rafeh R, Arbabi M. Data mining techniques to diagnose diabetes using blood lipids. *J Ilam Univ Med Sci* 2015;23(4):239-47.
3. Janahmadi Z, Nekooeian A, Mozafari M. Hydroalcoholic extract of Allium eriophyllum leaves attenuates cardiac impairment in rats with simultaneous type 2 diabetes and renal hypertension. *Res Pharm Sci* 2015;10(2):125-33.
4. World Health Organization (WHO). Diabetes country profiles 2016. [Internet] Geneva: WHO; 2016 [cited 2019 Aug 05]. Available from: https://www.who.int/diabetes/country-profiles/irn_en.pdf?ua=1
5. Mahmoodi MS, Mahmoodi SA, Haghighi F, Mahmoodi SM. Determining the stage of breast cancer by data mining algorithms. *Iran J Breast Dis* 2014;7(2):36-44.
6. Mahmoudi I, Askari Moghadam R, Moazzam MH, Sadeghian S. Prediction model for coronary artery disease using neural networks and feature selection based on classification and regression tree. *J Shahrekord Univ Med Sci* 2013;15(5):47-56.
7. Liao S-H, Chu P-H, Hsiao P-Y. Data mining techniques and applications: A decade review from 2000 to 2011. *Expert Syst Appl* 2012;39(12):11303-11.
8. Komi M, Li J, Zhai Y, Zhang X, editors. Application of data mining methods in diabetes prediction. In: 2nd International Conference on Image, Vision and Computing (ICIVC), 2017.
9. Sasibhushana Rao G, Vimala Kumari G, Prabhakara Rao B. Image compression using neural network for biomedical applications. In: Bansal J, Das K, Nagar A, Deep K, Ojha A, editors. Soft Computing for Problem Solving. Advances in Intelligent Systems and Computing. Singapore: Springer; 2019. P. 107-19.
10. Shahrabi J, Zolghadr Shojaee A. Advanced Data Mining: Concepts and Algorithms. Tehran, Iran: Jahad-e-Daneshgahi Publishing Inc.; 2009.
11. Tapak L, Mahjub H, Hamidi O, Poorolajal J. Real-data comparison of data mining methods in prediction of diabetes in iran. *Healthc Inform Res* 2013;19(3):177-85.
12. Choi SB, Kim WJ, Yoo TK, Park JS, Chung JW, Lee Y-h, et al. Screening for prediabetes using machine learning models. *Comput Math Methods Med* 2014;ID:618976.
13. Li CP, Zhi XY, Ma J, Cui Z, Zhu ZL, Zhang C, et al. Performance comparison between Logistic regression, decision trees, and multilayer perceptron in predicting peripheral neuropathy in type 2 diabetes mellitus. *Chin Med J (Engl)* 2012;125(5):851-7.
14. Meng XH, Huang YX, Rao DP, Zhang Q, Liu Q. Comparison of three data mining models for predicting diabetes or prediabetes by risk factors. *Kaohsiung J Med Sci* 2013;29(2):93-9.
15. Priya R, Aruna P. SVM and neural network based diagnosis of diabetic retinopathy. *Int J Comput Appl* 2012;41(1):6-12.
16. Yu W, Liu T, Valdez R, Gwinn M, Khoury MJ. Application of support vector machine modeling for prediction of common diseases: the case of diabetes and pre-diabetes. *BMC Med Inform Decis Mak* 2010;10:16.
17. Dreiseitl S, Ohno-Machado L. Logistic regression and artificial neural network classification models: a methodology review. *J Biomed Inform* 2002;35(5-6):352-9.
18. Esmaily H, Tayefi M, Ghayour-Mobarhan M, Amirabadizadeh A. Comparing three data mining algorithms for identifying the associated risk factors of type 2 diabetes. *Iran Biomed J* 2018;22(5):303-11.

19. Esmaily H, Tayefi M, Doosti H, Ghayour-Mobarhan M, Nezami H, Amirabadizadeh A. A comparison between decision tree and random forest in determining the risk factors associated with type 2 diabetes. *J Res Health Sci* 2018;18(2):e00412.
20. Bays HE, Chapman RH, Grandy S; SHIELD Investigators' Group. The relationship of body mass index to diabetes mellitus, hypertension and dyslipidaemia: comparison of data from two national surveys. *Int J Clin Pract* 2007;61(5):737-47.
21. Kirkman MS, Briscoe VJ, Clark N, Florez H, Haas LB, Halter JB, et al. Diabetes in older adults. *Diabetes Care* 2012;35(12):2650-64.
22. Narayan KM, Boyle JP, Thompson TJ, Gregg EW, Williamson DF. Effect of BMI on lifetime risk for diabetes in the U.S. *Diabetes Care* 2007;30(6):1562-6.
23. Selvin E, Parrinello CM. Age-related differences in glycaemic control in diabetes. *Diabetologia* 2013;56(12):2549-51.

TUMJ Tehran University Medical Journal, August 2019; Vol. 77, No. 5: 301-307 Original Article

Comparison of the efficiency of data mining methods in predicting type 2 diabetes

Hossein Tireh M.Sc.[1]
Mohammad Taghi Shakeri Ph.D.[2]
Sadegh Rasoulinezhad M.Sc.[1]
Habibollah Esmaily Ph.D.[2]
Razieh Yousefi M.Sc.[1*]

1- Department of Epidemiology and Biostatistics, Faculty of Health, Mashhad University of Medical Sciences, Mashhad, Iran.
2- Social Determinants of Health Research Center, Mashhad University of Medical Sciences, Mashhad, Iran.

[*] Corresponding author: Department of Epidemiology and Biostatistics, Faculty of Health, Daneshgah St., Mashhad, Iran.
Postal Code: 5137673119
Tel: +98 51 31892700
E-mail: yousefir9@mums.ac.ir

Abstract

Received: 02 Mar. 2019 Revised: 09 Mar. 2019 Accepted: 12 Aug. 2019 Available online: 22 Aug. 2019

Background: Diabetes mellitus as a chronic disease is the most common disease caused by metabolic disorders and it is one of the most important health issues all around the world. Nowadays, data mining methods are applied in different fields of sciences due to data mining methods capability. Therefore, in this study, we compared the efficiency of data mining methods in predicting type 2 diabetes.

Methods: In this cross-sectional study, the data of 7,000 participants in the Diabetes Screening Project in Samen, Mashhad City, Iran, were considered in 2016. There were 540 untreated diabetic patients. The Samen Project was included in the routine examinations of diabetes patients like blood glucose, eyes health, nephropathy, and legs health. So, in order to maintain balance, 600 healthy individuals were selected in a proportional volume sampling in this study. Therefore, the total sample size was 1140 people. In this study, people with diabetes aged over 30 years old were enrolled and participants with the previous history of type 2 diabetes, with normal blood glucose due to drug use or other issues at the time of the study, were excluded.

Results: All three models (Logistic regression, simple Bayesian and support vector machine models) had the same test accuracy (86%), however, in terms of area under the receiver operating characteristic (ROC) curve (AUC), logistic regression and simple Bayesian models had better performance (AUC=90% against AUC=88%). In the simple Bayesian model and logistic regression, body mass index (BMI) and age variables were the most important variables, while BMI and blood pressure variables were the most important factors in the support vector machine model.

Conclusion: According to the results, all three models had the same accuracy. In terms of area under the curve (AUC), logistic and simple Bayes models had better performance than the support vector machine model. Totally all three models had almost the same performance. Based on all three models, BMI was the most important variable.

Keywords: data mining, diabetes mellitus, metabolic diseases, sensitivity and specificity.