مجله دیابت و متابولیسم ایران. مهر و آبان ۱۴۰۰؛ دوره ۲۱ (شماره ۴) ۲۷۵-۲۶۴

مقاله پژوهشی

مقایسه‌ی کارایی الگوریتم‌های داده‌کاوی در پیش‌بینی تشخیص بیماری دیابت

فاطمه دکامینی[1*]، محمد احسانی فر [2]

چکیده

مقدمه: دیابت یکی از مشکلات اساسی سلامت در ایران بوده و حدود ٤/٦ میلیون نفر از بزرگسالان به این بیماری مبتلا هستند. ضعف در تشخیص این بیماری سبب شده نیمی از این تعداد از بیماری خود اطلاعی نداشته باشند. در سالیان اخیر همزمان با به‌کارگیری رایانه در تحلیل و ذخیره‌سازی اطلاعات، حجم و پیچیدگی داده‌ها به‌صورت چشمگیری افزایش یافته است.

روش‌ها: در سازمان‌های سلامت داده‌ها نقش اساسی در ارزش سازمان ایفا می‌کنند. از این‌رو داده‌کاوی به یکی از پُرکاربردترین فرآیندها در حوزه‌ی سلامت و تشخیص بیماری‌ها تبدیل شده است. در این پژوهش اطلاعات ۷٦۸ نفر از مراجعین آزمایشگاهی در تهران با حفظ محرمانگی و برای شناسایی متغیرهای تأثیرگذار در ابتلا به بیماری دیابت از نظرات خبرگان استفاده شده است.

یافته‌ها: یافته‌ها حاکی از بررسی ۵ الگوریتم مورد نظر بر روی داده‌های ارائه شده است که با پیاده‌سازی ۵ الگوریتم داده‌کاوی J48، بیز، بگینگ، کوهن و خوشه‌بندی ساده جهت دسته‌بندی داده‌ها، کارایی این الگوریتم‌ها از نظر سرعت و دقت در محاسبات بررسی گردید.

نتیجه‌گیری: مجموعه داده‌ها جهت دسته بندی، بانک داده‌های یک آزمایشگاه است که این مجموعه شامل ۷٦۸ نمونه با ۹ مشخصه است. نهایتاً الگوریتم J48 به‌دلیل سرعت بالا، دقت مورد قبول و عدم وجود حساسیت به داده‌های اولیه، جهت داده‌کاوی داده‌های بیماری دیابت پیشنهاد می‌شود.

واژگان کلیدی: داده‌کاوی، دیابت، کارایی، کشف دانش

۱- گروه مدیریت صنعتی، دانشکده‌ی مدیریت، واحد اراک، دانشگاه آزاد اسلامی، اراک، ایران

۲- گروه مهندسی صنایع، دانشکده‌ی فنی مهندسی، واحد اراک، دانشگاه آزاد اسلامی، اراک، ایران

*__نشانی__: اراک، میدان امام خمینی(ره)، بلوار امام خمینی(ره)، کیلومتر ۳ جاده خمین، شهرک دانشگاهی امیرکبیر، دانشگاه آزاد اسلامی اراک، صندوق پستی: ٥٦٧/٣٨١٣٥، تلفن: ۹-۳٤۱۳۲٤٥۱-۰۸٦ کدپستی: ۹۱۳۱-۱-۳۸۳٦۱، پست الکترونیک: s_dekamin@yahoo.com

تاریخ دریافت: ۱۴۰۰/۰۵/۱۰ **تاریخ درخواست اصلاح: ۱۴۰۰/۰۷/۲۰** **تاریخ پذیرش: ۱۴۰۰/۰۷/۲۸**

مقدمه

دیابت یکی از مهلک‌ترین، ناتوان کننده‌ترین و پُرهزینه‌ترین بیماری‌های مشاهده شده‌ی حال حاضر است و میزان آن به شکل هشدار دهنده‌ای در حال افزایش است [۱].

دیابت بیماری متابولیکی است که در آن افراد از کمبود یا کاهش قابلیت استفاده از انسولین بدنشان رنج می‌برند. عللی چون چاقی، استرس، بالا رفتن سطح کلسترول و چربی، تغذیه‌ی نامناسب و سبک زندگی بدون تحرک، عواملی هستند که در شیوع بیماری دیابت نقش دارند [۲]. عدم تشخیص به موقع و یا ضعف در تشخیص این بیماری ازجمله مشکلات عمده‌ی دیگری است که در رابطه با این بیماری وجود دارد [۳].

تاکنون درمان مشخصی برای بیماری دیابت کشف نشده است [٤]. این بیماری چهارمین علت مرگ و میر در بیشتر کشورهای توسعه یافته است [۱]. بیش از نیمی از مبتلایان به دیابت از بیماری خود بی‌خبر هستند [۳]، بنابراین پیاده‌سازی روشی که بتواند تشخیص صحیح ابتلا یا عدم ابتلا به بیماری دیابت را آسان نماید گامی مهم در جهت پیشگیری و کنترل، به‌خصوص در مراحل ابتدایی بیماری به حساب می‌آید [۵].

در شرایطی که استفاده از فناوری اطلاعات برای به‌کارگیری دانش بالقوه ضروری است، داده‌کاوی[1] پاسخی مناسب برای استخراج این دارایی به حساب می‌آید. داده‌کاوی راهی است برای تحلیل اتوماتیک داده‌ها و شناسایی الگوهای پنهان که انجام این امر به‌صورت دستی ممکن نیست [٦]. داده‌کاوی می‌تواند در پیش‌بینی و تشخیص سریع و کم‌هزینه‌ی بیماری‌ها به‌طور مؤثری استفاده شود. یک فایده‌ی آن این است که انبار داده برای برآورده کردن نیازهای تجزیه و تحلیل‌گر طراحی شده است، و مجموعه‌ای از داده‌های مناسب و متناسب زمان را در خود دارد و برای Query با کارآیی زیاد بهینه شده است [۷].

ساختن یک محل فیزیکی جداگانه برای ذخیره‌سازی اطلاعات مزیت‌های زیادی دارد، اول اینکه داده‌های انباره همیشه قابل دستیابی است، به‌طوری‌که حتی منابع اصلی هم همیشه قابل دسترسی نیستند، همچنین یک محل ذخیره‌ی جداگانه، تجزیه و تحلیل پردازش‌ها را آسان می‌کند و از طرفی اطلاعات خلاصه شده و قدیمی در انباره موجود است که در منابع اصلی موجود نیست.

ظرفیت داده‌کاوی در تحلیل داده‌ها و استخراج مدل که منجر به تولید دانش می شود، قدم اصلی فرآیند کشف دانش[2] از پایگاه داده است. منبع اصلی درآمد یک گروه یا سازمان نشان دهنده‌ی صنعت مربوطه است. فرآیند داده‌کاوی شامل تعریف مسئله، آماده‌سازی داده‌ها، بررسی و اعتبار سنجی داده‌ها، تهیه‌ی مدل‌ها، بررسی و ارزیابی مدل‌ها، اجرا و روزآمد سازی مدل‌ها است [۸].

ایده‌ی اصلی داده‌کاوی بر این امر استوار است که داده‌های قدیمی حاوی اطلاعات مفید هستند که در آینده مورد استفاده قرار خواهند گرفت [۹].

روش‌های مختلف داده‌کاوی، به‌طور گسترده‌ای در تحقیقات پزشکی به‌کار رفته است که این امر می‌تواند در تشخیص بیماری و کاهش اشتباهات به پزشکان کمک شایانی نماید [۱۱ ،۱۰]. روش‌های مختلفی از جمله استفاده از روش‌های تکاملی [۱۲]، مدل‌سازی فازی [۱۳]، روش‌های مبتنی بر بیزین [١٤]، تشخیص الگو در استخراج ویژگی [۱۵]، تشخیص دیابت به کمک بردار ماشین پشتیبان [١٦]، استفاده از درخت تصمیم در تشخیص دیابت نوع دو [۱۷]، روش خوشه‌بندی [۱۹ ،۱۸]. روش‌های بینایی ماشین [۲۰] و تشخیص بیماری دیابت با استفاده از شبکه‌های عصبی و فازی عصبی [۲۲ ،۲۱ ،٦] به‌کار گرفته شده است.

در جوامع مختلف یافتن الگوهای مفید در داده‌ها با عناوین متعدد نظیر استخراج دانش، کشف اطلاعات، برداشت اطلاعات و پردازش الگوهای داده‌ها مطرح می‌شود [۲۳].

یکی از با ارزش‌ترین دارایی‌های سازمان‌های سلامت در عصر اطلاعات، داده‌ها هستند و جمع‌آوری، ذخیره و تحلیل این داده‌ها می‌تواند یکی از عوامل موفقیت این سازمان‌ها محسوب شود [٢٤]. به‌وسیله داده‌کاوی امکان پی بردن به روابط، روندها

[1] Data mining

[2] Knowledge discovery

و الگوهای مخفی بین داده‌ها و در نهایت دستیابی به دانش نوین میسر خواهد شد [۹]. یکی از بیشترین کاربردهای داده‌کاوی در پزشکی، تشخیص بیماری‌ها است. جهت تشخیص بیماری دیابت، متغیرهای فیزیکی و خونی برخی بیماران دیابتی و افراد سالم به الگوریتم‌های رده‌بندی داده‌کاوی داده می شود [۲۵]. این الگوریتم‌ها مدل‌هایی را جهت رده‌بندی بیماران به دو رده‌ی "بیمار دیابتی" و "فرد سالم" ایجاد می‌نمایند. از مُدل‌های ایجاد شده می‌توان به‌منظور رده‌بندی مراجعان جدید و افراد مشکوک به بیماری دیابت استفاده و وضعیت سلامت افراد جدید را پیش‌بینی نمود [۲۶].

پژوهش پیش‌رو، با ترکیب روش های داده کاوی از قبیل سیستم‌های خوشه‌بندی، فضای حالتی جدید از بیماران به‌منظور تعیین میزان دیابتی بودن یا سالم بودن افراد به‌صورت خوشه‌ای بیان می‌کند. سپس با اعمال ورودی‌ها و خروجی‌های سیستم داده‌کاوی، سطح فیزیکی بین ورودی‌های سیستم، میزان گلوکز ناشتا و شاخص جرم بدن خروجی هر الگوریتم به‌دست می‌آید و الگوریتم‌های J48 (رگرسیون)، بیز، بگینگ، کوهن و خوشه‌بندی ساده از لحاظ صحت، دقت و سرعت با یکدیگر مقایسه می‌شوند. در انتها نیز الگوریتم کارا با خطای کمتر، سرعت بیشتر و صحت نتایج بالاتر در پیش‌بینی بیماری دیابت معرفی خواهد شد.

پیشینه‌ی پژوهش

در مطالعه‌ی Si و همکاران که در سال ۲۰۱۸ از مرکز دیابت استان انگلستان جمع‌آوری شده است تعداد ۲۵۰ رکورد با ۱۲ ویژگی پس از پالایش داده‌ها ثبت شده است. این ویژگی‌ها الزاماً مشابه نمونه‌های پایگاه داده PID نیستند، با این وجود در مطالعه‌ی مورد بررسی محقق ضمن پیش‌بینی وضعیت دیابت نشان داده است که چه پارامترهایی در تشخیص بیماری از اهمیت بالاتری برخورد هستند [۲۷]. در مطالعه‌ی دیگری، علاوه بر ویژگی‌های ثبت شده در PID، پارامترهای دیگری نیز مانند وزن و دور کمر به‌عنوان ویژگی‌های مهم تشخیص دیابت ذکر شده‌اند [۲۸].

Sabbagh Gol و همکاران نیز با استفاده از الگوریتم درخت تصمیم C4.5 به تشخیص بیماری دیابت با استفاده از چربی خون پرداختند [۲۹].

Burfei (۲۰۱۶)، مطالعه‌ای بر روی ۱۳۴۲۳ نفر از شرکت‌کنندگان بالای ۲۵ سال که هیچ‌کدام دیابت کنترل شده‌ای نداشتند، انجام داد. با استفاده از مدل شبکه‌ی عصبی مصنوعی سه لایه، مدلی ارائه داد که دقت سطح زیر منحنی ۷۲/۷٪ و صحت پیش‌بینی آموزش ۹۲٪ و صحت پیش بینی آزمون ۹۶/۶٪ بود. نتایج پژوهش حاکی از این بود که با توجه به عدم نیاز مدل شبکه‌ی عصبی مصنوعی به پیش فرض‌های معمول، روش‌های کلاسیک آماری و درصد صحت پیش‌بینی شبکه‌های عصبی بیشتر است [۳۰].

همچنین در سال ۲۰۱۰، عده‌ای از محققان تایلندی با استفاده از درخت تصمیم توانستند با دقت بیش از ۹۰ درصد سندرم متابولیک را در افراد تشخیص دهند. آنها از داده‌های مربوط به ۵۶۳۸ نفر استفاده کردند [۳۱].

از سوی دیگر، در یک مطالعه که در سال ۱۳۹۳ با عنوان پیش‌بینی به ابتلای بیماری دیابت با استفاده از شبکه‌ی عصبی مصنوعی صورت گرفت [۳۰]، تعداد ویژگی‌های ثبت شده‌ی هر بیمار ۸ مورد بود که با ویژگی‌های پایگاه داده PID متفاوت است.

در اواخر دهه‌ی ۸۰ میلادی، با پی بردن به ارزش اطلاعات نهفته در پایگاه‌های داده، تلاش جهت استفاده از این داده‌ها آغاز گردید. فرآیند داده‌کاوی در اوایل دهه‌ی ۹۰ میلادی با نگرشی متفاوت نسبت به سیستم‌های دستی سنتی مطرح شد. اولین کارگاه‌های کشف دانش از پایگاه داده در سال‌های ۱۹۸۹ و ۱۹۹۱ میلادی توسط پیاتتسکی[1] و سپس در فاصله‌ی سال‌های ۱۹۹۱ تا ۱۹۹۴ میلادی توسط فایاد و پیاتتسکی برگزار شدند. به‌طور رسمی اصطلاح داده‌کاوی برای اولین بار توسط فایاد[2] در اولین کنفرانس بین‌المللی "کشف دانش و داده‌کاوی"

[1] Piatsky

[2] Fayad

در سال ۱۹۹۵ میلادی مطرح و به‌طور جدی وارد مباحث آمار شد [۳۲].

Fang (۲۰۰۸) با استفاده از تکنیک های مختلف داده کاوی بیماران را براساس مبتلا بودن به دیابت خوشه‌بندی کرده است. ویژگی‌هایی که در این مدل‌ها مهم شناخته شدند عبارتند از سن، سابقه‌ی خانوادگی و وزن. دقت مدل ایجاد شده با استفاده از خوشه‌بندی ۸۰ درصد است [۳۳].

Han و همکاران (۲۰۰۸) با به کار بردن الگوریتم درخت تصمیم، وجود دیابت را در پایگاه داده بیماران پیش‌بینی کرده‌اند [۳٤]. همچنین، در مطالعه‌ای دیگر شبکه‌ی عصبی مصنوعی و درخت تصمیم ساخته شده از الگوریتم C4.5 به‌کار بسته شد تا وجود دیابت در افراد بر اساس ویژگی‌هایی مثل سن و فشار خون تشخیص داده شود [۳۵]. به‌علاوه، در پژوهشی دیگر با استفاده از ترکیب الگوریتم‌های C4.5 و EM[1] (حداکثر انتظار) سیستم پردازش داده‌های دیابت نوع دو را ایجاد کرده‌اند [۳٦].

در سال ۲۰۰۱ میلادی نیز در فرانسه، تحقیقی بر روی دیابت جهت مدیریت درمان و کنترل دیابت و عوامل قابل اصلاح خطرات قلبی- عروقی در بیماران دارای دیابت نوع دو تحت مراقبت مشخص اجرا شد. داده‌های استفاده شده برای انجام این تحقیق شامل تاریخچه‌ی دیابت و بیماری های قلبی – عروقی افراد، فشار خون، عوامل خطرات قلبی – عروقی، سطح هموگلوبین A_{1c}، سطح کلسترول و جزئیات داروهای استفاده شده بوده‌اند [۵].

Breault و همکاران (۲۰۰۲) با استفاده از طبقه‌بندی به کمک متد CART[2] وابستگی بین برخی ویژگی ها شامل سن، جنسیت، بیماری‌های قلبی – عروقی، فشار خون و از این قبیل را در تشخیص دیابت معرفی نمودند [۲۵].

در مطالعه‌ای دیگر عوامل مؤثر بر پیش‌بینی و تشخیص بیماری دیابت بررسی شد و با استفاده از روش طبقه‌بندی درختی ترتیب اهمیت درمان های بالینی که در نهایت منجر به کاهش عوارض بیماری دیابت می‌شوند ارائه گردید [۳۷].

در یکی از مطالعات صورت گرفته، با استفاده از روش تجزیه و تحلیل رگر سیون خطی ارتباط بین قند خون و هموگلوبین A_{1c} در دیابت نوع یک بررسی گردیده است که هدف‌گذاری روزانه جهت دستیابی به سطح مشخص قند خون و هموگلوبین A_{1c} مستلزم شناسایی این ارتباط است [۳۸].

روش‌ها

در این پژوهش اطلاعات ۷٦۸ نفر از مراجعین آزمایشگاهی در تهران با حفظ محرمانگی و برای شناسایی متغیرهای تأثیرگذار در ابتلا به بیماری دیابت از نظرات خبرگان استفاده شده است. مجموعه داده‌ها شامل ۹ مشخصه‌ی تعداد دفعات بارداری، غلظت قند پلاسما در یک دو ساعت در تست تغییر گلوکز زبانی، فشارخون در انبساط قلب، ضخامت لایه‌ی پوستی ماهیچه سه سر بازو، دو ساعت سرم انسلین، ضریب جرمی بدن، تابع منشاء دیابت، سن و متغیر دسته هستند.

لازم به ذکر است که روش داده‌کاوی دارای یازده مرحله است [۳۹]:

۱. تبدیل مسأله مورد نظر به یک مساله داده‌کاوی
۲. انتخاب داده‌های مناسب
۳. شناخت داده‌ها
٤. ساخت مجموعه مدل
۵. رفع مشکلات داده‌ها
٦. تبدیل داده‌ها برای استخراج اطلاعات
۷. ساخت مدل‌ها
۸. ارزیابی مدل‌ها
۹. پیاده‌سازی مدل‌ها
۱۰. ارزیابی نتایج
۱۱. شروع دوباره

فرآیند داده‌کاوی از ترکیب چندین رشته برای شناسایی و استخراج دانش کمک می‌گیرد. آمار و ریاضی، یادگیری ماشین، روش‌های بهینه‌سازی، روش‌های تشخیص و شناخت الگو، بانک اطلاعاتی، تجسم‌سازی، شبکه‌های عصبی، بازیابی

[1] Expectation-Maximization

[2] Classification and Regression Trees

اطلاعات، الگوریتم ژنتیک و هوش م صنوعی فنونی ه ستند که داده‌کاوی از آنها بهره می‌برد [۴۰].

آزمایش‌ها با استفاده از مجموعه داده‌های آموختن انجام و این مجموعه داده‌ها به دو بخش تقسیم شده‌اند؛ به‌طوری‌که ۶۰ درصد داده‌ها برای آموختن و ۴۰ درصد برای تست به‌کار رفته است. داده‌ها پس از آماده‌سازی و پیش‌پردازش‌های لازم در الگوریتم‌های J48، بیز، بگینگ، کوهن و خوشه‌بندی ساده اجرا و در نهایت مقایسه‌ی بین این الگوریتم‌ها انجام گرفته است.

یافته‌ها

مراحل اجرایی پژوهش به دو قسمت معرفی داده‌ها و تحلیل داده‌ها تفکیک شده است که در ادامه جزئیات آنها تشریح خواهد شد.

معرفی داده‌ها

بیماری دیابت سطوح، علائم و روش‌های پیشگیری مختلفی دارد. علل ابتلا به دیابت تاکنون ناشناخته‌اند. با این حال تصور می‌شود ترکیبی از عوامل ژنتیکی، سابقه‌ی خانوادگی و عوامل محیطی در ایجاد دیابت دخالت داشته باشند. مجموعه داده‌ی آموختن برای طبقه‌بندی، داده‌های آزمایشگاهی شهر تهران است. مجموعه‌ی داده‌ها ۷۶۸ نمونه است که هر کدام ۹ مشخصه دارند. لازم به ذکر است جهت تحلیل داده‌ها از نرم‌افزار وکا ۳,۷ استفاده شده است. میز کار وکا مجموعه‌ای از الگوریتم‌های روز ماشینی و ابزارهای پیش‌پردازش داده‌ها است. این نرم‌افزار به گونه‌ای طراحی شده که می‌تواند به سرعت، روش‌های موجود را به‌صورت انعطاف‌پذیری روی مجموعه‌های جدید داده، آزمایش نماید.

جدول ۱ جزئیاتی از مشخصه‌های انتخاب شده برای تحلیل داده‌ای دیابت ارائه می‌دهد.

جدول ۱- مشخصه‌های مجموعه داده دیابت

مشخصه	نوع
تعداد دفعات بارداری	پیوسته
غلظت قند پلاسما در یک دو ساعت در تست تغییر گلوکز زبانی	پیوسته
فشار خون در انبساط قلب (mm Hg)	پیوسته
ضخامت لایه پوستی ماهیچه سه سر بازو	پیوسته
۲ ساعت سرم انسولین (mu U/ml)	پیوسته
ضریب جرمی بدن ((Kg/m)^2)	پیوسته
تابع منشاء دیابت	پیوسته
سن (سال)	پیوسته
متغیر دسته بیمار/سالم (صفر یا یک)	ناپیوسته

تحلیل داده‌ها

در این بخش، ۵ الگوریتم مورد نظر بررسی شده، بر روی داده‌های ارائه شده پیاده‌سازی و نتایج زیر حاصل شده است.

لازم به ذکر است صحت دسته‌بندی داده‌ها عبارت است از دقت درصد چندتایی‌هایی که به درستی توسط دسته‌بندی کننده انتخاب شده‌اند که با استفاده از ماتریس اغتشاش ارائه شده در جدول ۲ محاسبه شده است.

جدول ۲- ماتریس اغتشاش

	C1	C2
C1	True Positives	False Negatives
C2	False Positives	True Negatives

درست مثبت (TP) به چندتایی‌های مثبت برمی‌گردد که به‌طور درست توسط دسته‌بندی کننده برچسب‌گذاری شده‌اند. درست منفی (TN) به چندتایی‌های منفی اشاره می‌کند که به درستی توسط دسته‌بندی کننده برچسب خورده‌اند. غلط مثبت (FP) به چندتایی‌های منفی اشاره دارد که به غلط توسط دسته‌بندی کننده برچسب خورده‌اند و غلط منفی (FN) به چندتایی‌های مثبتی اطلاق می‌شود که اشتباه توسط دسته‌بندی کننده دسته‌بندی شده‌اند.

الگوریتم J48

این الگوریتم یک الگوریتم منبع باز است که در ابزار داده‌کاوی نرم افزار وکا وجود دارد. درخت تصمیم یکی از رایج‌ترین تکنیک‌های داده‌کاوی است. این الگوریتم در پژوهش‌های متنوع، از جمله در دسته‌بندی میزان خسارات، به‌عنوان مؤثرترین تکنیک داده‌کاوی ارزیابی شده است [۴۱].

این روش یادگیری برای توابع گسسته و داده‌های خطادار به‌کار می‌رود و به کشف دانش کمک می‌کند [۴۲]. یک درخت تصمیم‌گیری یک مدل طبقه‌بندی مناسب با استفاده از مجموعه‌ای از داده‌ها است [۴۳]. در مقاله‌ای با عنوان "آنالیز تکنیک‌های مختلف داده‌کاوی در تشخیص دیابت قندی" صحت پیش‌بینی در درخت تصمیم حدود ۸۶ درصد و در شبکه‌ی عصبی حدود ۷۴ درصد تخمین زده شد [۴۴].

یک درخت تصمیم از داده‌های ورودی می‌تواند به کمک برنامه‌ی رگرسیون ایجاد شود. درخت تصمیم ایجاد شده با این روش می‌تواند برای دسته‌بندی استفاده شود و به‌عنوان یک دسته‌بندی کننده‌ی آماری شناخته شود [۴۵]. در این روش ابتدا از فیلتراسیون نرمال‌سازی، جهت بهبود وضعیت داده‌ها استفاده شده است و سپس داده‌های به‌دست آمده توسط الگوریتم J48 اجرا شده‌اند. نتایج به‌دست آمده در جدول ۳ شرح داده شده است.

جدول ۳- نتایج الگوریتم J48

عنوان	مقدار	توضیحات
نمونه	۷۶۸	
مشخصه	۹	
تعداد برگ	۲۰	
سایز درخت	۳۹	
زمان صرف شده برای ساخت مدل	۰/۰۱	Second
تعداد نمونه‌هایی که صحیح دسته‌بندی شده‌اند	۶۴۶	۸۴/۱۱۴۶٪
تعداد نمونه‌هایی که صحیح دسته‌بندی نشده‌اند	۱۲۲	۱۵/۸۸۵۴٪
دقت	۰/۸۰۸	
بازخوانی	۰/۸۸۵	
صحت	۰/۸۴۱	Accuracy= (TP+TN)/(TP+TN+FP+FN)

الگوریتم طبقه‌بندی بیز

این الگوریتم یکی از روش‌های بیز اســـت و به داده‌های اولیه بسیار حساس است. در صورت عدم تعریف درست داده‌ها و عدم فیلترینگ مناسب داده‌ها، این الگوریتم دچار مشکل زیادی خواهد شـــد. نتایج به‌دســـت آمده از این الگوریتم در جدول ٤ ارائه شده است.

جدول ۴- نتایج الگوریتم طبقه بندی بیز

عنوان	مقدار	توضیحات
نمونه	٧٦٨	
مشخصه	۹	
زمان صرف شده برای ساخت مدل	۰/۰۷	Second
تعداد نمونه‌هایی که صحیح دسته‌بندی شده‌اند	٦۰۱	۷۸/۲۵۵۲٪
تعداد نمونه‌هایی که صحیح دسته‌بندی نشده‌اند	١٦٧	۲۱/۷٤٤۸٪
دقت	۰/۷۸۳	
بازخوانی	۰/۷۸۳	
صحت	۰/۷٦۳	Accuracy= (TP+TN)/(TP+TN+FP+FN)

الگوریتم دسته‌بندی بگینگ

این الگوریتم برای ترکیب رده‌بندی‌های پیش‌بینی شـــده از چند مدل به‌کار می‌رود. در این روش از چند الگوریتم مستقل از هم اســـتفاده می‌شـــود. نکته قابل توجه در این روش نزدیکی زیاد صـــحت و دقت داده‌های الگوریتم اســـت. نتایج پیاده‌ســـازی الگوریتم بگینگ در جدول ۵ ارائه شده‌اند.

جدول ۵- نتایج الگوریتم دسته‌بندی بگینگ

عنوان	مقدار	توضیحات
نمونه	٧٦٨	
مشخصه	۹	
زمان صرف شده برای ساخت مدل	۰/۳٦	build model / Second
تعداد نمونه‌هایی که صحیح دسته‌بندی شده‌اند	٦٨۹	۸۹/۷۱۳۵٪
تعداد نمونه‌هایی که صحیح دسته‌بندی نشده‌اند	۷۹	۱۰/۲۸٦۵٪
دقت	۰/۸۹۷	
بازخوانی	۰/۸۹۷	
صحت	۰/۸۹۷۱	Accuracy= (TP+TN)/(TP+TN+FP+FN)

الگوریتم کوهن

این الگوریتم براســـاس کلاس‌بندی نزدیک‌ترین همســـایه کار می‌کند و داده‌های نزدیک را در دو کلاس دســـته‌بندی می‌نماید. نتایج به‌دست آمده در جدول ٦ بر حسب بیماران و افراد سالم با دقت ۱۰۰ درصـد اسـت که با توجه به نوع الگوریتم می‌توان گفت اگرچه تشـــخیص درســـتی دارد ولی این عامل نشـــان دهنده‌ی واقعیت و صحت قطعی این الگوریتم در ساختارهای مشابه نیست.

جدول ۶- نتایج الگوریتم کوهن

عنوان	مقدار	توضیحات
نمونه	۷۶۸	
مشخصه	۹	
زمان صرف شده برای ساخت مدل	۰/۱۹	test model / Second
تعداد نمونه‌هایی که صحیح دسته‌بندی شده‌اند	۷۶۸	۱۰۰٪
تعداد نمونه‌هایی که صحیح دسته‌بندی نشده‌اند	۰	۰٪
دقت	۱	
بازخوانی	۱	
صحت	۱	Accuracy= (TP+TN)/(TP+TN+FP+FN)

الگوریتم خوشه‌بندی[1] ساده (K-Means)

K-Means یک الگوریتم ساده است که کاربردهای مختلفی دارد. همچنین یکی از ساده‌ترین الگوریتم‌های یادگیری بدون نظارت است که مسائل خوشه‌بندی را حل می‌کند. این الگوریتم از یک شیوه‌ی تکراری برای دسته‌بندی مجموعه داده‌ها در یک تعداد از پیش مشخص شده‌ی خوشه استفاده می‌کند. اگرچه ثابت شده که این الگوریتم همیشه پایان می‌پذیرد اما لزوماً جواب بهینه را نمی‌یابد. این الگوریتم دارای حساسیت زیادی به مراکز خوشه اولیه است که به‌صورت تصادفی انتخاب می‌شوند. برای کاهش این تأثیر می‌توان الگوریتم را چند بار اجرا کرد. نتایج حاصل از این الگوریتم در جدول ۷ ارائه شده است.

جدول ۷- نتایج الگوریتم خوشه‌بندی ساده

عنوان	مقدار	توضیحات
نمونه	۷۶۸	
مشخصه	۹	
تعداد تکرار	٤	
زمان صرف شده برای ساخت مدل	۰/۱۱	test model / Second
تعداد دسته	۲	

بحث و نتیجه‌گیری

بیماری دیابت درمان قطعی ندارد، بنابراین از یک سو شناسایی عوامل خطر این بیماری و پیشگیری از ابتلا به آن و از سوی دیگر تشخیص زود هنگام که به طرز چشمگیری از عوارض دیابت می‌کاهد، اهمیت بالایی دارد. دنیای پزشکی نیاز به یک روش پیش‌بینی قابل اطمینان برای تشخیص بسیاری از بیماری‌ها از جمله دیابت دارد.

پژوهش‌های مبتنی در داده‌کاوی می‌تواند به پزشکان در تشخیص بیماری‌های مختلف از جمله دیابت کمک کند. هدف نهایی یک سیستم شناسایی الگو دستیابی به بالاترین نرخ طبقه‌بندی ممکن برای مسئله مورد نظر است.

در مطالعه‌ی حاضر با بررسی ۵ الگوریتم داده‌کاوی مورد نظر و پیاده‌سازی آنها روی داده‌ها می‌توان نتیجه گرفت الگوریتم‌های کلاس‌بندی دارای سرعت بیشتر و به مراتب هزینه تولید و تست

[1] Clustering

مدل کمتر نسبت به سایر روش‌ها است. همچنین دقیق‌ترین روش، الگوریتم بگینگ است که دقت آن در حدود ۸۹ درصد است. اگرچه روش‌های خوشه‌بندی دارای دقت بالایی هستند ولی همان‌طور که مطرح شد جواب این روش‌ها برای تمامی تست‌ها همیشگی نیست و به داده‌های اولیه حساسیت زیادی دارند.

نتایج این پژوهش یافته‌های مطالعات قبلی مانند Wang و همکاران را تأیید می‌کند. آنها دو مدل شبکه‌ی مصنوعی و رگرسیون لجستیک چند متغیره را با هم مقایسه کرده‌اند. نتایج آنها نشان داد که مدل شبکه عصبی از لجستیک دارای دقت بالاتری است [٤٦].

Langarizadeh و همکاران با استفاده از شبکه‌ی عصبی به پیش‌بینی تولد نوزاد نارس در مادران باردار شده پرداخته‌اند که نتایج به‌دست آمده از این پژوهش، نشان داد استفاده از شبکه‌ی پرسپترون چندلایه برای پیش‌بینی نتیجه‌ی زایمان از نظر تولد نوزاد ترم یا نوزاد نارس در مادران باردار شده از طریق فناوری‌های کمک باروری می‌تواند در پیشگیری از عوارض تولد نوزاد نارس کمک کننده باشد [٤٧].

Huang و همکاران در مطالعه خود، با استفاده از الگوریتم‌های درخت تصمیم، جنگل تصادفی، ماشین بردار پشتیبان، شبکه‌ی بیزین به پیش‌بینی بیماری دیابت نوع دو پرداخته‌اند و با استفاده از این روش به دقت بالای ٦٠ درصد رسیدند [٤٨].

Balakrishnan و همکاران نیز برای پیشبینی بیماری دیابت، از الگوریتم‌های Bayes Naïve و SVM استفاده کردند. آنها با کار بر روی دیتاست pima و با استفاده از نرم‌افزار وکا، به بررسی دقت الگوریتم‌های داده‌کاوی پرداخته و یک روش انتخاب ویژگی برای پیدا کردن یک زیرمجموعه ویژگی مطلوب پیشنهاد دادند که دقت طبقه‌بندی bayes naïve و svm را بالا برده است. تحلیل‌های خود را در نمودار سطح زیر منحنی مقایسه کرده که در نهایت الگوریتم SVM به دقت بالاتری رسید [٤٩].

پس از انجام مقایسات پیشنهاد می‌شود جهت داده‌کاوی بیماری دیابت به‌دلیل سرعت بسیار بالا و دقت مورد قبول از الگوریتم J48 استفاده شود.

سپاسگزاری

نویسندگان از مدیریت آزمایشگاه‌های تهران و مراجعین آزمایشگاهی که در جمع آوری اطلاعات این پژوهش ما را یاری رساندند، کمال تشکر و امتنان را دارند.

مآخذ

1. Nazarzadeh M, Bidel Z, Sanjari Moghaddam A. Meta analysis of diabetes mellitus and risk of hip fractures small study effect. *Osteoporos Int* 2015; 26:123-9.
2. Asadollahi K, Delpisheh A, Asadollahi P, Abangah G. Hyperglycaemia and its related risk factors in Ilam province west of Iran a population based study. *J Diabetes Metab Disord* 2015; 14:81.
3. Janahmadi Z, Nekooeian AA, Mozafari M. Hydroalcoholic extract of Allium eriophyllum leaves attenuates cardiac impairment in rats with simultaneous type 2 diabetes and renal hypertension. *Res Pharm Science* 2015; 10:125-33.
4. Mahdavi O, Hashemi S, Boostani N & Zokaee H. A new method to evaluate fasting plasma glucose by salivary glucose measurement. *Iranian Journal of Diabetes and Obesity* 2012; 4(3):127-133.
5. Charpentier G, Gene`s N, Vaur L, Amar, J, Clerson, P, Cambou JP & Gue´ret P. Control of diabetes and cardiovascular risk factors in patients with type 2 diabetes: a nationwide French survey. *Diabetes & Metabolism* 2003 29 (2):152–158.
6. Elsappagh S, Elmogy M, Riad AM. A fuzzy ontology oriented case based reasoning framework for semantic diabetes diagnosis. *Artif Intell Med* 2015;14:92-5.
7. Amirthalingam G, Shaheen R, Kousar M, Bilfaqih SM. Integrated Data Mining and Knowledge Discovery Techniques in ERP. International. *Journal of Advanced Research in Computer Science & Technology* 2014; 2(4):210-214.
8. Christy A, Joy A, Umamakeswari L, Priyatharsini, Neyaa A. RFMr ranking an effective approach to costumer segmentation. *J King Saudi University Comput Inf Sci.* 2018.
9. Burbidge R & Buxton B. An Introduction to Support Vector Machines for Data Mining. In: Proc. 12th Conference Young Operational Research (YOR12), Nottingham, UK, 27-29 March: 2002; 3-15.

10. Habibi M, Ahmadifard A. Feature Selection Using Taboo Search, Genetic Algorithm and KNN for Diagnosis of Diabetes. *12th Iranian Conference on Intelligent Systems*, 2013; [In Persian].
11. Dehghan P, Mogharabi M, Zabbah I, Layeghi K, Maroosi A. Modeling Breast Cancer Using Data Mining Methods. *Journal of Health and Biomedical Informatics* 2018; 4(4):266-278.
12. Saiti F, Naini A, Shoorehdeli A, Teshnehlab MA. Thyroid disease diagnosis based on genetic algorithms using PNN and SVM. In Bioinformatics and Biomedical Engineering; ICBBE 2009. *3rd International Conference on* 2009; (pp: 1-4). IEEE.
13. Petrich W, Dolenko B, Früh J, Ganz M, Greger H, Jacob S, Keller F, Nikulin AE, Otto M, Quarder O, Somorjai RL. Disease pattern recognition in infrared spectra of human sera with diabetes mellitus as an example. *Applied optics* 2000; 39(19):3372-9.
14. Ling J, Cheng P, Ge L, Zhang DH, Shi AC, Tian JH, Chen YJ, Li XX, Zhang JY, Yang KH. The efficacy and safety of dipeptidyl peptidase-4 inhibitors for type 2 diabetes: a Bayesian network meta-analysis of 58 randomized controlled trials. *Acta diabetologica* 2018; 21:1-24.
15. Rezaei M, Zandkarimi E, Hashemian A. Comparison of Artificial Neural Network, Logistic Regression and Discriminant Analysis Efficiency Determining Risk Factors of Type 2 Diabetes.*World Applied sceinces Journal* 2013; 23(11):1522-9.
16. Santhanam T, Padmavathi MS. Application of K-means and genetic algorithms for dimension reduction by integrating SVM for diabetes diagnosis. Procedia Computer Science. 2015; 1;47:76-83.
17. Esmaily H, Tayefi M, Doosti H, Nezami H, Amirabadizadeh A. A Comparison between Decision Tree and Random Forest in Determining the Risk Factors Associated with Type 2 Diabetes. *Journal of research in health sciences* 2018; 24:18(2).
18. Firuzi Jahantigh F, Ameri H. The investigation of TB patients features with K-Means clustering. *Journal of Health and Biomedical Informatics* 2015; 2(3):149-159.
19. Ghasemzadeh F, Arab-kheradmand A, Daklan S, Shabaninezhad A, Garajei A, Etminani K. Determination of the Most Important Factors Affecting Non-Melanoma Skin Cancer Using Data Mining Algorithms. Journal of Health and Biomedical Informatics. (2017); 4 (1) :39-47.
20. Zabbah I, Hassaanzadeh M, kohjani Z. The Effect of Continuous Parameters on the Diagnosis of Coronary Artery Disease Using Artificial Neural Networks. *Journal of Torbat Heydariyeh University of Medical Sciences* 2017; 4(4):29-39.
21. Mirsharif M, Rouhani S. Data Mining Approach based on Neural Network and Decision Tree Methods for the Early Diagnosis of Risk of Gestational Diabetes Mellitus. *Journal of Health and Biomedical Informatics* 2017; 4(1):59-68.
22. Ahamad MG, Ahmed MF, Uddin MY. Clustering as Data Mining Technique in Risk Factors Analysis of Diabetes, Hypertension and Obesity. *European Journal of Engineering Research and Science* 2018; 27; 1(6):88-93.
23. Craig ME, Jones TW, Silink M & Ping, YJ. Diabetes care, glycemic control, and complications in children with type 1 diabetes from Asia and the Western Pacific Region. *Journal of Diabetes and its Complications* 2007; 21 (5):280– 287.
24. Latour KM & Eichenwald S. *Health Information Management: Concepts, Principles, and Practice.* Chicago: AHIMA 2002; 478-480.
25. Breault JL, Goodall CR & Fose PJ. Data mining a diabetic data warehouse. *Artificial Intelligence in Medicine 2002*; 26 (1-2): 37–54.
26. Yildirim EG, Karahoca A. & Uçar T. Dosage planning for diabetes patients using data mining methods. *Procedia Computer Science* 2011; 3:1374-1380.
27. Si J, Zhang Y, Hu S, Sun L, Li S, Yang H, et al. Comparison of LVQ and BP Neural Network in the Diagnosis of Diabetes and Retinopathy. *International Conference of Pioneering Computer Scientists, Engineers and Educators* 2018; Springer.
28. Jia`ng M, Jiang L, Jiang D, Xiong J, Shen J, Ahmed SH, Luo J and Song H. Dynamic Measurement Errors Prediction for Sensors Based on Firefly Algorithm Optimize Support Vector Machine. *Sustainable Cities and Society* 2017; 35: 250-256.
29. Sabbagh Gol H. A Detection of Type2 Diabetes using C4.5 Decision Tree. *Journal of Health and Biomedical Informatics* 2018; 5 (2):293-303.
30. Burfei F, Salehi M, Najafi I. Anticipating Diabetes Using an Artificial Neural Network. *Razi Medical Journal* 2016; 22(135):29-37.
31. Worachartcheewan A, Shoombuatong W, Pidetcha P, Nopnithipat W, Prachayasittikul V, Nantasenamat C. Predicting metabolic syndrome using the random forest method. *Scientific World J* 2015; 581501.
32. Almazyad AS, Ahamad MG, Siddiqui, MK & Almazyad AS. Effective hypertensive treatment using data mining in Saudi Arabia. *Journal of Clinical Monitoring and Computing* 2010; 24 (6): 391– 401.
33. Fang X. Are you becoming a diabetic? A data mining approach. In: Proc. *The 6th International Conference on Fuzzy Systems and Knowledge Discovery, Tianjin, China* .2009: 14-16: 18-22.
34. Han J, Rodriguez JC & Beheshti M. Diabetes data analysis and prediction model discovery using rapidminer. *In: Future Generation Communication and Networking, FGCN'08, Second International Conference on, Hainan Island, China*, 2008; 13-15, 3: 96-99.
35. Anbananthen KSM, Sainarayanan G, Chekima A & Teo J. Artificial Neural Network Tree Approach in Data Mining. *Malaysian Journal of Computer Science* 2007; 20(1):51-62.

36. Juan G, Luo S, Jia H, Zhang T & Han Y. Type 2 diabetes data processing with EM and C4. 5 algorithm. In: Complex Medical Engineering, CME 2007. *IEEE/ICME International Conference on, Beijing, China*, 2007; 371-377.
37. Miyaki K, Takei I, Watanabe K, Nakashima H, Watanabe K, & Omae K. Novel statistical classification model of type 2 diabetes mellitus patients for tailor-made prevention using data mining algorithm. *Journal of epidemiology* 2002; 12 (3):243-248.
38. Rohlfing CL, Wiedmeyer HM, Little RR, England JD, Tennill A & Goldstein DE. Defining the relationship between plasma glucose and HbA_{1c}: analysis of glucose profiles and HbA_{1c} in the Diabetes Control and Complications Trial. *Diabetes care* 2002; 25(2):275-278.
39. Shahrabi J. *Data Mining Book*. Gita Data Processors Research Institute and Jihad University Amirkabir Industrial Branch. First Edition, 2008.
40. Turban E, Aronson JE, Liang TP & Sharda R. *Decision support and business intelligence systems (8th ed.)*, Pearson Prentice Hall, New Jersey, USA, 2007.
41. Duarte de Araujo FH. Evaluation of Classifiers Based on Decision Tree for Learning Medical Claim Process. *Latin America Transactions, IEEE (Revista IEEE America Latina)* 2015; 13(1): 299 – 30.
42. San PP, Ling SH, Nguyen HT. Intelligent detection of hypoglycemic episodes in children with type 1 diabetes using adaptive neural-fuzzy inference system. *Conf Proc IEEE Eng Med Biol Soc* 2012:6325-8.
43. David SK, Saeb AT, Al Rubeaan K. Comparative analysis of data mining tools and classification techniques using weka in medical bioinformatics. C*omputer Engineering and Intelligent Systems* 2013; 4(13):28-38.
44. Devi MR, Shyla JM. Analysis of various data mining techniques to predict diabetes mellitus. *International Journal of Applied Engineering Research* 2016; 11(1):727-30.
45. Patil TR & Sherekar SS. Performance analysis of Naive Bayes and J48 classification algorithm for data classification. *International Journal of Computer Science and Applications* 2013; 6 (2): 256-261.
46. Wang C, Li L. Evaluating the risk of type 2 diabetes mellitus using artificial neural network: An effective classification approach. *Diabetes Research and Clinical Practice* 2013; 100 (1):111–118.
47. Langarizadeh M, Ghazi Saeedi M. Predicting Premature Birth in Pregnant Women via Assisted Reproductive Technologies using Neural Network. *Journal of Health Administration* 2016; 18(62): 42–51 [In Persian].
48. Huang GM, Huang KY, Lee TYi, Weng J. An interpretable rule-based diagnostic classification of diabetic nephropathy among type 2 diabetes patients. *BMC Bioinformatics* 2015; 16(suppl1): S5.
49. Balakrishnan S, Ramaraj N, Savarimuthu N, Samikannu R. SVM ranking with backward search for feature selection in type II diabetes databases. *IEEE International Conference on Systems, Man and Cybernetics* 2008; 12- 15; Singapore; 2008.

Comparison of the Efficiency of Data Mining Algorithms in Predicting the Diagnosis of Diabetes

Fatemeh Dekamini*[1], Mohammad Ehsanifar[2]

1. *Department of Industrial Management, Faculty of Management, Arak Branch, Islamic Azad University, Arak, Iran*
2. *Department of Industrial Engineering, Faculty of Engineering, Arak Branch, Islamic Azad University, Arak, Iran*

ABSTRACT

Background: Diabetes is one of the major health problems in Iran and about 4.6 million adults suffer from this disease. Poor diagnosis of this disease has caused half of this number to be unaware of their disease. In recent years, along with the use of computers in data analysis and storage, the volume and complexity of data has increased dramatically.
Methods: In health organizations, data play an essential role in the value of the organization. Therefore, data mining has become one of the most widely used processes in the field of health and disease diagnosis. In this study, the information of 768 laboratory clients in Tehran was kept confidential and the opinions of experts were used to identify the variables affecting the incidence of diabetes.
Results: The findings indicate the study of 5 algorithms on the presented data, which by implementing 5 data mining algorithms J48, Bayes, Beginning, Cohen and simple clustering to classify the data, the efficiency of these algorithms in terms of speed and accuracy in calculations was evaluated.
Conclusion: The data set for classification is the database of a laboratory, which includes 768 samples with 9 characteristics. Finally, J48 algorithm is recommended for data mining of diabetes due to high speed, acceptable accuracy and lack of sensitivity to raw data.

Keywords: Data Mining, Diabetes, Efficiency, Knowledge Discovery

* Imam Khomeini Square, Imam Khomeini Boulevard, 3 km Khomein Road, Amirkabir University Town, Arak Islamic Azad University, Arak, Iran. PO Box: 38135/567, Tel: +989334132451-086. Postal Code: 9131-1-38361, Email: s_dekamin@yahoo.com