مجله دانشگاه علوم پزشکی تربت حیدریه
دوره ۷. شماره ۱. بهار ۱۳۹۸

مقاله پژوهشی

تشخیص بیماری دیابت با استفاده از روش‌های مبتنی بر داده‌کاوی با تکیه بر داده های بومی

ایمان عابدیان[1*]، علی ایوبی[1]، حمیدرضا غفاری[1]، ایمان ذباح[2]

۱. گروه مهندسی کامپیوتر، دانشگاه آزاد اسلامی واحد فردوس، فردوس، ایران
۲. گروه مهندسی کامپیوتر، دانشگاه آزاد اسلامی واحد تربت حیدریه، تربت حیدریه، ایران

چکیده

زمینه و هدف: بیماری دیابت یکی از شایع ترین و پر هزینه ترین بیماری ها می باشد که تشخیص به موقع آن می‌تواند منجر به کاهش پیشرفت این بیماری و عوارض ناشی از آن شود. این پژوهش با هدف تعیین وضعیت بیماری دیابت ازنظر ابتلا و یا عدم ابتلا به آن، با استفاده از تکنیک‌های داده‌کاوی انجام شده است.

روش‌ها: این مطالعه از نوع تحلیلی بوده و پایگاه داده آن شامل ۲۵۴ رکورد مستقل مبتنی بر ۱۳ ویژگی و جمع‌آوری‌شده توسط محققین طرح از یکی از مراکز تخصصی دیابت شهرستان مشهد می‌باشد.

نتایج: پس از پیش‌پردازش داده‌ها روش‌های مختلف تشخیص الگو مورد بررسی قرار گرفتند، با استفاده از شبکه عصبی پرسپترون چندلایه MLP، شبکه عصبی LVQ، بردار پشتیبان SVM و روش خوشه‌بندی K_means، میانگین حداقل مربعات خطا محاسبه گردید. صحت عملکرد هر یاد گیر به ترتیب ۹۴٪، ۹۲٪، ۹۶٪ و ۹۳٪ محاسبه گردید.

نتیجه‌گیری: نتایج مطالعه حاکی از آن است که روش SVM عملکرد بهتری نسبت به سایر روش‌ها در تشخیص بیماری دیابت دارد.

کلید واژه‌ها:
دیابت شیرین، شبکه عصبی مصنوعی، بردار یادگیر پشتیبان، خوشه‌بندی

مقدمه

امروزه دیابت تبدیل به یک بیماری گسترده در سراسر دنیا شده است که میلیون‌ها نفر به آن مبتلا هستند. طبق آخرین آمارهای منتشرشده سازمان بهداشت جهانی و فدراسیون بین‌المللی دیابت، از هر ٤ نفر بالای سن ٦٠ سال یک نفر به این بیماری مبتلا می‌باشند. از این‌رو این بیماری به یکی از چالش‌های مهم نظام‌های بهداشت و درمان کشورهای مختلف دنیا، چه درحال‌توسعه و چه پیشرفته، تبدیل‌شده است (۱). عدم تشخیص به موقع و یا ضعف در تشخیص این بیماری ازجمله مشکلات عمده دیگری است که در رابطه با این بیماری وجود دارد (۲). این امر تا حدی به دلیل عدم انتخاب الگوی مناسب توسط پزشک و یا عدم استفاده مناسب از الگوهای استاندارد موجود است؛ بنابراین پیاده‌سازی روشی که بتواند هر فرد را در تشخیص صحیح ابتلا یا عدم ابتلا به این بیماری یاری رساند می تواند گام مهمی در جهت پیشگیری و کنترل این بیماری بخصوص در مراحل ابتدایی آن تلقی گردد (۳).

داده‌کاوی روشی است که برای تحلیل منظم داده‌ها و شناسایی الگوهای پنهان در بین آن‌ها بکار می‌رود. عملی که انجام آن به‌صورت دستی امکان‌پذیر نیست. روش‌های مختلف داده‌کاوی، به‌طور گسترده‌ای در تحقیقات پزشکی بکار رفته است که این امر می‌تواند در تشخیص بیماری و کاهش اشتباهات به پزشکان

*آدرس نویسنده مسئول: فردوس، دانشگاه آزاد اسلامی واحد فردوس، گروه مهندسی کامپیوتر
آدرس پست الکترونیک: iman.abedian@gmail.com

کمک شایانی نماید (٥،٤). متدهای مختلفی ازجمله استفاده از روش‌های تکاملی (٦)، مدل‌سازی فازی (٧)، روش‌های مبتنی بر بیزین (٨)، تشخیص الگو در استخراج ویژگی (٩)، تشخیص دیابت به کمک بردار ماشین پشتیبان (Support vector machine) (١٠)، استفاده از درخت تصمیم در تشخیص دیابت نوع دو (١١) و تشخیص بیماری دیابت با استفاده از شبکه‌های عصبی و فازی عصبی (١٢) بکار گرفته شده است.

ضرورت فرآیند کشف الگو و تبدیل داده‌ها به دانش، به‌منظور کمک به تصمیم‌گیری به داده‌کاوی نسبت داده‌شده است و از آن به منظور استخراج الگو استفاده می‌شود. اگرچه علاوه بر داده‌کاوی سایر روش‌ها، مانند استفاده از سیگنال‌های قلبی (١٣) و روش‌های بینایی ماشین (١٤) برای طراحی سیستم شناسایی بیماری دیابت موردتوجه قرارگرفته‌اند، اما هزینه بالا و کارایی کمتری نسبت به روش‌های مبتنی بر داده‌کاوی دارند. به طورکلی وظیفه کاوش داده به دو بخش اصلی تقسیم می‌گردد: روش‌های توصیفی و روش‌های پیش بینانه. روش توصیفی خواص عمومی داده‌ها را مشخص می‌کند و هدف آن پیدا کردن الگوهای قابل تفسیر توسط انسان برای داده‌ها است، اما روش پیش بینانه، پیش‌بینی رفتار آینده آن‌ها را مشخص می‌کند و هدف آن به‌کارگیری چند متغیر یا ویژگی در پایگاه داده برای پیش‌بینی مقادیر آینده یا ناشناخته دیگر است. یکی از انواع پیش‌بینی، دسته‌بندی است. دسته‌بندی فرآیند دستیابی به مدلی است که با تشخیص دسته‌ها یا مفاهیم داده می‌تواند دسته ناشناخته دیگری را پیش‌بینی کند. درواقع دسته‌بندی یک تابع یادگیری است که یک قلم داده را به یکی از دسته‌های از قبل تعریف‌شده ثبت می‌کند. به‌کارگیری متدهای داده‌کاوی بیماری دیابت می‌تواند با هدف تشخیص و پیش‌بینی دیابت، پیش‌بینی بروز عوارض و یا تعیین میزان دارو و نوع درمان انجام شود. هدف این مطالعه استفاده از شیوه‌های مختلف داده‌کاوی و مقایسه آن‌ها با یکدیگر است. ضمن این‌که به‌منظور بومی‌سازی مدل‌سازی تمرکز این پژوهش بر روی داده‌های بومی صورت گرفته است. کلاس‌بندی بیماری دیابت با متدهای مبتنی برداده کاوی می‌تواند به پزشک در تشخیص این بیماری کمک کرده و به عبارتی دستیار پزشک باشد.

روش‌ها

این مطالعه از نوع توصیفی تحلیلی است که به‌صورت مقطعی در سال ١٣٩٧ انجام‌شده است. جامعه پژوهش، بیماران مبتلا به دیابت بودند و محیط پژوهش نیز یکی از مراکز تخصصی بیماری دیابت بود. یکی از پایگاه داده‌های مورد استفاده در اکثر مطالعات داده‌کاوی بیماری دیابت، مجموعه داده‌های (Pelvic inflammatory disease) است که در سایت اینترنتی دانشگاه کالیفرنیای آمریکا قابل‌دسترس است. این نمونه‌ها به عنوان مرجعی جامع جهت بررسی الگوریتم‌های یادگیری ماشین در بیماری دیابت در بسیاری از پژوهش‌ها استفاده می‌شود و در سال ١٩٩٨ جمع‌آوری شده و در مرجع (١٥) قابل‌دسترس است. این نمونه‌ها دارای ٨ ویژگی: گلوکز، تعداد دفعات بارداری، فشارخون دیاستولیک و ضخامت پوست ماهیچه‌ای سه سر بازویی، انسولین شاخص توده بدنی، سابقه بیماری، سن مربوط به ٥٠٠ زن سالم و ٢٦٨ زن مبتلابه بیماری دیابت می‌باشد که مطابق با شاخص‌ها و استانداردهای سازمان بهداشت جهانی جمع‌آوری گردیده است. در یک مطالعه که در سال ١٣٩٣ با عنوان پیش‌بینی به ابتلای بیماری دیابت با استفاده از شبکه عصبی مصنوعی صورت گرفت (١٦) تعداد ویژگی‌های ثبت‌شده هر بیمار ٨ مورد بود که با ویژگی‌های پایگاه داده PID متفاوت است. همچنین در مطالعه عامری و همکاران که در سال ١٣٨٨ از مرکز دیابت استان گلستان جمع‌آوری‌شده است تعداد ٢٥٠ رکورد با ١٢ ویژگی پس از پالایش داده‌ها ثبت شده است. این ویژگی‌ها الزاماً مشابه نمونه‌های پایگاه داده PID نیستند، بااین‌وجود در مطالعه مورد بررسی محقق ضمن پیش‌بینی وضعیت دیابت نشان داده است که چه پارامترهایی در تشخیص بیماری از اهمیت بالاتری برخوردار می‌باشند (١٧). در مطالعه حاضر که با تکیه بر داده‌های بومی انجام‌شده است علاوه بر ویژگی‌هایی که در پایگاه داده PID ثبت شده، پارامترهای دیگری نیز مانند وزن و

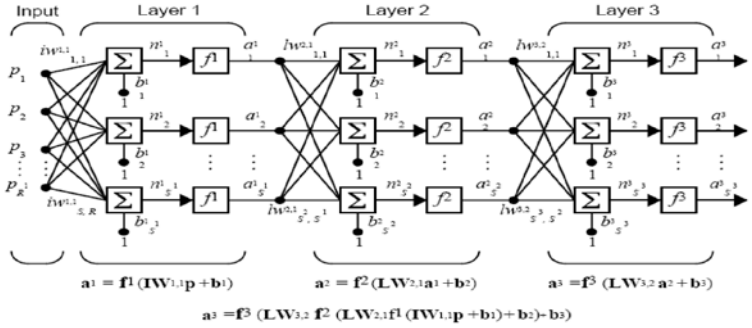
دور کمر که در سایر مطالعات (۱۸،۱۹) به‌عنوان ویژگی‌های مهم تشخیص دیابت ذکر شده‌اند نیز ثبت گردیده است. علاوه بر این نه تنها فشارخون دیاستولیک بلکه فشارخون سیستولیک هم به‌عنوان یکی دیگر از پارامترهای بالینی بیمار اندازه‌گیری و ثبت‌شده است. به این ترتیب در این مطالعه در مجموع ۱۳ پارامتر اندازه‌گیری و ثبت گردید: میانگین سیستول، میانگین دیاستول، مقدار وزن، قد، نمایه توده بدنی، دور کمر، کلسترول، تری گلیسیرید، HDL-C، LDL-C، گلوکز ۲ ساعت پس از صبحانه، قند خون ناشتا، هموگلوبین گلیکوزیله. داده‌ها توسط نرم‌افزار Excel (2013) ثبت و با SPSS (23) تحلیل و توسط MATLAB (2016) مورد برنامه‌نویسی قرارگرفته‌اند. جدول ۱ آماره‌های توصیفی مربوط به بیماران را نشان می‌دهد.

جدول ۱. آماره‌های توصیفی مربوط به متغیرهای کیفی بیماران دیابتی

ردیف	نام ویژگی	نوع	حداقل	حداکثر	میانگین ± انحراف معیار
۱	میانگین سیستول	سالم	۱۰۰	۱۳۰	۱۱۶/۲۶ ± ۵/۱۰
		بیمار	۱۰۰	۱۹۰	۱۳۵/۶۴ ± ۲۳/۸۳
۲	میانگین دیاستول	سالم	۶۰	۷۵	۶۷/۵۳ ± ۵/۱۵
		بیمار	۶۰	۱۱۰	۸۰/۶۶ ± ۱۳/۷۰
۳	مقدار وزن	سالم	۷۱	۷۹	۷۵/۶۸ ± ۲/۳۸
		بیمار	۵۷	۱۱۴	۷۸/۹۱ ± ۱۱/۴۰
۴	قد	سالم	۱۵۳	۱۷۸	۱۶۵/۶۶ ± ۶/۳۹
		بیمار	۱۳۵	۱۹۲	۱۶۷/۳۳ ± ۱۱/۳۳
۵	نمایه توده بدنی	سالم	۱۸	۲۴	۲۰/۹۷ ± ۱/۹۰
		بیمار	۲۰.۳	۳۸/۱	۲۸/۱۸ ± ۲/۸۹
۶	دور کمر	سالم	۸۰	۸۹	۸۴/۶۸ ± ۲/۵۷
		بیمار	۷۸	۱۲۸	۹۸/۵۴ ± ۱۰/۹۲
۷	کلسترول	سالم	۱۵۹	۱۹۹	۱۸۰/۶ ± ۱۱/۲۸
		بیمار	۹۲	۲۶۵	۱۶۸/۳۵ ± ۲۹/۷۳
۸	تری گلیسرید	سالم	۱۵۵	۱۹۹	۱۷۸/۹۷ ± ۱۲/۰۸
		بیمار	۳۷	۱۹۰	۱۱۲/۷۸ ± ۳۴/۸۴
۹	HDL-C	سالم	۴۰	۵۹	۹۲/۵۰ ± ۵/۴۲
		بیمار	۳۱	۳۳۷	۵۹/۱۲ ± ۲۶/۶۴
۱۰	LDL-C	سالم	۱۰۰	۱۲۹	۱۱۸/۹۸ ± ۷
		بیمار	۵۳	۱۹۸	۱۲۳/۹۲ ± ۳۴/۱۳
۱۱	گلوکز ۲ ساعت پس از صبحانه	سالم	۱۱۷	۱۳۹	۱۲۷/۹۸ ± ۶/۱۹
		بیمار	۱۲۹	۳۶۵	۲۶۰/۹۹ ± ۴۲/۳۶
۱۲	قند خون ناشتا	سالم	۷۰	۹۹	۸۴/۰۶۶ ± ۸/۳۸
		بیمار	۸۰	۳۱۷	۱۶۳/۳۳ ± ۳۹/۳۴
۱۳	همو گولو بین گلیکوزیله	سالم	۴.۷	۷	۵/۸۷ ± ۰/۶۰
		بیمار	۱۲.۵	۵/۱	۸۷/۲۲ ± ۱/۲۷

داده‌ها از پرونده بیماران جمع‌آوری گردید. تعداد پرونده‌های اولیه و حاوی اطلاعات دموگرافیک و بالینی ۵۲۱ پرونده بود. از این تعداد حدود ۲۶۷ مورد به دلیل نقص در پرونده و یا عدم مراجعه بیمار و به کمک پزشک متخصص ناقص تشخیص داده شد و به‌عنوان داده‌های گم شده حذف گردیدند؛ پس از پالایش اولیه، داده‌های قابل‌استفاده به ۲۵۴ مورد تقلیل پیدا کرد. از این بین ۷۵ نفر به تشخیص پزشک سالم و ۱۷۹ نفر مبتلابه دیابت تشخیص داده شده بودند. در این مطالعه انتخاب ویژگی با شیوه‌های رایج صورت نگرفته است و کل ویژگی‌های ثبت‌شده جهت آموزش به سیستم‌های یادگیری اعمال گردید. روش مطالعه برای دسته‌بندی بیماران به دو صورت یادگیری نظارت‌شده و غیر نظارت شده بود. در روش ناظر از شبکه‌های عصبی مصنوعی پرسپترون چندلایه Multi-Layer Perceptron و شبکه عصبی بردار یادگیر Learning Vector Quantization و نیز بردار ماشین پشتیبان Support Vector Machine و در روش بدون ناظر از روش K_means استفاده گردید.

تشخیص بیماری دیابت با استفاده از شبکه عصبی مصنوعی پرسپترون (MLP): شبکه‌های عصبی سیستم های دینامیکی هستند که با پردازش روی داده‌های تجربی، دانش یا قانون نهفته در ورای داده‌ها را به ساختار شبکه منتقل می‌کنند. این سیستم‌ها بر اساس محاسبات انجام شده بر روی داده‌های اولیه و جدید ارائه شده به آن قوانین کلی را فرا می‌گیرند. پردازش اطلاعات در شبکه‌های عصبی روشی مشابه مغز انسان دارد (۲۰). هر شبکه عصبی شامل یک‌لایه ورودی است که هر گره در این لایه معادل یکی از ویژگی‌های مسئله می‌باشد. گره‌های موجود به تعدادی گره در لایه نهان متصل و سپس به همه گره‌های لایه نهان وصل می‌شوند. هر یال بین نودها دارای یک وزن است. این وزن‌ها در محاسبات لایه‌های میانی استفاده می‌شوند. وزن یال‌ها پارامترهای نامعینی هستند که توسط تابع آموزش و داده‌های آموزشی که به سیستم داده می‌شود تعیین می‌شوند. تعداد گره‌ها و تعداد لایه‌های نهان و نحوه وصل شدن گره‌ها به یکدیگر معماری شبکه عصبی را مشخص می‌کند. در طراحی شبکه‌های عصبی باید تعداد نودها، تعداد لایه‌های نهان، تابع فعال‌سازی را مشخص کرد. مدل پرسپترون چندلایه کاربرد موفقیت‌آمیزی در حل برخی مسائل ازجمله شناسایی الگو و تخمین تابع داشته است. شبکه سه لایه به کاربرده شده در این مقاله با ورودی و خروجی در شکل ۱ نشان داده‌شده است.

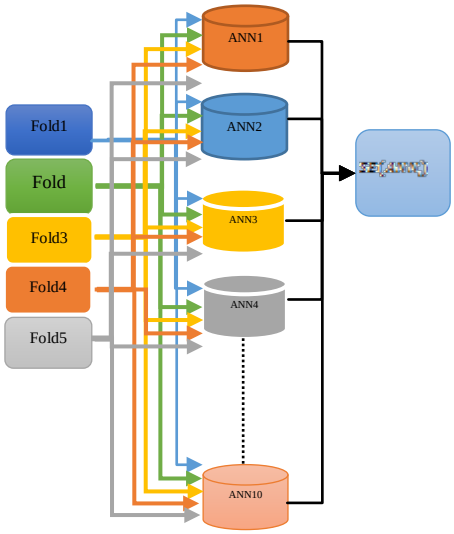


شکل ۱. معماری شبکه‌های عصبی پرسپترون سه لایه مورداستفاده در این مطالعه

ازآنجایی‌که معماری‌های مختلف از شبکه‌های عصبی با توجه به تعداد نورون در لایه مخفی قابل طراحی است، در این مطالعه شبکه عصبی پرسپترون چندلایه با معماری‌های متنوع استفاده گردید. همچنین الگوریتم پس انتشار خطا (Feed forward Back propagation) برای محاسبه وزن‌های شبکه به کار گرفته شد.

انتخاب نمونه‌های آموزش و آزمون: یکی از مواردی که همیشه در مدل سازی به کمک سیستم های یاد گیر با آن مواجه هستیم نحوه انتخاب داده‌های آموزش است. به‌صورت معمول حدود ۷۰٪ از داده‌ها به‌عنوان داده‌های آموزش مورد استفاده قرار می‌گیرند و انتخاب نمونه‌ها جهت انجام فرآیند

آموزش به‌صورت تصادفی انجام می‌شود. این مسئله باعث می‌شود که در اجراهای متعدد شبکه‌های عصبی، نتایج مختلف حاصل شود. اگر نمونه‌های آموزشی به‌اندازه کافی متنوع نباشند، یعنی اینکه از تمام رده‌ها به‌اندازه کافی نمونه جهت آموزش سیستم وجود نداشته باشد، عمل یادگیری به‌درستی انجام نخواهد شد. برای رفع این مشکل ۲ روش پیشنهاد می‌گردد: ۱- تغییر ساختارهای متعدد سیستم یادگیر (شبکه عصبی) ازنظر تعداد نورون‌ها و محاسبه میانگین خطا. و ۲- استفاده از روش اعتبار سنجی متقابل (kfold) به‌منظور خوراندن نمونه‌های متعدد به سیستم و محاسبه میانگین خطا.



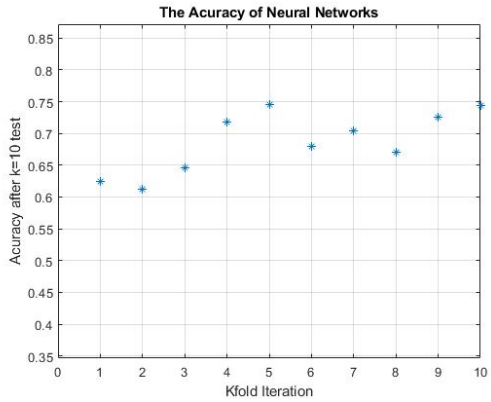
شکل ۲. الف: اجرای kfold بر روی ۱۰ شبکه عصبی

تشخیص بیماری دیابت با استفاده از بردار پشتیبان (SVM):

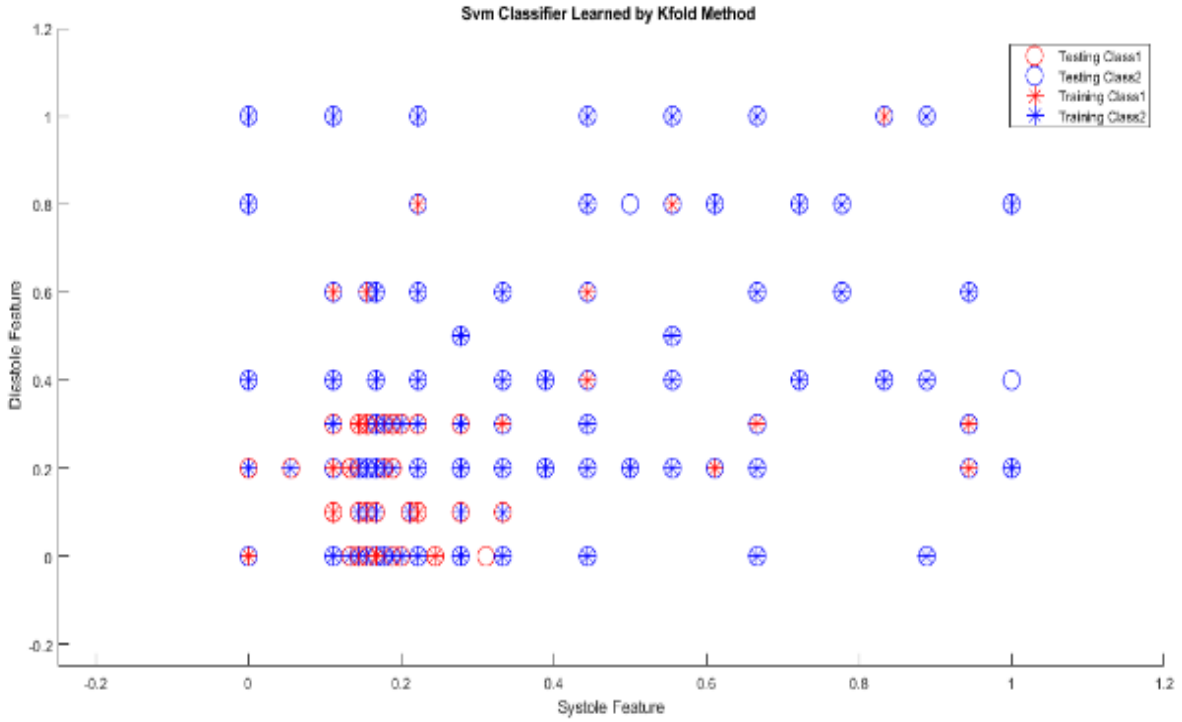
ماشین بردار پشتیبانی یکی از الگوریتم‌های یادگیری با نظارت است که اغلب برای دسته‌بندی های باینری استفاده می‌شود (۱۹،۲۰). ازآنجایی‌که در تشخیص بیماری دیابت با یک مسئله ۲ کلاسه مواجه هستیم، لذا در بخش دوم این مطالعه از طبقه‌بندی SVM استفاده شده است. نظر به اینکه تعداد ویژگی‌ها در بیماری دیابت ۱۳ مورد است، لذا یک طبقه بند غیرخطی مورد نیاز است. باز هم جهت اطمینان از صحت عملکرد svm نمونه‌ها توسط روش kfold به ۵ قسمت تقسیم‌شده و مانند قبل هر بار از ۳ قسمت جهت train یک قسمت جهت test استفاده گردید و درنهایت میانگین عملکرد محاسبه گردید. شکل ۳ نحوه عملکرد svm در کلاس بندی بیماران دیابتی را نشان می‌دهد. به‌منظور ترسیم نمونه‌ها، بر اساس ۲ ویژگی سیستول و دیاستول ترسیم شده‌اند.

در این مطالعه از هر دو روش فوق برای افزایش دقت تشخیص وضعیت بیماری دیابت استفاده شده است.

به این ترتیب تعداد ۲۵٤ نمونه با استفاده از روش اعتبارسنجی متقابل به ۵ قسمت تقسیم گردید که هر بار از ٤ قسمت برای آموزش (Train) شبکه و ۱ قسمت به عنوان آزمون (Test) مورداستفاده قرار گرفت. تعداد نورون‌ها در لایه‌های شبکه عصبی بین ۱۰ تا ۲۰ نورون تغییر و ۱۰ بار، عملیات صحت سنجی تکرار گردید. شکل ۲ الف و ب به ترتیب نحوه اجرای Kfold بر روی شبکه‌های عصبی و متوسط عملکرد ۱۰ شبکه را نشان می‌دهد.



شکل ۲. ب: متوسط صحت عملکرد ۱۰ شبکه عصبی



شکل ۳. تفکیک نمونه‌ها توسط SVM به دو کلاس بیمار و سالم بر اساس دو ویژگی سیستول و دیاستول

ازآنجایی‌که فضای هر نمونه یک بردار با ۱۳ ویژگی است، لذا امکان نمایش همه ویژگی‌ها و نحوه کلاس‌بندی آن توسط شکل وجود ندارد. لذا شکل ٤ فقط محدود به ۲ ویژگی رسم شده است تا بتواند تا حدی عملکرد SVM را نمایش دهد. دوایر، مربوط به آزمون و ستارها مربوط به فرآیند آموزش هستند. رنگ‌های آبی و قرمز نیز بیانگر ۲ کلاس بیمار و سالم است. اگر دایره آبی بر ستاره آبی و دایر قرمز بر ستاره قرمز منطبق شده باشد به معنی صحت کلاس‌بندی توسط SVM برای آن نمونه است.

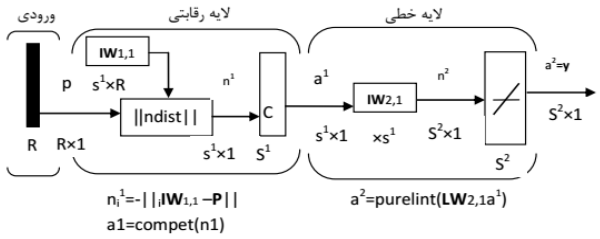
تشخیص بیماری دیابت با استفاده از شبکه عصبی LVQ:

شبکه عصبی LVQ دارای دولایه رقابتی و خطی می‌باشد. لایه رقابتی دسته‌بندی کردن بردارهای ورودی را یاد گرفته و در نهایت لایه خطی کلاس‌های لایه رقابتی را به دسته‌های هدف که توسط کاربر تعیین‌شده ثبت می‌کند. شکل ٤. الف معماری شبکه‌های LVQ را نشان می‌دهد. در این شکل S1 و S2 به ترتیب تعداد نورون های لایه رقابتی و خطی و R تعداد عضو-های بردار ورودی می‌باشند. هردو لایه رقابتی و خطی دارای یک نورون به ازای هر کلاس می‌باشند (۱۷). در این مطالعه از ۱۰ شبکه عصبی LVQ استفاده شد تا بتوان میانگین خطای معماری‌های مختلف شبکه جهت تشخیص بیماری را محاسبه نمود. در تمامی شبکه‌ها از ۸۰٪ داده جهت آموزش شبکه و ۱۰٪ جهت اعتبارسنجی و ۱۰٪ به‌منظور آزمون استفاده گردید. تعداد تکرارها (Epochs) برای تمامی آن‌ها ۳۰ و الگوریتم یادگیری از نوع LVQ2 انتخاب‌شده است. شکل ٤.ب وضعیت خطای شبکه-های عصبی در ازای تغییر تعداد نورون‌های لایه رقابتی را نشان می‌دهد.

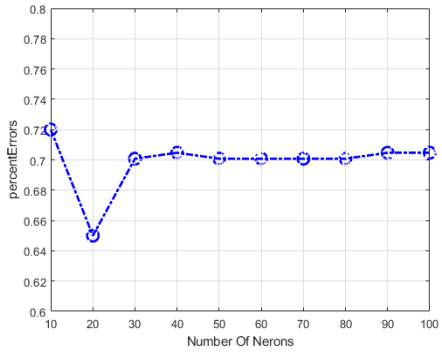
شکل ٤. ب حاکی از آن است زمانی که تعداد نورون‌ها در لایه رقابتی ۲۰ انتخاب‌شده است رفتار الگوریتم LVQ2 عملکرد بهتری داشته است. هم‌چنین نتایج نشان می‌دهد که الزاماً با افزایش تعداد نورون‌های لایه رقابتی وضعیت شبکه بهتر نشده است.

تشخیص بیماری دیابت با استفاده از خوشه بندی K_means:

خوشه‌بندی، یک روش یادگیری بدون نظارت است که روی دسته‌های از قبل تعریف شده و یا ویژگی خاصی



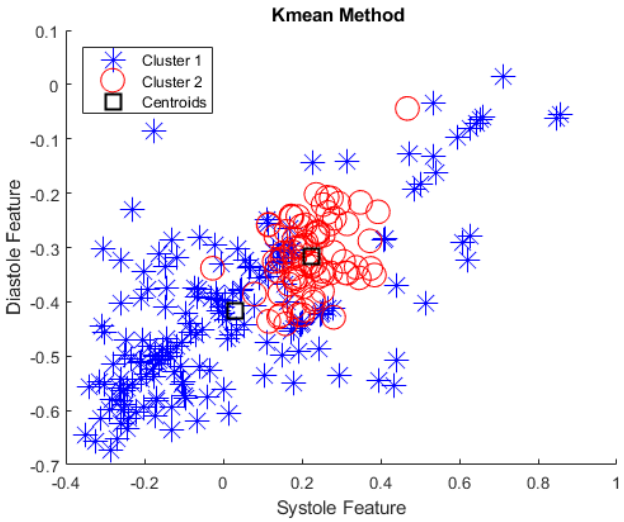
شکل ۴. الف: معماری LVQ به‌منظور تشخیص دیابت



شکل ۴. ب: نمودار خطا در تشخیص دیابت با تغییر نورون‌های لایه رقابتی با الگوریتم LVQ2

به‌عنوان هدف تکیه ندارد، بلکه نمونه‌های مشابه با هم در یک حجم داده را گروه‌بندی می‌کند. اگرچه در مسئله تشخیص بیماری دیابت با یک مسئله نظارت‌شده روبرو هستیم و روش‌های با ناظر از عملکرد بهتری جهت مسائل تشخیصی دارند. لیکن مطالعاتی نیز جهت استفاده از روش‌های غیر نظارت شده بر روی تشخیص بیماری‌ها صورت گرفته است که ازاین‌بین می‌توان به پژوهش فیروزی که از روش K_means به منظور بررسی ویژگی‌های بیماران مبتلا به سل (۲۱) استفاده کرد و یا پژوهش قاسم‌زاده که به تعیین عوامل مؤثر بر سرطان پوست غیر ملانومایی با استفاده از K_means است (۲۲) اشاره نمود.

مهم‌ترین عامل در خوشه‌بندی، معیار شباهت است. به این معنی که اشیاء داخل یک خوشه به هم شبیه هستند. شباهت هر خوشه نسبت به متوسط اشیاء آن خوشه سنجیده می‌شود. خوشه‌ای به‌عنوان خوشه بهینه شناخته می‌شود که اشیاء هر خوشه در دسته‌های مجزا قرار داشته باشند و تداخلی باهم نداشته باشند. هر چه خوشه‌ها متمرکزتر باشند، عمل خوشه‌بندی بهینه‌تر انجام‌شده است. شکل ۵ وضعیت خوشه‌بندی به روش K_means در شکل زیر دیده می‌شود.



شکل ۵. خوشه‌بندی به روش K_means و رسم نمودار بر اساس دو ویژگی سیستول و دیاستول

نتایج

در یک بررسی آماری، ضریب همبستگی ویژگی‌های نمایه توده بدنی و هموگلوبین محاسبه گردید که به ترتیب برابر ۰/۷۸ و ۰/۶۹ اندازه‌گیری شدند. ضریب همبستگی سایر ویژگی‌ها عبارتند از: میانگین سیستول: ۰/۴۵، میانگین دیاستول:۰/۴۰، وزن ۰/۱۴، قد: ۰/۰۷، دور کمر: ۰/۵۶، کلسترول: ۰/۲۱-، تری گلیسیرید: ۰/۰۷-، HDL: ۰/۱۳، LDL: ۰/۰۷، گلوکز :۰/۸۶ و قند خون:۰/۷۳. هدف نهایی یک سیستم شناسایی الگو رسیدن به بالاترین نرخ طبقه‌بندی ممکن برای مسئله موردنظر است. ازآنجایی‌که هیچ الگوریتم طبقه‌بندی وجود ندارد که به تنهایی به طور کامل برای تمام مسائل مناسب باشد، بررسی روش‌های مختلف تشخیص الگو به‌عنوان یک‌راه حل برای افزایش کارایی آن‌ها پیشنهاد شده است. در مطالعات مختلف تحت عنوان داده-کاوی پزشکی، راهکارهای متعددی جهت کشف روابط بین عوامل بیماری دیابت ارائه شده است. استفاده از شبکه‌های

عصبی مصنوعی جهت تشخیص دیابت در مطالعات مختلف مورد بررسی قرارگرفته است. در این پژوهش سعی شده است برای بهبود تشخیص بیماری به بررسی روش‌های با و یا بدون ناظر پرداخته شود.

در ابتدا به پالایش نمونه‌ها پرداخته و داده‌های ناقص را از مجموعه پایگاه داده حذف گردید. اختلاف مقدار در ویژگی‌های نمونه‌ها بارز می‌باشد. به عنوان مثال میانگین سیستول در افراد سالم مطابق جدول ۱ برابر ۱۱۶ و در مورد افراد بیماری برابر ۱۳۵ می‌باشد. درحالی‌که میانگین هموگلوبین در افراد سالم ۵/۸ و در افراد بیمار ۸/۲ است. این اختلاف عملاً موجب کاهش دقت یاد گیرها و کاهش سرعت همگرایی آن‌ها خواهد شد. لذا عمل نرمال‌سازی داده‌ها مطابق رابطه (۱) انجام‌گرفته:

رابطه (۱):

$$xn = \frac{x - x_{min}}{x_{max} - x_{mhn}}$$

در این رابطه، X نمونه‌های نرمال نشده، X_n داده‌های نرمال شده و X_{max} و X_{min} حداقل و حداکثر ورودی می‌باشند. همچنین از معیار مجذور مربعات خطا (RMSE) مطابق رابطه ۲ جهت ارزیابی عملکرد ماشین بردار و شبکه‌های عصبی استفاده‌شده است.

رابطه (۲):

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{N}}$$

پس از پالایش و نرمال‌سازی نمونه‌ها در اولین مرتبه مدل‌سازی از یک شبکه عصبی پرسپترون چندلایه با الگوریتم پس انتشار خطا استفاده شد. از این شبکه عصبی، به‌عنوان یک طبقه بند برای کلاس‌بندی داده‌ها به دو دسته سالم بیمار استفاده گردید و کلیه ۱۳ فاکتور مربوط به بیماری دیابت به شبکه عصبی اعمال شد. آموزش شبکه با ۷۰٪ داده‌ها و آزمون آن با ۳۰٪ از دادها انجام گردید. هم‌چنین معماری‌های مختلف شبکه عصبی برای حصول بهترین کلاس‌بندی مورد بررسی قرار گرفت. از آنجایی‌که نتایج شبکه‌های عصبی در هر بار آموزش متفاوت است، لذا از روش Kfold به منظور تنوع نمونه‌های آموزشی استفاده شده است.

پس از پالایش و نرمال‌سازی نمونه‌ها در اولین مرتبه مدل‌سازی از یک شبکه عصبی پرسپترون چندلایه با الگوریتم پس انتشار خطا استفاده شد. از این شبکه عصبی، به‌عنوان یک طبقه بند برای کلاس‌بندی داده‌ها به دو دسته سالم بیمار استفاده گردید و کلیه ۱۳ فاکتور مربوط به بیماری دیابت به شبکه عصبی اعمال شد. آموزش شبکه با ۷۰٪ داده‌ها و آزمون آن با ۳۰٪ از دادها انجام گردید. هم‌چنین معماری‌های مختلف شبکه عصبی برای حصول بهترین کلاس‌بندی مورد بررسی قرار گرفت. از آنجایی‌که نتایج شبکه‌های عصبی در هر بار آموزش متفاوت است، لذا از روش Kfold به منظور تنوع نمونه‌های آموزشی استفاده‌شده است. در مرتبه دوم مدل‌سازی، از بردار یادگیر پشتیبان استفاده شد و نمونه‌ها جهت آموزش با روش kfold به یادگیر اعمال گردید. در مرحله سوم از شبکه عصبی LVQ استفاده گردید. علاوه بر بررسی شبکه‌های عصبی با معماری-های مختلف تعداد ورودی های بیماری در ورودی شبکه عصبی نیز موردبررسی قرار گرفت. در مرحله نهایی از روش یادگیری بدون ناظر K_means استفاده شد. شکل ٦ الگوریتم ارائه‌شده در این پژوهش را نشان می‌دهد. جدول ۲ نتایج برخی از پژوهش‌هایی که در حوزه تشخیص بیماری دیابت انجام‌شده است به همراه صحت عملکرد و پایگاه داده مورداستفاده را نشان می‌دهد.در این مطالعه تلاش شده است که تشخیص بیماری دیابت بر روی‌داده‌های بومی جمع‌آوری شده توسط محققین طرح انجام شود. سپس با استفاده از یادگیری‌های مختلف به پیش بینی بیماری دیابت پرداخته شد. اگرچه شبکه عصبی ممکن است بتواند به تنهایی به‌دقت مطلوب در تشخیص بیماری دیابت دست پیدا کند، لیکن استفاده از یادگیری‌های متنوع باعث اطمینان، بیشتر از صحت عملکرد خواهد شد. عملکرد یادگیرهای مختلف در جدول ۳ نمایش داده شده است.

جدول ۲. مقایسه صحت عملکرد مطالعات مختلف

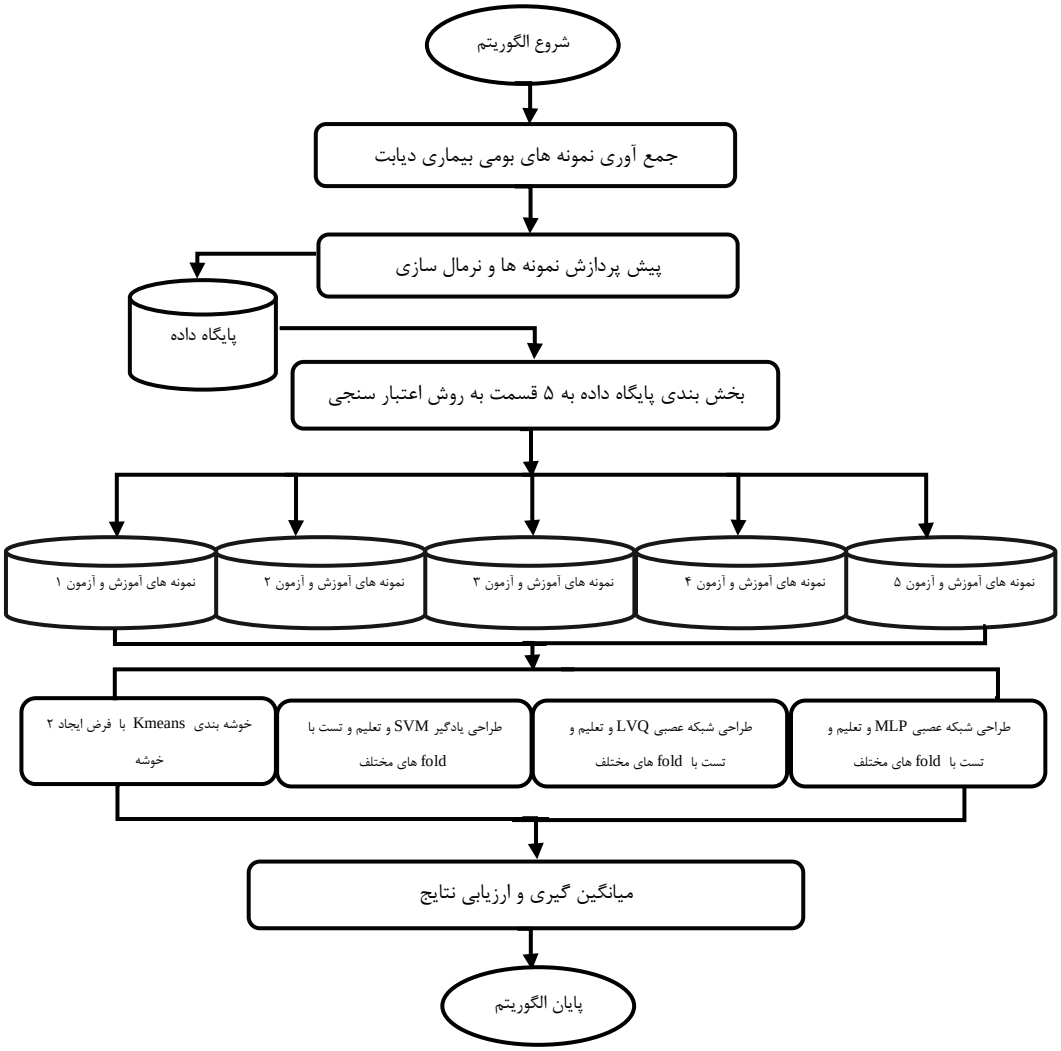
روش موردمطالعه	پایگاه داده مورد استفاده	صحت عملکرد	سال پژوهش
شبکه‌های فازی (۲۶)	۲۱۴	٪۷۹/۶۹	۲۰۱۱
الگوریتم تکاملی (۴)	۲۰۰	٪۷۶/۱۷	۲۰۱۳
k نزدیک‌ترین همسایه K_means (۱۰)	۱۹۷	٪۷۵/۵۵	۲۰۱۵
درخت تصمیم decision tree (۲۷)	۲۱۰	٪۷۸/۱۷	۲۰۱۱
رگرسیون لجستیک logistic regression (۹)	۲۷۴ بیمار دیابت نوع ۲	٪۸۳	۲۰۱۳
شبکه‌های عصبی پرسپترون MLP (۱۹)	۲۷۴ بیمار دیابت نوع ۲	٪۸۸	۲۰۱۳
شبکه عصبی شعاعی RBF (۲۳)	(PID) ۷۶۸	٪۷۳/۳۲	۲۰۰۵
بردار ماشین پشتیبان SVM (۱۰)	(PID) ۷۶۸	٪۸۲/۵	۲۰۱۶
مدل‌سازی درخت تصمیم c4.5 (۲۹)	(PID) ۷۶۸	٪۷۹	۲۰۱۸
شبکه عصبی پرسپترون (۲۰)	۱۰۰۰ نمونه کلینیک تخصصی	٪۹۰/۶	۲۰۱۷
بردار ماشین پشتیبان (پژوهش حاضر)	۲۵۴ نمونه	٪۹۶	۲۰۱۸

جدول ۳. مقایسه عملکرد روش‌های مختلف در این مطالعه (در هر روش میانگین یادگیرهای مختلف ثبت شده است)

شبکه عصبی MLP	شبکه عصبی LVQ	روش SVM	روش K_means
٪۹۴	٪۹۱	٪۹۶	٪۹۳

بحث

برای داشتن یک داده‌کاوی مؤثر علاوه بر نیاز به داده‌های مرتبط، باید از یک فرآیند و روش داده‌کاوی مناسب نیز بهره‌مند گردید. روشی که کلیه مراحل داده‌کاوی آن اعم از آماده‌سازی داده، مدل‌سازی و ارزیابی مبتنی بر داده های بومی باشد. تاکنون مطالعات بسیاری به منظور تشخیص بیماری دیابت انجام‌شده است. از این میان می‌توان به پژوهش سیتی (Siti) و همکاران در استفاده از شبکه های عصبی پایه شعاعی (Radial Basis Functions(RBF)) (۲۳) و شبکه‌های عصبی پرسپترون چند لایه (۲۴) اشاره کرد که به ترتیب به دقت‌های ٪۷۳/۳۲ و ٪۷۶/۸۹ دست یافتند. Breault و همکاران در سال ۲۰۰۸ روش Classification and Regression Trees (CART) را با هدف بررسی عوارض میکروواسکولار دیابت مورد مطالعه قرار دادند (۲۵). Fang با روش خوشه بندی بدون نظارت شده بیماران دیابتی را بررسی کرد (۲۶). Jarullah با استفاده از الگوریتم درخت تصمیم J48 بر روی داده‌ها در نرم‌افزار وکا (weka) به کلاس بندی بیمارن دیابتی پرداخت (۲۷). Antonelli و همکاران در سال ۲۰۱۳ یک چارچوب تجزیه و تحلیل مبتنی بر خوشه‌بندی چند سطحی برای شناسایی مسیرهای درمان و معاینه بیماران دیابتی ارائه کردند. پایگاه داده مورد مطالعه شامل ۸۵۶ رکورد از بیماران یکی از بیمارستان‌های انگلستان بوده و روش پیشنهادی در شناسایی گروه‌های بیماران با تاریخچه بیماری مشابه و افزایش شدت عوارض آن‌ها عملکرد خوبی داشته است (۲۸).



شکل ۶. الگوریتم ارائه شده در این پژوهش

صباغ گل و همکاران با استفاده از الگوریتم درخت تصمیم C4.5 به تشخیص بیماری دیابت با استفاده از چربی خون پرداختند (۲۹). برفه فرزانه (۱۳۹٤) با مطالعه بر روی ۱۳٤۲۳ نفر از شرکت‌کنندگان بالای ۲۵ سال که هیچ‌کدام دیابت کنترل‌شده‌ای نداشتند و با استفاده از مدل شبکه عصبی مصنوعی سه لایه مدلی ارائه دادند به دقت سطح زیر منحنی راک ۷/۷۲٪ و صحت پیش‌بینی آموزش ۹۲٪ و صحت پیش‌بینی آزمون ۹۱/٦٪ دست یافت. آنان نتیجه گرفتند که با توجه به عدم نیاز مدل شبکه عصبی مصنوعی به پیش فرض‌های معمول روش‌های کلاسیک آماری و صحت پیش‌بینی، شبکه‌های عصبی از آماری بالاتر است (۳۰). نکته قابل تأمل در پژوهش حاضر استفاده از متدهای مختلف داده کاوی بر روی داده های بومی است. این داده ها در چندین پارامتر متفاوت هستند و به عبارتی ویژگی های اضافه تری دارند که قابلیت تقریب زنی را نسبت به مطالعات مشابه افزایش داده و همان‌طور که در جدول ۳ مشخص‌شده است با متدهای مختلف داده‌کاوی تمامی دقت‌ها به بالای ۹۰٪ رسیده است درحالی‌که استفاده از همین روش ها در پایگاه دادهای مشابه دقت‌های کمتری را به همراه داشته است.

نتیجه‌گیری

بیماری دیابت درمان قطعی ندارد، بنابراین از یک‌سو 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. به همین منظور، این مطالعه با هدف بررسی روش های مختلف داده‌کاوی، در تشخیص

افراد مبتلا به دیابت از افراد سالم انجام شد. پژوهش‌های مبتنی در داده‌کاوی می‌توانند 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. ازآنجایی‌که هیچ الگوریتم طبقه‌بندی وجود ندارد که به تنهایی به‌طور کامل برای تمام مسائل مناسب باشد، استفاده از طبقه بندهای مختلف می‌تواند منجر به اطمینان از صحت طبقه‌بندی شود. درنهایت می‌توان از آن‌ها به‌عنوان دستیار پزشک در تشخیص بیماری‌ها جهت پیش‌گویی امکان ابتلای افراد به بیماری‌ها استفاده کرد.

تضاد منافع

در این پژوهش هیچ گونه تعارض منافعی توسط نویسندگان گزارش نشده است.

References

1. Nazarzadeh M, Bidel Z, Sanjari Moghaddam A. Meta-analysis of diabetes mellitus and risk of hip fractures small study effect. Osteoporos Int (2016) 27: 229.

2. 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. Research in pharmaceutical sciences. 2015; 10(2):125.

3. Haddadnia J, Vahidi J, Gharakhani A, Fouzi A. Fuzzy Diagnosis of Diabetes Mellitus Based on the Comprehension Rules and Characteristics Based on Combination of Data Mining Systems and Intelligent Intelligent Algorithms, International Conference on Nonlinear Modeling and Optimization, 2011. [Persian]

4. Habibi M, Ahmadifard A. Feature Selection Using Taboo Search, Genetic Algorithm and KNN for Diagnosis of Diabetes.12th Iranian Conference on Intelligent Systems, 2013. [Persian]

5. 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.

6. Saiti, F, Naini A, Shoorehdeli A, Teshnehlab M.A. Thyroid disease diagnosis based on genetic algorithms using PNN and SVM. In Bioinformatics and Biomedical Engineering, 2009. ICBBE 2009. 3rd International Conference on (pp. 1-4). IEEE.

7. 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. 1;39(19):3372-9.

8. 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.

9. 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.

10. 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.

11. 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).

12. Elsappagh S, Elmogy M, Riad AM. A fuzzy ontology oriented case based reasoning framework for semantic diabetes diagnosis. Artif Intell Med2015;14:92-5.

13. Nazari M, Zamani Dehkordi M, Kiumarsi Dehkordi M. Evaluation of Diagnosis of Diabetes Mellitus based on ECG Signal Information Using Artificial Neural Networks, Journal of Shahrekord University of Medical Sciences, 2012.4(19): 64-77.

14. 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.

15. Blake C, Merz C. Repository of machine learning databases, Irvine, CA: University of California, Department of Information and Computer Science. 1998.

16. Burfei F, Salehi M, Najafi I. Anticipating Diabetes Using an Artificial Neural Network. Razi Medical Journal. 2015; 22 (135): 29-37.

17. Si J, Zhang Y, Hu S, Sun L, Li S, Yang H, et al., editors. 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.

18. Jia`ng, M. Jiang, L. Jiang, D. Xiong, J. Shen, J. Ahmed, S.H. 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, pp. 250-256

19. 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.

20. 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

21. 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

22. 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

23. Jaafar SF, Ali DM. Diabetes mellitus forecast using artificial neural network (ANN). Asian Conference on Sensors and the International Conference on New Techniques in Pharmaceutical and Biomedical Research 2005; 5, 135-139.

24. 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.

25. Breault JL, Goodall CR, Fos PJ. Data mining a diabetic data warehouse. Artificial intelligence in medicine. 2002 Sep 1;26(1-2):37-54.

26. Fang X. Are you becoming a diabetic a data mining approach? In Fuzzy Systems and Knowledge Discovery. FSKD'09. Sixth International Conference on. 2009 ;(5) 18-22. IEEE.

27. Al Jarullah, Asma A. Decision tree discovery for the diagnosis of type II diabetes. International Conference on Innovations in Information Technology (IIT), 2011; pp. 303-307. IEEE.

28. Antonelli D. Baralis E. Bruno G.Cerquitelli T.Chiusano S. & Mahoto N. Analysis of diabetic patients through their examination history. Expert Systems with Applications, 2013; 40(11), 4672–4678.

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 Farzaneh, Salehi Masoud, Najafi Iraj. Anticipating Diabetes Using an Artificial Neural Network. Razi Medical Journal. 2015; 22 (135): 29-37.

Journal of Torbat Heydariyeh University of Medical Sciences
2019, Volume 7, No.1: 1-14.

Original article

Diagnosis of diabetes by using a data mining method based on native data

Iman Abedian[1*], Ali Ayoobi[1], Hamidreza Ghaffary[1], Iman Zabbah[2]

1. Department of Computer Engineering, Islamic Azad University, Ferdows Branch,Ferdows, Iran

2. Department of Computer Egineering, Islamic Azad University, Torbat Heydariyeh Branch, Torbat Heydariyeh, Iran

Corresponding author: iman.abedian@gmail.com

Keywords:

Diabetes mellitus,

Artificial neural network,

Support vector Machine,

Clustering

Abstract

Background & Aim*:* Detecting the abnormal performance of diabetes and subsequently getting proper treatment can reduce the mortality associated with the disease. Also, timely diagnosis will result in irreversible complications for the patient. The aim of this study was to determine the status of diabetes mellitus using data mining techniques.

Methods**:** This is an analytical study and its database contains 254 independent records based on 13 characteristics. Data is collected by a researcher from one of the specialized diabetes centers in Mashhad.

Results**:** After preprocessing of the obtained data, different methods of pattern recognition were applied. Using multilevel MLP neural networks, LVQ neural networks, SVM support vector and Kmeans clustering method, the mean square error (RMSE) was calculated. The correctness of each learner's performance is 94%, 92%, 96%, and 93%, respectively.

Conclusion**:** Reducing the diagnosis of diabetes is one of the goals of the researchers. Using data mining techniques can help to reduce this error. In this study, among different protocols used for pattern recognition, SMV method displayed a significantly better performance.

How to Cite this Article: Abedian I, Ayoobi A, Ghaffary H, Zabbah I. Diagnosis of diabetes by using a data mining method based on native data. Journal of Torbat Heydariyeh University of Medical Sciences. 2019;7(1):1-14.