مقاله پژوهشی

سیستم پشتیبانی تصمیم در تشخیص سندرم متابولیک با بهره‌گیری از راه‌کار داده‌کاوی

نویسندگان:
مهدی ادریسی[1]، مژگان قاری‌پور[2]، آزاده فاروقی[3]، فاطمه جاوری[3]، بهروز شاهقلی[3]، امین قاری‌پور[4]، نضال صراف‌زادگان[*5]

۱- بخش مهندسی پزشکی، دانشکده فنی و مهندسی، دانشگاه اصفهان، اصفهان، ایران
۲- مرکز تحقیقات بازتوانی قلبی، پژوهشکده قلب و عروق اصفهان، دانشگاه علوم پزشکی اصفهان، اصفهان، ایران
۳- بخش مهندسی فناوری اطلاعات، دانشکده فنی و مهندسی، دانشگاه اصفهان، اصفهان، ایران
۴- خانه ریاضیات اصفهان، دانشگاه صنعتی اصفهان، اصفهان، ایران
۵- مرکز تحقیقات قلب و عروق، پژوهشکده قلب و عروق اصفهان، دانشگاه علوم پزشکی اصفهان، اصفهان، ایران

فصلنامه دانشگاه علوم پزشکی جهرم، دوره نهم، ویژه‌نامه قلب و عروق، ۱۳۹۰

چکیده:

مقدمه: با توجه به اهمیت نقش سندرم متابولیک در کنترل بیماری‌های قلبی – عروقی و دیابت، در مطالعه حاضر به کاوش مهم‌ترین ویژگی‌های موثر در ابتلا به سندرم متابولیک و بررسی ویژگی‌های افراد مبتلا به این سندرم با روش داده کاوی پرداخته شد. در این مطالعه همچنین مدل‌های داده‌کاوی به کار گرفته شده، ارزیابی شدند.

روش کار: با به‌کارگیری اطلاعات مرحله سوم برنامه قلب سالم اصفهان در سال ۱۳۸۶، تعداد ۹۵۷۲ نفر مورد بررسی قرار گرفتند. کارایی سه مدل اصلی داده کاوی شامل مدل درخت تصمیم، مدل مبتنی بر قاعده بیز و مدل شبکه عصبی مقایسه و بهترین مدل در تشخیص سندرم متابولیک انتخاب شد.

یافته ها: از میان ویژگی‌های مورد بررسی در هر سه مدل، ویژگی شاخص توده بدنی به عنوان مهم‌ترین عامل ابتلا به سندرم متابولیک تشخیص داده شد. با توجه به نتایج مدل درخت تصمیم مشخص شد که ۶۷/۲ درصد از افراد گروه سنی بین ۵۰ سال تا ۷۰ سال با شاخص توده بدنی بزرگ‌تر از ۳۰ مبتلا به سندرم متابولیک هستند.

نتیجه گیری: نتایج این مطالعه نشان داد که مدل شبکه عصبی از دقت بالاتری (۰/۸۲) نسبت به دو مدل درخت تصمیم (۰/۸۱) و مدل مبتنی بر قاعده بیز (۰/۷۷) برخوردار است. اما مدل مبتنی بر قاعده‌ی بیز در تشخیص مبتلایان به سندرم متابولیک پیش‌بینی بهتری ارائه می‌دهد. مناسب بودن داده‌ها، به‌کارگیری پیش‌پردازش مناسب و راه کار مناسب داده‌کاوی منجر به کسب نتایج بهتر در ارتباط با داده‌های پزشکی می‌شود.

واژگان کلیدی: داده‌کاوی، درخت تصمیم، قضیه بیز، شبکه‌های عصبی، بیماری قلبی– عروقی

مقدمه:

بهره‌گیری از راه‌کار داده کاوی با رویکرد استخراج دانش از اطلاعات موجود در راستای اهدافی چون نحوه تشخیص بیماری، کاهش هزینه های درمانی و میزان خطای نحوه درمان موثر بوده و موجب بهبود عملکرد سازمان‌های بهداشتی می‌شود. بیماری‌های قلبی– عروقی از شایع‌ترین بیماری‌ها در ایران است که نه تنها شیوع خود بیماری بلکه شیوع عوامل خطرساز آن نیز رو به افزایش است [۱]. مطالعات مختلف نشان داده‌اند که وجود سندرم متابولیک تأثیر به سزایی روی فراوانی مرگ و میر ناشی از بیماری‌های قلبی– عروقی دارد. سندرم متابولیک به عنوان مجموعه‌ای از اختلالات متابولیکی شامل پرفشاری خون، چاقی، اختلال لیپیدها و افزایش مقاومت به انسولین از جمله عوامل تأثیرگذار روی بروز مرگ و میر و ناتوانی ناشی از بیماری‌های قلبی– عروقی تلقی می‌شود [۲]. مطالعات مختلف نشان داده‌اند که شیوع این سندرم نه فقط در کشورهای غربی بلکه در کشورهای آسیایی نیز بالا است. فراوانی این بیماری در ایران ۲۳/۳ درصد می‌باشد [۳]. مطالعات بسیاری در خصوص ارتباط این سندرم با وجود عوامل خطر ساز بیماری‌های قلبی– عروقی

* نویسنده مسئول، آدرس: اصفهان، میدان جمهوری اسلامی، خیابان خرم، مجتمع مراکز درمانی تحقیقاتی حضرت صدیقه طاهره (س)، مرکز تحقیقات قلب و عروق، پژوهشکده قلب و عروق اصفهان، صندوق پستی: ۱۱۴۸– ۸۱۴۶۵
تلفن: ۳۳۵۹۶۹۶ و ۳۳۵۹۷۹۷–۰۳۱۱ نمابر: ۳۳۷۳۴۳۵–۰۳۱۱ پست الکترونیک: nsarrafzadegan@gmail.com

تاریخ دریافت: ۱۳۹۰/۰۷/۰۴ تاریخ پذیرش: ۱۳۹۰/۰۹/۱۳

از جمله دیابت، پر فشاری خون، اختلالات لیپیدی، چاقی، بی تحرکی و سیگار انجام شده است و در برخی از موارد به نظر می‌رسد این اختلالات از دوران جنینی آغاز شده و عامل ژنتیک در بروز آن موثر باشد [۴]. اگرچه وجود سندرم متابولیک با افزایش احتمال مرگ و میر ناشی از بیماری‌های قلبی- عروقی ارتباط دارد اما به نظر می‌رسد که این احتمال خطر، مستقل از سایر عوامل خطر ساز مثل سن، سطح LDL کلسترول سرمی و استعمال دخانیات عمل می‌کند [۵ و ۶]. در این راستا در مطالعه حاضر کوشش شده است تا با به‌کارگیری راه کارهای داده‌کاوی، بر اساس علائم اولیه و نتایج آزمایشات ساده و زود پاسخ ده، احتمال ابتلای فرد به سندرم متابولیک از اولین لحظات مراجعه‌ی فرد به پزشک تا قبل از آماده شدن نتایج آزمایشات نهایی برآورد شود و بدین ترتیب دانش پزشک برای اتخاذ تمهیدات ضروری ارتقا یابد.

روش کار:

در این تحقیق از راه کارها داده‌کاوی به منظور تشخیص سندرم متابولیک و عوامل موثر روی آن استفاده شده است. با اعمال مدل‌های داده‌کاوی روی تعداد زیادی از داده‌های افراد مختلف، الگوهایی برای تشخیص سندرم متابولیک به دست آمده است.

مدل‌های مورد بررسی

با توجه به لزوم بررسی و بهره گیری از روش‌های گوناگون در فرایند اخذ تصمیم و تشخیص بیماری، در این پژوهش توانایی مدل درخت تصمیم، مدل مبتنی بر قاعده‌ی بیز و مدل شبکه‌ی عصبی پیش رونده ارزیابی و شرح مختصری از هر یک بیان می‌شود.

مدل درخت تصمیم‌گیری

با توجه به این که درخت تصمیم‌گیری در حین فرایند یادگیری ساخته و از آن برای تحلیل و پیش بینی داده های جدید استفاده می‌شود، یکی از معروف‌ترین و ساده‌ترین روش‌های قابل فهم در داده‌کاوی است [۷]. در این مدل، داده‌ها بر اساس مقادیر متغیرها جداسازی می‌شوند. مزیت این مدل سرعت بالاتر و قابل فهم بودن آن می‌باشد [۸].

مدل مبتنی بر قاعده‌ی بیز

استدلال بیزی یکی از ابزارهای داده‌کاوی است که در طبقه‌بندی، دسته‌بندی و رگرسیون استفاده می‌شود. مدل بر اساس این فرض بنا شده است که مقادیر مورد توجه از یک توزیع احتمال پیروی می‌کنند و تصمیم‌های بهینه می‌توانند با استدلال روی این احتمالات به همراه داده‌های مشاهده شده اتخاذ شوند [۹].

مدل شبکه‌ی عصبی

شبکه‌های عصبی، مجموعه ای از گره های مرتبط با ورودی‌ها، خروجی‌ها و عملکرد پردازش در هر گره می‌باشند. تعدادی از لایه‌های مخفی پردازش مابین لایه‌های ورودی و خروجی قابل مشاهده، وجود دارند. مدل عصبی باید شبکه‌ای را روی یک مجموعه داده‌های در حال آزمایش، مورد بررسی و امتحان قرار داده و از آن برای انجام پیش‌بینی‌های مختلف استفاده کند [۱۰].

مجموعه داده

اطلاعات استفاده شده در این پژوهش از مجموعه داده‌های مربوط به مرحله سوم طرح تحقیقاتی برنامه‌ی قلب سالم اصفهان در سه شهرستان اصفهان، نجف آباد و اراک در سال ۱۳۸۶ روی ۹۵۷۲ نفر که به طور تصادفی از سطح جامعه انتخاب شده اند انجام گرفت. داده‌ها ابتدا به وسیله پایگاه داده‌های نرم‌افزار Microsoft Access جمع‌آوری شد و بعد از بررسی توسط متخصصان به نرم‌افزار SPSS انتقال یافت. طبق مطالعات انجام شده و نظر متخصصین، ۱۹ ویژگی شامل شاخص توده بدنی، جنس، سن، تحصیلات، میزان فعالیت فیزیکی، نحوه مقابله با استرس، سابقه‌ی فامیلی دیابت، سابقه فامیلی فشارخون، سابقه فامیلی‌چربی‌خون، سابقه فامیلی سکته قلبی، سابقه فامیلی سکته مغزی، ضربان قلب، سابقه سکته مغزی، در آمد، شغل، کلسترول (LDL) و سابقه مصرف سیگار برای داده‌کاوی انتخاب و تأثیر این ویژگی‌ها روی ویژگی هدف یعنی تشخیص بیماران مبتلا به اختلال سندرم متابولیک بررسی شد.

یافته‌ها:

به منظور نشان دادن کارایی هر یک از روش‌های داده‌کاوی، نتایج بدست آمده در قالب چهار هدف به شرح زیر توصیف شده‌اند. نتایج نشان می‌دهند که هر سه مدل درخت تصمیم، مدل مبتنی بر قاعده بیز و مدل شبکه عصبی اهدافی را بیان می‌کنند و می‌توانند در تصمیم‌گیری پزشک، تشخیص بیماری و یافتن داروی مرتبط با اختلال سندرم متابولیک موثر واقع شوند. در این مطالعه برای ارزیابی کارایی مدل‌ها از دو روش ارزیابی اهداف داده کاوی و بررسی میزان کارایی استفاده شده است.

۱- ارزیابی اهداف داده‌کاوی

این چهار هدف به شرح زیر می‌باشند:

۱-۱- پیش‌بینی اختلال سندرم متابولیک

هدف این است که با دادن مشخصات بالینی بیمار، پیش‌بینی شود که وی مبتلا به اختلال سندرم متابولیک می‌باشد یا خیر. این هدف به وسیله هر سه مدل برآورده شده است. برای مثال با در نظر گرفتن ویژگی‌های جنس = زن و سن = بیش از ۷۰ سال و تحصیلات =۰ تا ۵ سال، احتمال ابتلا بیمار به سندرم متابولیک بر اساس مدل درخت تصمیم ۵۱/۴۶ درصد و برای مدل مبتنی بر قاعده‌ی بیز و شبکه عصبی پیش خور به ترتیب برابر با ۹۹/۲۶ درصد و ۴۱/۵۵ درصد بدست آمده است. با توجه به بالا بودن مقادیر احتمال، پزشک می‌تواند برای این چنین بیماران انجام آزمایشات بیش‌تری را توصیه کند و بدین ترتیب از بروز این بیماری جلوگیری کرد.

۱-۲- تحلیل ویژگی‌های موثر

هدف، تحلیل روابط و تشخیص ویژگی‌های موثر روی ویژگی هدف (اختلال سندرم متابولیک) می‌باشد. در دو مدل درخت تصمیم و مدل مبتنی بر قاعده بیز ویژگی شاخص توده بدنی دارای بیش‌ترین تأثیر می‌باشد. ویژگی های مهم دیگر به ترتیب کلسترول (LDL)، سن، تحصیلات، شغل، جنس، فعالیت بدنی، سابقه فامیلی دیابت، سابقه فامیلی فشار خون، سابقه فامیلی سکته مغزی، سابقه سکته مغزی، درآمد، مصرف سیگار، سابقه فامیلی چربی خون، سابقه فامیلی سکته قلبی، ضربان قلب می‌باشند. مدل مبتنی بر قاعده‌ی بیز اطلاعات معنادارتری از همه‌ی ویژگی‌ها در مقایسه با مدل درخت تصمیم ارائه می‌دهد. این اطلاعات در تحلیل ویژگی‌هایی که بیش‌ترین و یا کم‌ترین تأثیرات را در ارتباط با اختلال سندرم متابولیک دارند توسط پزشک قابل استفاده می‌باشند.

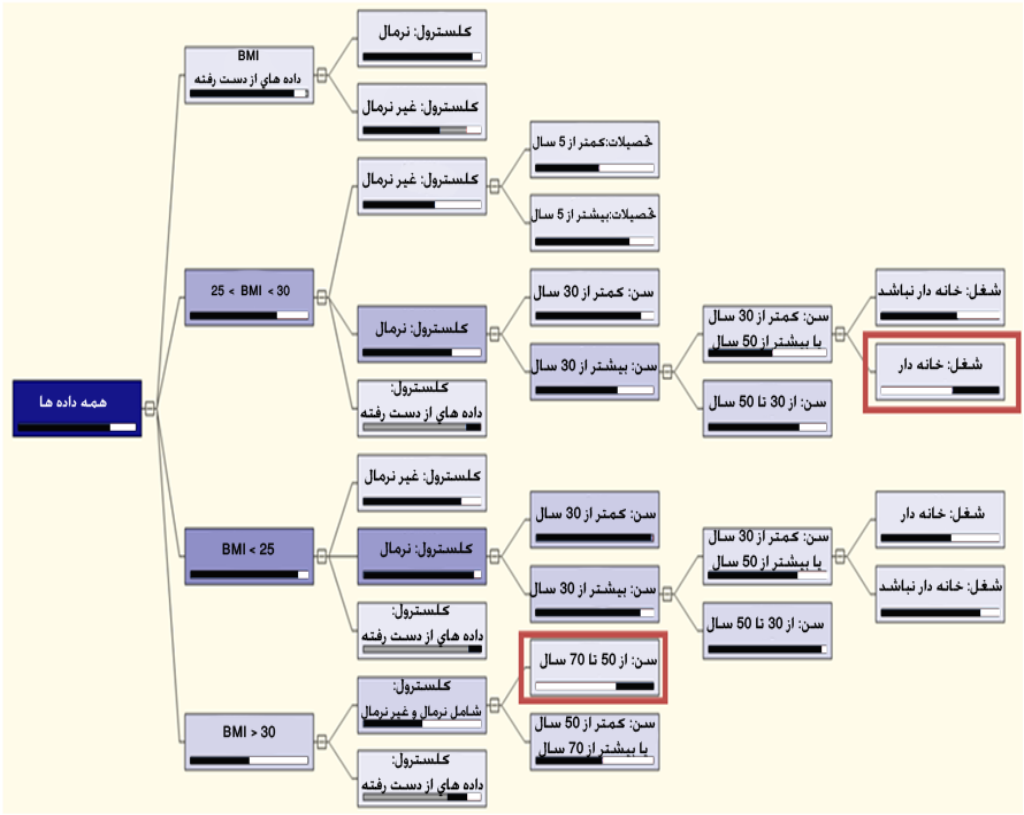
۱-۳- مشخص کردن تأثیر و رابطه‌ی ویژگی‌ها

هدف، مشخص کردن تأثیر و رابطه‌ی ویژگی‌های ورودی نسبت به ویژگی هدف (اختلال سندرم متابولیک) است. این هدف فقط توسط مدل درخت تصمیم برآورده شده است. درخت کلی حاصل شده با تمامی مقادیر در شکل ۱ نشان داده شده است. تیرگی پس‌زمینه در هر گره بیانگر تعداد افراد مورد بررسی در آن گره و نوار رنگی در هر گره بیانگر درصد حالات ویژگی خروجی می‌باشد. نوار رنگی سفید نشان‌دهنده درصد فراوانی افراد مبتلا به سندرم متابولیک و نوار رنگی مشکی بیانگر درصد فراوانی افراد غیر مبتلا به سندرم متابولیک است. با توجه به رنگ پس‌زمینه و نوارهای رنگی گره‌های درخت شکل ۱، دو گره برگ با برچسب «شغل: خانه‌دار» و گره‌برگ با برچسب «سن: ۵۰ تا ۷۰ سال» دارای بیش‌ترین فراوانی مبتلایان به سندرم متابولیک هستند.

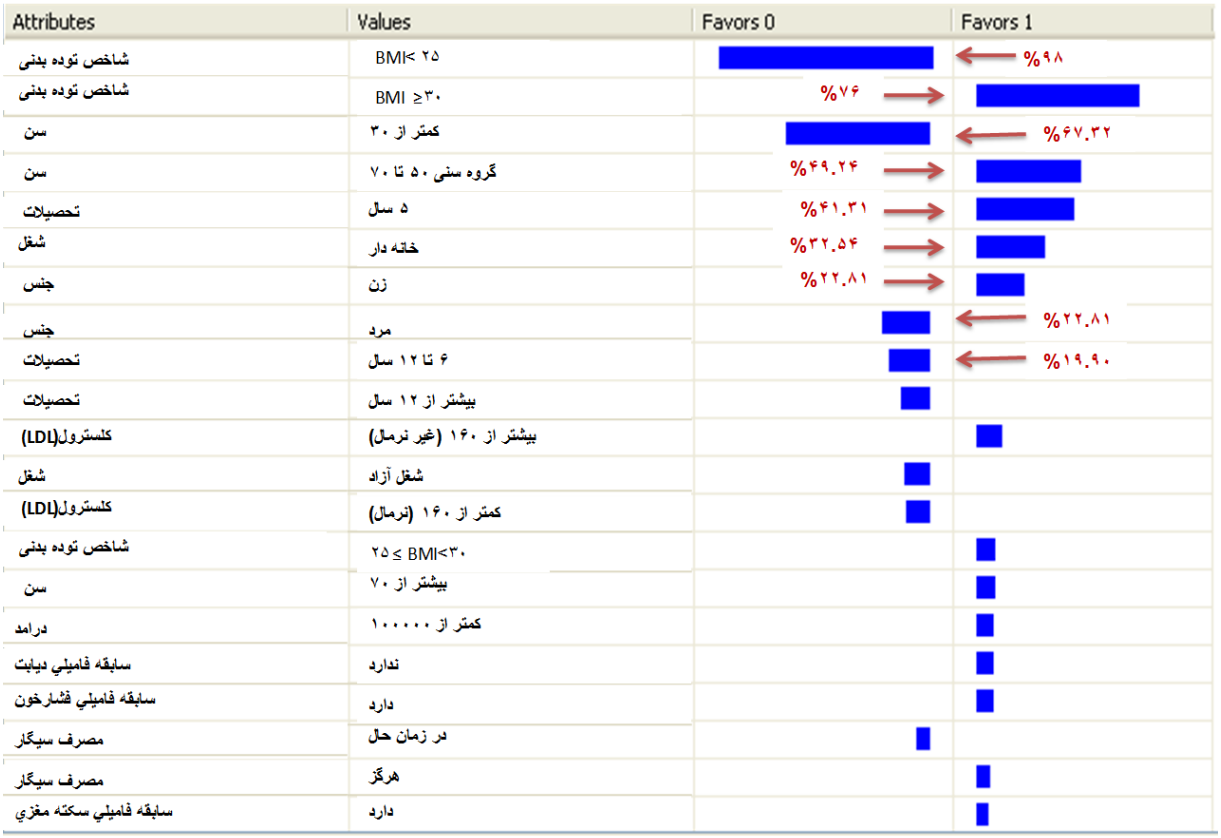
۱-۴- تعیین میزان وابستگی مقادیر ویژگی‌ها

هدف، تعیین میزان وابستگی مقادیر ویژگی‌های ورودی به متغیر هدف در دو حالت ۱- مبتلا بودن به اختلال سندرم متابولیک ۲- مبتلا نبودن به اختلال سندرم متابولیک می‌باشد. در نمودار مدل مبتنی بر قاعده‌ی بیز شکل ۲، ویژگی‌های با بیش‌ترین تأثیر روی ویژگی هدف به ترتیب نزولی نشان داده شده است.

با توجه به نتایج به دست آمده از این نمودار، افراد با ویژگی وزن نرمال با احتمال ۸۵ درصد، افراد با ویژگی سن کم‌تر از ۳۰ سال با احتمال ۶۷/۳۲ درصد و مردها با احتمال ۸۱/۲۲ درصد به اختلال سندرم متابولیک مبتلا نمی‌شوند. اما افراد چاق با احتمال ۷۶ درصد و افراد در گروه سنی ۵۰ سال تا ۷۰ سال با احتمال ۴۹/۲۴ درصد و افراد با سطح تحصیلات کم با احتمال ۴۱/۳۱ درصد به اختلال سندرم متابولیک مبتلا می‌شوند.

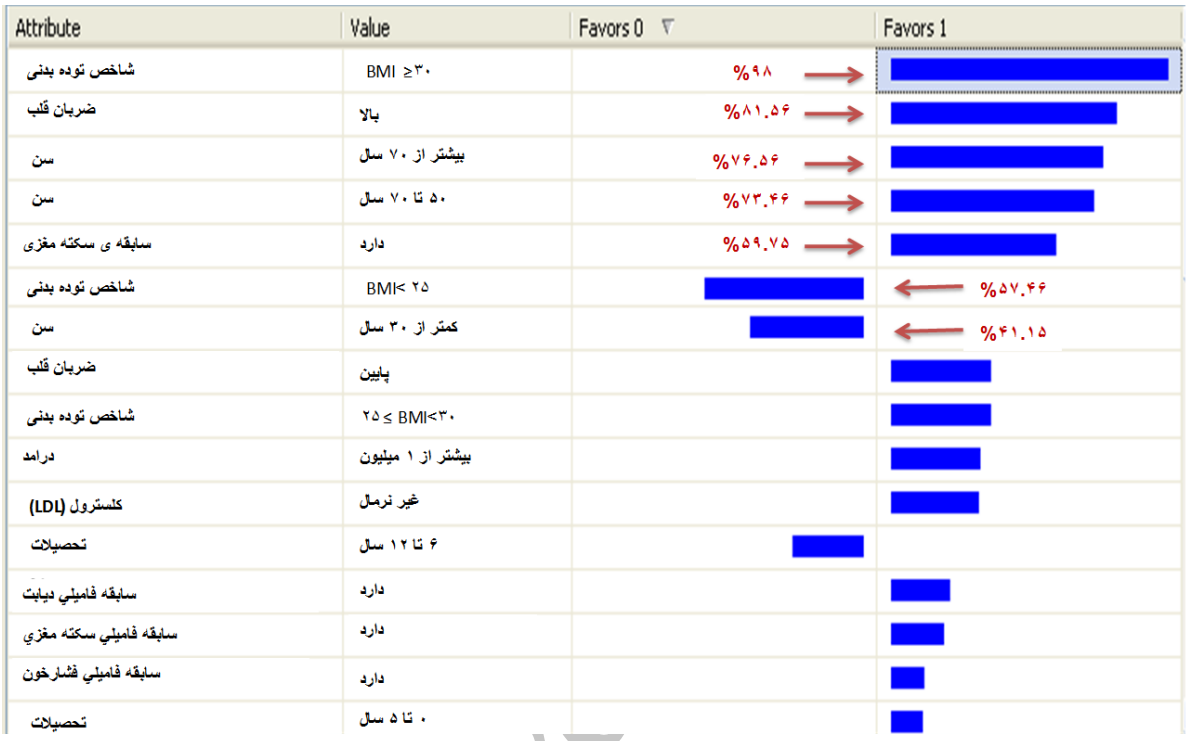


شکل ۱: نمایش درخت تصمیم‌گیری- استفاده از شرح نمودار داده کاوی برای کشف درصد احتمال وقوع سندرم متابولیک



شکل ۲: نمایش تمایز ویژگی‌ها (Attribute Discrimination) مدل مبتنی بر قاعده بیز

همچنین در نمودار مدل شبکه عصبی پیش‌خور شکل ۳، سیر نزولی تأثیر ویژگی‌ها با ارزش‌های متفاوت روی ویژگی هدف مشخص شده است. اولین ویژگی با بیش‌ترین تأثیر روی ابتلا به سندرم متابولیک، ویژگی شاخص توده بدنی بیش‌تر از ۳۰ می‌باشد. با توجه به شکل ۳، ویژگی‌های شاخص توده بدنی بیش‌تر از ۳۰ و ضربان قلب غیر نرمال و سن بالاتر از ۵۰ سال، به ترتیب دارای بیش‌ترین اثر روی سندرم متابولیک هستند. ویژگی‌های وزن نرمال و سن کم‌تر از ۳۰ سال و سطح تحصیلات بالا، ویژگی هدف را با مقدار صفر (عدم مبتلا به سندرم متابولیک) پیش‌بینی می‌کنند.



شکل ۳: نمایش تمایز ویژگی‌ها مدل شبکه عصبی

۲- بررسی میزان کارایی مدل‌ها

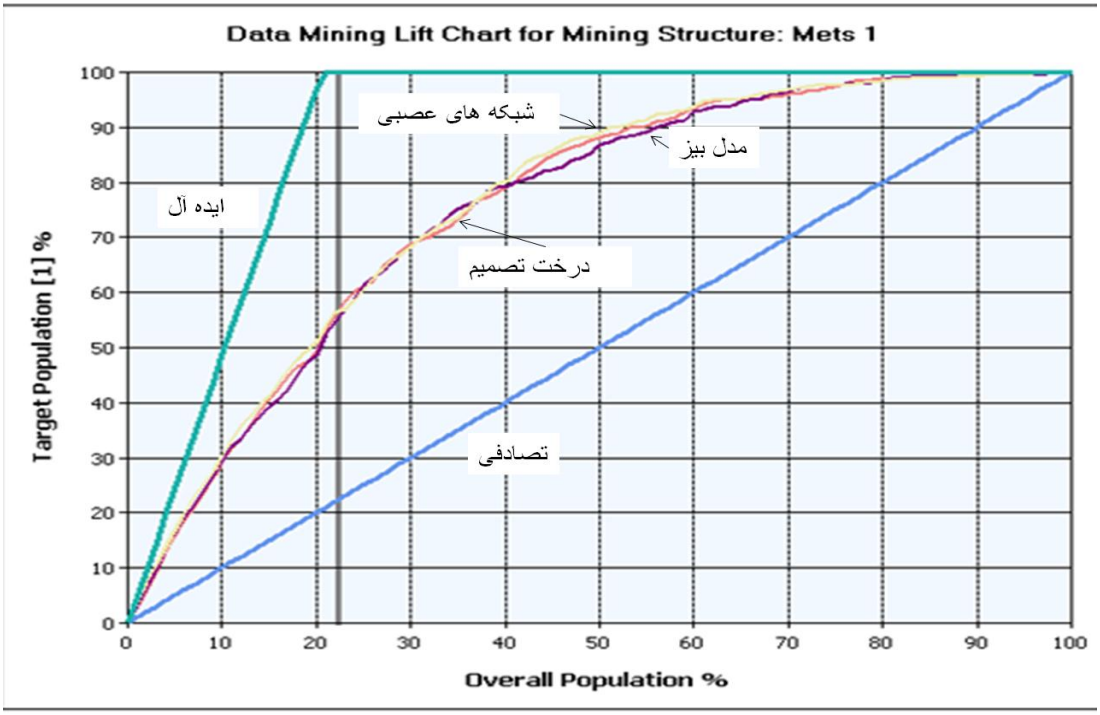
کارایی مدل‌ها با استفاده از دو روش نمودار Lift و ماتریس طبقه بندی ارزیابی شده است. هدف، مشخص کردن مدلی با بالاترین دقت و بالاترین درصد پیش‌بینی درست در تشخیص افراد مبتلا به اختلال سندرم متابولیک می‌باشد.

۲-۱- نمودار Lift

این نمودار برای نشان دادن عملکرد و دقت پیش‌بینی مدل‌های داده‌کاوی کاربرد دارد. برای این کار دو نوع نمودار وجود دارد. در نوع اول، متغیر هدف مشخص می‌شود ولی در نوع دوم، بررسی بدون مشخص کردن متغیر هدف انجام می‌شود.

۲-۱-۱- نمودار Lift با مشخص‌بودن مقدار متغیر مورد پیش‌بینی در شکل۴ خروجی نمودار درحالت مشخص‌بودن مقدار متغیر هدف یعنی افراد مبتلا به سندرم متابولیک (۱= Metabolic syndrome) نشان داده شده است. محور افقی نمودار، بیانگر درصد مجموعه داده‌های آزمایشی و محور عمودی نشان‌دهنده‌ی درصد مقادیری است که با موقعیت مشخص شده از ویژگی هدف (۱=Metabolic syndrome) پیش‌گویی شده‌اند [۱۱]. این نمودار حاوی یک منحنی برای هر مدل انتخاب شده، به اضافه یک منحنی ایده‌آل و یک منحنی تصادفی است. با توجه به این که در ۲۱ درصد از داده‌های آزمون، افراد مبتلا به سندرم متابولیک هستند (۱=Metabolic syndrome)، پس در دنیای واقعی باید ۲۱ درصد از داده‌های آزمون برای پیدا کردن همه‌ی بیماران مبتلا به سندرم متابولیک بررسی شوند [۱۱]. به همین دلیل، منحنی ایده‌آل ۱۰۰ درصد از محور عمودی روی ۲۱ درصد از محور افقی است. خط مستقیم مشخص کننده حالت تصادفی، نشان‌دهنده‌ی این است که اگر به صورت تصادفی نتیجه‌ی هر مورد حدس زده شود، ۵۰ درصد از جمعیت هدف، از ۵۰ درصد داده‌های آزمون به دست می‌آید. هر سه منحنی مربوط به مدل‌ها که بین منحنی‌های تصادفی و ایده‌آل می‌باشند، اطلاعات معناداری از الگوهای یادگیری در ارتباط با حالت مشخص شده از متغیر پیش‌بینی را به دست می‌دهند. در نمودار شکل ۴، ۲۲/۳۳ درصد از جمعیت هدف مورد بررسی قرار

گرفته است. با توجه به جدول ۱ مدل شبکه عصبی پیش خور، ۵۶/۲۵ درصد از جمعیت هدف (افراد مبتلا به سندرم متابولیک)، مدل درخت تصمیم ۵۵/۵۷ درصد از جمعیت هدف و مدل مبتنی بر قاعده‌ی بیز ۵۴/۵۶ درصد از جمعیت هدف را یافته اند. هر مدلی که با بررسی درصد کم‌تری از داده‌ها، بیش‌ترین درصد از جمعیت هدف را بدست آورد، مدل بهتر تلقی می‌شود. اما در مقایسه‌ی مدل‌ها علاوه بر درصد جمعیت هدف، باید احتمال پیش‌بینی هر یک از مدل‌ها هم در نظر گرفته شود. با توجه به جدول ۱ در مدل شبکه عصبی پیش خور که ۵۶/۲۵ درصد از جمعیت هدف را بدست آورده، با احتمال ۳۴ درصد ممکن است افرادی غیرمبتلا به سندرم متابولیک در این جمعیت وجود داشته باشند. برای مقایسه‌ی مدل‌ها بهتر است از پارامتر دقت (Score) که در شرح نمودار داده‌کاوی (Mining Legend) وجود دارد و میزان دقت مدل‌ها را نشان می‌دهد استفاده کرد. با توجه به نتایج جدول ۱، شبکه عصبی پیش‌خور با دقت ۰/۸۵ دقیق‌تر از دو مدل دیگر (با دقت ۰/۸۴) است.



شکل ۲: نمودار Lift با مشخص کردن مقدار متغیر مورد پیش‌بینی

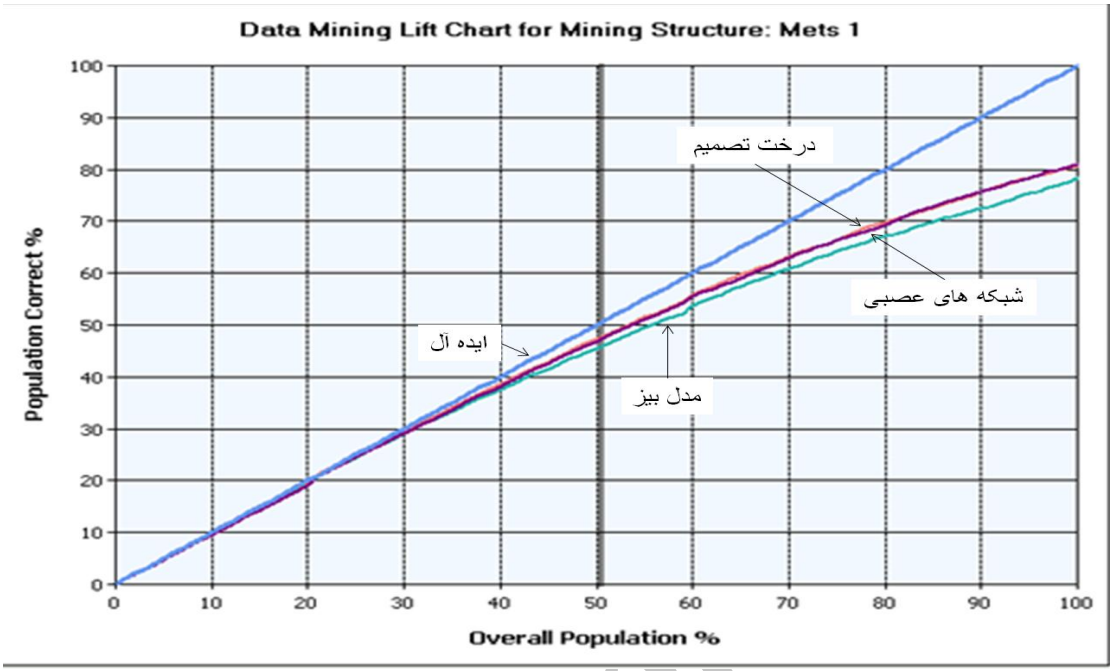
جدول ۱: شرح نمودار داده‌کاوی (Mining Legend) با در نظر گرفتن ٪۲۲/۳۳ جمعیت

مدل	دقت	احتمال پیش بینی	جمعیت هدف
مدل درخت تصمیم	٪۰/۸۴	٪۴۴/۱۳	٪۵۵/۵۷
مدل مبتنی بر قاعده بیز	٪۰/۸۴	٪۵۵/۶۳	٪۵۴/۵۶
مدل شبکه عصبی	٪۰/۸۵	٪۳۴/۰۱	٪۵۶/۲۵
مدل تصادفی			٪۲۲/۰۰
مدل ایده آل			٪۱۰۰/۰۰

۲-۱-۲- نمودار Lift بدون مشخص کردن مقدار متغیر مورد پیش‌بینی

این نمودار هم مانند نمودار قبلی است با این تفاوت که محور عمودی نشان‌دهنده‌ی درصد درستی پیش‌بینی متغیر هدف است. محور افقی بیانگر درصد داده‌های آزمون مورد بررسی است. خط ایده‌آل زاویه‌ی ۴۵ درجه در نمودار دارد و به این معنی است که اگر ۵۰ درصد از داده‌ها بررسی شود، همگی درست پیش‌بینی شده‌اند [۱۱]. در شکل ۱۰ نمودار Lift با در نظر گرفتن ۵۰ درصد جمعیت نشان داده شده است. با توجه به جدول ۲ مدل درخت تصمیم دارای بیش‌ترین درصد (۴۸/۱۷

درصد) پیش‌بینی درست است. شبکه عصبی پیش خور با ۴۷/۷۵ درصد و مدل مبتنی بر قاعده‌ی بیز با ۴۶/۲۶ درصد پیش‌بینی‌ها را درست انجام می‌دهند.



شکل ۳: نمودار Lift بدون مشخص کردن مقدار متغیر مورد پیش‌بینی (۵۰٪ جمعیت)

جدول ۲: شرح نمودار داده کاوی (Mining Legend) با در نظر گرفتن ۵۱٪ جمعیت

مدل	دقت	احتمال پیش بینی	جمعیت هدف
مدل درخت تصمیم	٪۰/۹	٪۸۲/۲۲	٪۴۸/۱۷
مدل مبتنی بر قاعده بیز	٪۰/۸۷	٪۹۱/۴۶	٪۴۶/۲۶
مدل شبکه عصبی	٪۰/۸۹	٪۸۹/۸۵	٪۴۶/۲۶
مدل ایده آل			٪۵۱

با توجه به نتایج بدست آمده از نمودارها با محدوده جمعیت‌های مختلف، مشخص می‌شود که هر یک از سه مدل می‌توانند بهتر یا بدتر از مدل‌های دیگر باشند، به همین دلیل در مقایسه‌ی کلی مدل‌ها از پارامتر دقت (Score) که در شرح نمودار داده-کاوی (Mining Legend) وجود دارد استفاده شده است. هرچه عدد مذکور مدل به یک نزدیک‌تر باشد، آن مدل دقیق‌تر است. با توجه به شکل ۵ مدل درخت تصمیم بیش‌ترین دقت (۰/۹) و بعد از آن شبکه عصبی پیش خور با دقت ۰/۸۹ و مدل مبتنی بر قاعده‌ی بیز با دقت ۰/۸۷ قرار دارند.

۲-۲- ماتریس طبقه‌بندی (Classification Matrix)

ماتریس طبقه‌بندی، فراوانی درستی و نادرستی پیش‌بینی‌ها را نشان می‌دهد. در این ماتریس مقادیر واقعی در مجموعه داده‌های آزمون با مقادیر پیش‌بینی شده مقایسه می‌شوند. در مجموعه داده‌ی آزمون، ۵۹۲ نفر دارای اختلال سندرم متابولیک و ۲۲۴۹ نفر فاقد اختلال سندرم متابولیک هستند. در جدول‌های ۱، ۲ و ۳ نتایج ماتریس طبقه‌بندی برای سه مدل درخت تصمیم و مدل مبتنی بر قاعده‌ی بیز و مدل شبکه عصبی پیش خور نشان داده شده است. سمت چپ‌ترین ستون در هر ماتریس نشان دهنده‌ی مقادیر پیش‌بینی شده از طریق مدل‌ها و دیگر ستون‌ها بیانگر مقادیر واقعی داده‌های آزمون هستند (۱ برای مبتلایان به اختلال سندرم متابولیک و ۰ برای غیرمبتلایان به اختلال سندرم متابولیک). قطر ماتریس تعداد پیش‌بینی‌های درست را نشان می‌دهد.

جدول ۳: ماتریس طبقه بندی مدل درخت تصمیم

	پیش بینی	واقعی=۰	واقعی=۱
مدل درخت تصمیم	۰	۲۱۳۷	۴۱۵
مدل درخت تصمیم	۱	۱۱۲	۱۷۷
مدل مبتنی بر قاعده بیز	۰	۱۸۸۹	۲۳۲
مدل مبتنی بر قاعده بیز	۱	۳۶۰	۳۶۰
مدل شبکه عصبی	۰	۲۱۵۳	۴۱۸
مدل شبکه عصبی	۱	۹۶	۱۷۴

معیارهای متفاوتی برای ارزیابی عملکرد روش های داده‌کاوی وجود دارد [۱۰]. یکی از این معیارها، اندازه‌گیری دقت پیش‌بینی است که در رابطه ۱ با محاسبه‌ی درصد درستی پیش‌بینی هر دو حالت ۰ و ۱ (غیرمبتلا و مبتلا به سندرم متابولیک) به دست می‌آید.

Accuracy = (TP+TN)/(TP+TN+FP+FN) (۱)

که در آن مثبت واقعی (TP) تعداد مقادیر مثبت درست پیش‌بینی شده، منفی واقعی (TN) تعداد مقادیر منفی درست پیش‌بینی شده، منفی کاذب (FP) تعداد مقادیر مثبت غلط پیش‌بینی شده و منفی کاذب (FN) تعداد مقادیر منفی است که غلط پیش‌بینی شده‌اند. با استفاده از نتایج ماتریس طبقه‌بندی، مقادیر دقت (accuracy) هر یک از مدل‌ها محاسبه و در جدول ۴ نشان داده شده است. برای ارزیابی مدل‌ها در پیش‌بینی درست افراد مبتلا به سندرم متابولیک از معیار دقت و صحت (F-measure) استفاده شده است [۱۲]. مقدار این معیار از رابطه‌ی ۲ بدست می‌آید:

$$\text{F-measure}=\frac{2\times(\text{recall}\times\text{precision})}{(\text{recall}-\text{precision})} \quad (۲)$$

که در آن recall به معنی درصد پیش‌بینی درست بیماران مبتلا به سندرم متابولیک در میان افرادی است که به عنوان بیمار پیش‌بینی شده‌اند و از رابطه‌ی ۳ بدست می‌آید. همچنین precision به معنی درصد پیش‌بینی درست افراد مبتلا به سندرم متابولیک در میان افرادی که به عنوان بیمار شناسایی شده‌اند و از رابطه‌ی ۴ بدست می‌آید. اندازه‌گیری دقت و صحت هر یک از مدل‌ها در جدول ۴ نشان داده شده است.

Recall = TP/(TP+ FN) (۳)

Precision = TP/(TP+ FP) (۴)

جدول ۴: نتایج دقت (Accuracy) و اندازه گیری دقت و صحت (F-Measure) الگوریتم‌ها

مدل	دقت (Accuracy)	اندازه گیری دقت و صحت (F-Measure)
درخت تصمیم	٪۰/۸۱	۰/۴
مدل بیز	٪۰/۷۷	۰/۵۴
شبکه‌های عصبی	٪۰/۸۲	۰/۴

با توجه به نتایج جدول ۴ مدل شبکه عصبی پیش خور دارای بیش‌ترین دقت (۰/۸۲) و پس از آن مدل‌های درخت تصمیم با دقت ۰/۸۱ و مدل مبتنی بر قاعده‌ی بیز با دقت ۰/۷۷ قرار دارند. همچنین از مقادیر اندازه گیری دقت و صحت موجود در جدول ۴ می‌توان نتیجه گرفت که هر سه مدل قادر به ایجاد الگوهایی در جهت پیش‌بینی متغیر هدف هستند، اما مدلی که دارای بهترین پیش‌بینی در رابطه با بیماران مبتلا به سندرم متابولیک می‌باشد مدل مبتنی بر قاعده‌ی بیز و بعد از آن مدل‌های شبکه عصبی پیش خور و درخت تصمیم می‌باشند.

بحث:

هر یک از مدل‌ها در حد توانایی‌شان اطلاعات ریز و دقیقی را ارائه می‌دهند. مدل درخت تصمیم و مدل مبتنی بر قاعده‌ی بیز قادر به پاسخ گویی به سه تا از این اهداف و شبکه عصبی پیش

خور دو تا از این اهداف را پاسخ می‌دهند. بدون توجه به کارایی مدل‌ها، نتایج مدل درخت تصمیم آسان‌تر و قابل فهم تر هستند. همچنین جزئیات پروفایل بیمار فقط در مدل درخت تصمیم قابل دسترس است. از سوی دیگر مدل مبتنی بر قاعده‌ی بیز از این جهت که پیش‌بینی‌های با معنادارتری را ارائه می‌دهد نسبت به مدل درخت تصمیم رجحان دارد. فهمیدن روابط ایجاد شده بین ویژگی‌های مدل شبکه عصبی پیش خور مشکل می‌باشد. با توجه به این که قبلاً در هیچ مطالعه ای این سه روش آماری در تشخیص سندرم متابولیک مورد استفاده قرار نگرفته‌اند، اطلاعات قبلی در مورد کارایی روش‌های مذکور و رجحان هر یک به دیگری وجود ندارد. اما در مطالعه انجام شده توسط هلمینن و همکاران نشان داده شد که به طور کلی پزشکان در شناسایی دقیق بیماران مبتلا به سندرم متابولیک ناتوان و این دسته از بیماران نیز از وجود بیماری خود نا آگاه بوده‌اند. این یافته‌ها ضرورت پیدا کردن روشی دقیق برای تشخیص سندرم متابولیک را توجیه می‌کند [۱۳]. بسیاری از مطالعات قبلی تایید کرده‌اند که تجزیه و تحلیل با استفاده از درخت تصمیم ابزار موثر برای تجزیه و تحلیل داده های بالینی است [۱۴-۱۶]. مطالعه دیگری که در تایلند انجام شده است نشان داده که استفاده از درخت تصمیم‌گیری برای طبقه‌بندی خودکار اجزا سندرم متابولیک روش سودمندی است [۱۷] . همچنین یافته-های لمیوکس و همکاران نشان داد که تری گلیسرید افزایش یافته پارامتر مهمی است که می‌تواند به عنوان اولین فنوتیپ برای غربالگری و شناسایی مبتلایان به سندرم متابولیک استفاده شود [۱۸]. نتایج مطالعه دیگری در تایلند نشان داد که وجود ترکیب تری گلیسرید افزایش یافته به انضمام فشار خون بالا و یا قند خون بالا و فشار خون و همچنین تری گلیسرید افزایش یافته با قند خون بالا پیش بینی کننده های قدرتمندی در تشخیص سندرم متابولیک می‌باشند [۱۷]. نکته جالب توجه این است که تجزیه و تحلیل با روش درخت تصمیم گیری قادر به کشف سندرم متابولیک مطابق با معیارهای فدراسیون جهانی دیابت است [۱۷].

نتیجه‌گیری: بررسی‌های انجام‌شده روی مجموعه داده‌ها به همراه یافتن بهترین و کاراترین و همچنین بدترین الگوریتم در تشخیص سندرم متابولیک این حقیقت را مشخص می‌کند که نمی‌توان مدلی را با توجه به ساختار آن، همواره به عنوان مدل بهینه معرفی کرد و عوامل متعددی روی کارایی آن‌ها موثرند. با توجه به نتایج بدست آمده، مدل شبکه عصبی از دقت بالاتری (۰/۸۲) نسبت به دو مدل درخت تصمیم (۰/۸۱) و مدل مبتنی بر قاعده بیز (۰/۷۷) برخوردار است. اما مدل مبتنی بر قاعده‌ی بیز در تشخیص مبتلایان به سندرم متابولیک پیش‌بینی دقیق‌تری ارائه می‌دهد. وجود داده‌های مناسب، پیش‌پردازش مناسب و اعمال روش داده‌کاوی مناسب نتایج خوبی را در مورد داده‌های پزشکی ارائه می‌دهد.

References:

1. Maroco J, Silva D, Rodrigues A, et al. Data mining methods in the prediction of Dementia: A real-data comparison of the accuracy, sensitivity and specificity of linear discriminant analysis, logistic regression, neural networks, support vector machines, classification trees and random forests. BMC Res Notes 2011; 4: 299.
2. Jago R, Baranowski T, Buse J, et al. Prevalence of the metabolic syndrome among a racially/ethnically diverse group of U.S. eighth-grade adolescents and associations with fasting insulin and homeostasis model assessment of insulin resistance levels. Care 2008; 31(10): 2020-5.
3. Gharipour M, Kelishadi R, Baghaie M, et al. Metabolic Syndrome in an Iranian Adult Population. Eur Heart J 2006; 27: 250-1.
4. Kelishadi R, Ardalan G, Adeli K, et al. Factor analysis of cardiovascular risk clustering in pediatric metabolic syndrome: CASPIAN study. Ann Nutr Metab 2007; 51(3): 208-15.
5. Weatherley BD, Nelson JJ, Heiss G, et al. The association of the ankle-brachial index with incident coronary heart disease: the Atherosclerosis Risk in Communities (ARIC) study, 1987-2001. BMC Cardiovasc Disord 2007; 7: 3.
6. Sarraf-Zadegan N, Sadri G, Malek Afzali H, et al. Isfahan Healthy Heart Programme: a comprehensive integrated community-based programme for cardiovascular disease prevention and control. Design, methods and initial experience. Acta Cardiol 2003; 58(4): 309-20.
7. Kim TN, Kim JM, Won JC, et al. A decision tree-based approach for identifying urban-rural differences in metabolic syndrome risk factors in the adult Korean population. J Endocrinol Invest 2012 Jan 30. [Epub ahead of print]. DOI: DOI: 10.3275/8235.
8. Ho TJ. Data Mining and Data Warehousing", Prentice Hall, 2005.
9. Langley P, Iba W, Thompson K. An analysis of Bayesian classifiers, Proceedings of the 10th National Conference on Artificial Intelligence. Massachusetts: MIT Press; 1992: 223-8.
10. Witten HI, Frank E. Data mining, practical machine learning tools and techniques. 2nd ed. Massachusetts: Morgan Kaufmann Publ; 2005.
11. Larson B. Delivering business intelligence with Microsoft SQL Server 2008. 2nd ed. New York: McGraw-Hill Osborne Media; 2008.
12. Van Rijsbergen CJ. The geometry of information retrieval. 2nd ed. Oxford: Butterworth-Heinemann; 1979.
13. Helminen EE, Mantyselka P, Nykanen I, et al. Farfrom easy and accurate-detection of metabolic

syndrome by general practitioners. BMC Fam Pract 2009; 10: 76.
14. Szabo de Edelenyi F, Goumidi L, Bertrais S, et al. Prediction of the metabolic syndrome status based on dietary and genetic parameters using Random Forest. Genes Nutr 2008; 3(3-4): 173-6.
15. Firouzi F, Rashidi M, Hashemi S, et al. A decision tree-based approach for determining low bone mineral density in inflammatory bowel disease using WEKA software. Eur J Gastroenterol Hepatol 2007; 19(12): 1075-81.
16. Quentin-Trautvetter J, Devos P, Duhamel A, et al. Assessing association rules and decision trees on analysis of diabetes data from the Diab Care program in France. Stud Health Technol Inform 2002; 90: 557-61.
17. Worachartcheewan A, Nantasenamat C, Isarankura-Na-Ayudhya C, et al. Identification of metabolic syndrome using decision tree analysis. Diabetes Res Clin Pract 2010; 90(1): e15-8.
18. Lemieux I, Poirier P, Bergeron J, et al. Hypertriglyceridemic waist: a useful screening phenotype in preventive cardiology? Can J Cardiol 2007; 23(Suppl B): 23B-31B.

Research Article

Decision support in prediction of metabolic syndrome with data mining methods

Edrisi M[1], Gharipour M[2], Faroughi A[3], Javeri F[3], Shahgholi B[3], Gharipour A[4], Sarrafzadegan N[*5]

Received: 09/26/2011 Accepted: 12/04/2011

1. Dept. of Biomedical Engineering, School of Engineering, Isfahan University, Isfahan, Iran
2. Cardiac Rehabilitation Research Center, Isfahan Cardiovascular Research Center, Isfahan University of Medical Sciences, Isfahan, Iran
3. Dept. of Information Technology Engineering, School of Engineering, Isfahan University, Isfahan, Iran
4. Isfahan House of Mathematics, Isfahan University, Isfahan, Iran
5. Cardiovascular Research Center, Isfahan Cardiovascular Research Institute, Isfahan University of Medical Sciences, Isfahan, Iran

Journal of Jahrom University of Medical Sciences, Vol. 9, Suppl. 2, 2011

Abstract

Introduction:
The aim of this study was to find the most important risk factors which have a role in causing metabolic syndrome and also to evaluate the efficacy of different models by data mining.
Material and Methods:
We used the data of third phase of "Isfahan Healthy Heart Program" as data set, which was done on 9572 subjects in 2007. In this study, we evaluated the efficacy of 3 main algorithms including decision tree, Naïve Bayes and neural network to detect the subjects with metabolic syndrome.
Results:
The results of the study showed that BMI is the most significant factor leading to metabolic syndrome. The other risk factors included LDL-Cholesterol, age, education, type of employment, sex, physical activity, history of diabetes, hypertension, stroke, income, smoking history of hyperlipidemia, myocardial infarction, and heart rate.
Conclusion:
We found that the optimal algorithm might be different by the dataset and data preprocessing methods. Various factors have a role in the efficacy of algorithms; using data preprocessing methods increased the prediction accuracy of all the examined techniques.
Our results showed that artificial Neural Networks model has the highest accuracy in the pattern recognition while Naive Bayes was better in predicting the metabolic syndrome among healthy subjects.

Keywords: Data Mining, Decision Tree, Bayes Theorem, Neural Networks, Cardiovascular Disease

* Corresponding author, Email: nsarrafzadegan@gmail.com