پیش بینی بیماری قلبی با استفاده از تکنیک داده کاوی شبکه عصبی

مریم کاظمی[۱]، حسین مهدی زاده*[۲]، اردشیر شیری[۳]

۱) دانشگاه علوم پزشکی ایلام، ایلام، ایران

۲) گروه کارآفرینی و توسعه روستایی، دانشکده کشاورزی، دانشگاه ایلام، ایلام، ایران

۳) گروه مدیریت، دانشکده ادبیات و علوم انسانی، دانشگاه ایلام، ایلام، ایران

تاریخ دریافت: ۹۴/۱/۱۶ تاریخ پذیرش:۹۴/۱۲/۸

چکیده

مقدمه: داده کاوی به بررسی و تجزیه و تحلیل مقادیر عظیمی از داده ها به منظور کشف الگوها و قوانین معنی دار اطلاق میشود که عمدتا" از طریق ساختن مدل ها و الگوریتم ها، ورودی ها را با هدف خاصی مرتبط می نماید. گاهی تکنیک های داده کاوی منجر به شناسایی الگوریتم های معنادار می شوند که می توانند با استفاده از داده های موجود و در دسترس و با هزینه کم، زمینه های ابتلا، پیشگیری و درمان بیماری ها را در پزشکی فراهم آورده و پزشک ها را در تشخص به موقع یاری رساند.

مواد و روش ها: این مطالعه با هدف استفاده مدیران بیمارستان از نتایج حاصل از داده کاوی سیستم های اطلاعات بیمارستانی جهت پیش بینی دقیق تر و تصمیم گیری مؤثرتر در درمان بیماران صورت گرفته است. داده‌های مورد استفاده در این مطالعه، مربوط به اطلاعات۲۷۰ بیمار است که از انبار داده سایت UCI استخراج شده و شامل ۱۴ متغیر است. از مدل" شبکه عصبی" برای پیش-بینی مبتلا بودن به بیماری قلبی استفاده شده و دقت پیش‌بینی آن مورد بررسی و مقایسه قرار گرفته است.

یافته های پژوهش: بر اساس نتایج، مشاهده می‌شود، مدل شبکه عصبی با ساختار پرسپترون چند لایه با دقتی برابر با ۸۳,۳۳٪ عمل کلاس‌بندی را برای مجموعه مشاهدات آزمون انجام داده است.

بحث و نتیجه گیری: نتایج نشان داد که دقت مدل در کلاس‌بندی رکوردها از لحاظ متغیر پاسخ بیماری قلبی(*Heart-dis*)برای مجموعه رکوردهای مدل‌ساز ۸۷,۷۵٪ و برای مجموعه رکوردهای آزمون ۸۳,۳۳٪ می‌باشد. همچنین متغیرهای تعداد عروق بزرگ (*Nbr-ves*)، کاهش استرس(*ST-dep*)، نقص (*Defect*)، درد قفسه سینه(*Chest-pain*)، اوج استرس(*Peak-ST*)، ضربان قلب (*Heart-rate*)، آنژین(*Angina*)، جنسیت(*Sex*)، سن(*Age*) ،ایستایی نوار قلب (*Res-elec*)، فشار خون (*Blood-press*)، قندخون (*Blood-sugar*) و کلسترول سرم(*Serum-chol*) به ترتیب بیشترین اهمیت را از لحاظ مدل "شبکه عصبی با ساختار پرسپترون چند لایه" در پیش‌بینی متغیر پاسخ بیماری قلبی(*Heart-dis*) دارند.

کلمات کلیدی: داده کاوی، شبکه عصبی، بیماری قلبی

*نویسنده مسئول: گروه کارآفرینی و توسعه روستایی،دانشکده کشاورزی، دانشگاه ایلام، ایلام، ایران

Email:Hossein.mahdizadeh@ilam.ac.ir

[DOI: 10.29252/sjimu.25.1.20]

مقدمه

امروزه مدیریت یکی از ارکان اساسی اداره جوامع است، زیرا ترکیب و تلفیق مناسب عوامل موجود و ایجاد هماهنگی میان آنها در نتیجه تصمیم گیری صحیح برای رسیدن به هدف مورد نظر حاصل می شود. تصمیم گیری و مدیریت را می توان مترادف دانست. زیرا تصمیم گیری جزء اصلی مدیریت است(۱).

اطلاعات اساس تصمیم گیری است. مدیر برای تصمیم گیری صحیح نیاز به استفاده از اطلاعات درست و نیز ابزارها و مدل های تصمیم گیری دارد. همکاری متخصصان در زمینه کامپیوتر و پزشکی راه حل جدیدی را در تحلیل این داده ها و به دست آوردن الگوهای مفید و کاربردی ارائه می دهد که همان داده کاوی است. در داده کاوی بر خلاف علم آمار به دنبال پیشگویی هستند نه کشف یا اثبات. بدین معنا که با استفاده از روش های داده کاوی به دنبال تأیید آنچه از قبل وجود دارد نیستند، بلکه به دنبال مشخص کردن الگوهای از قبل شناخته نشده هستند(۲).

در واقع داده کاوی شکل پیشرفته پشتیبانی از تصمیم است و برخلاف ابزارهای پرس و جوی غیرفعال Passive، بدون الزام به طرح سوال از طرف کاربر، به تولید الگو، روندها و قواعد برنامه ریزی شده می پردازد. به عبارت دیگر قدرت داده کاوی در این است که می تواند الگوهایی را که در جستجوی کاربر مورد توجه قرار نگرفته است، افشا کند و پاسخ هایی را برای سوالاتی که هرگز درخواست نشده بود، تولید نماید(۳).

در حقیقت بیان کاربرد عملی داده کاوی در حوزه های مختلف، با استفاده از داده های ثبت شده در پایگاه داده است که به فراهم کردن اطلاعات ضروری و دانش مورد نیاز پزشکان در تصمیم گیری بهتر کمک می کند. یکی از مواردی که درآن با افزایش حجم داده روبرو هستیم، دانش پزشکی و صنعت سلامت است.

صنعت سلامت به طور مستمر در حال تولید میزان زیادی از داده ها می باشد. افرادی که با این نوع داد ه ها، مواجه هستند، دریافته اند که بین جمع آوری تا تفسیر آن ها شکاف وسیعی وجود دارد. حوزه ی به نسبت جوان و در حال رشد داده کاوی در سلامت، از جمله شیوه هایی است که می تواند این صنعت را از تحلیل عمیق این داده ها بهره مند سازد و به توسعه ی تحقیقات پزشکی و تصمیم گیری های علمی در زمینه‌ی تشخیص و درمان منتج شود(۳).

امروزه در دانش پزشکی جمع آوری داده های فراوان در مورد بیماری های مختلف از اهمیت فراوانی برخوردار است. مراکز پزشکی با مقاصد گوناگونی به جمع آوری این داده ها می پردازند. تحقیق روی این داده ها و به دست آوردن نتایج و الگوهای مفید در رابطه با بیماری ها، یکی از اهداف استفاده از این داده ها است. حجم زیاد این داده ها و سردرگمی حاصل از آن مشکلی است که مانع رسیدن به نتایج قابل توجه می شود. بنابراین از داده کاوی برای غلبه بر این مشکل و به دست آوردن روابط مفید بین عوامل خطر زا در بیماری های قلب و عروق استفاده می شود. این بیماری ها با توجه به شیوع و سهمی که در مرگ و میر انسان ها دارند از اهمیت بالایی برخوردارند. داده کاوی دانش استخراج روابط و الگوهای مفید پنهان در حجم زیاد داده است(۲).

تحول در صنعت سلامت به واسطه این هدف واحد که چگونه سازمانهای سلامت هزینه ها را کاهش و کیفیت را افزایش دهند و هم چنان رقابتی بمانند؟ به پیش می رود و این مقوله همواره یک چالش بزرگ محسوب می شود. بهبود کیفیت در صنعت سلامت را می توان به واسطه نیروهای محرکی که بر آن تاثیرگذار است بهتر تعریف نمود و از جمله این نیروهای محرک، داده های سلامت است، به عبارت دیگر در هر نوع برنامه ی بهبود کیفیت متمرکز بر بیمار، داده ها قلب آن برنامه به حساب می آید. داده ها در عصر امروزی یعنی عصر اطلاعات، عمده ترین دارایی برای سازمان های سلامت بوده و موفقیت های سازمان های سلامت در گروی جمع آوری، ذخیره و تحلیل آنها است. با این وجود، جمع آوری و ذخیره ی میزان زیادی از داده ها به شکل سودمند استفاده شده و تبدیل به یک منبع مالی برای سازمان گردد. برای تبدیل این ارزش بالقوه به اطلاعات استراتژیک، بسیاری از سازمان ها به داده کاوی روی آورده اند، چرا که به واسطه داده کاوی

[DOI: 10.29252/sjimu.25.1.20]

امکان کشف روابط، روندها و الگوهای مخفی بین داده ها و دستیابی به دانش نوین در زمینه چالش های آشکار و نهان سازمان میسر خواهد شد(۴).

در سالهای اخیر استفاده از روشهای داده کاوی روی حجم زیادی از داده ها با هدف تولید مدلها و الگوهای پیش بینی کننده در حیطه های متعدد پزشکی رواج یافته است. در حال حاضر مطالعات متعددی مؤکد این است که تکنیک های داده کاوی ابزار مؤثری را برای شناسایی الگوهای مهم سلامت از درون پرونده های پزشکی فراهم می کند. پرونده های سلامت کامپیوتری به واسطه ی دربرداشتن مجموعه ای از داده ها درباره ی تشخیص، درمان، اقدامات آزمایشگاهی و دارویی به طور بالقوه منبع غنی از دانش هستند اگر چه کشف دانش از انبوه داده ها در آن ها، هستند برای انسان غیر ممکن نیست، اما امری دشوار است و داده کاوی بهترین شیوه برای حل این چالش می باشد(۳).

بیماریهای قلبی عروقی از مهمترین عوامل مرگ و میر و از عمده ترین مشکلات بهداشتی در کشورهای پیشرفته و در حال توسعه می باشد(۵). بر اساس گزارش سازمان بهداشت جهانی ۴۱/۳ درصد کل مرگهای سال۲۰۰۵در ایران ناشی از بیماریهای قلبی عروقی بوده است و با کمال تأسف پیش بینی می شود تا سال ۲۰۳۰ این میزان به ۴۴/۸ درصد برسد(۶).

در ابتدای قرن بیستم ۱۰درصد کل مرگ و میرها به علت بیماری های قلبی عروقی بود. در انتهای همین قرن موارد مرگ و میر ناشی از بیماری های قلبی به ۲۵ درصد افزایش یافت و پیش بینی می شود با توجه به روند کنونی تا سال ۲۰۲۵ میلادی بیشتر از ۳۵ تا ۶۰ درصد موارد مرگ و میر در جهان از بیماری های قلبی عروقی ناشی شود(۲).

با توجه به شیوع و سهمی که بیماری های قلبی در مرگ و میر انسان ها دارند، در این پژوهش، اطلاعات مربوط به بیماری های قلبی، با استفاده از تکنیک های داده کاوی، مورد تحلیل قرار گرفته است تا بتوان با استفاده از این اطلاعات، به مدل و الگوی دقیق تری جهت پیش بینی بیماری قلبی دست یافت.

بیمارستان بازوی مهم ارائه خدمات بهداشتی درمانی و اولین سطح ارجاع با قلمرو مسئولیت های مشخص است و به این لحاظ مهمترین سازمان بهداشتی درمانی به شمار می آید. از یک سو به دلیل اهمیت شیوه های تصمیم گیری مدیران بیمارستان ها در پیشبرد اهداف بیمارستان و اهمیت قدرت پیش بینی آنان در حل مشکلات درمانی بیماران، و از سوی دیگر به دلیل عدم انجام مطالعات مشابه در این زمینه، این مطالعه با هدف استفاده مدیران بیمارستان ها از نتایج حاصل از داده کاوی سیستم های اطلاعات بیمارستانی (HIS) جهت پیش بینی دقیق تر و تصمیم گیری بهتر و مؤثرتر برای درمان بیماران صورت گرفته است. گاهی تکنیک های داده کاوی منجر به شناسایی الگوریتم های معنادار می شوند که می توانند با استفاده از داده های موجود و در دسترس و با هزینه کم، زمینه های ابتلا، پیشگیری و درمان بیماری ها را در پزشکی فراهم آورده و می تواند پزشک ها را در تشخص به موقع یاری رساند.

پیشینه تحقیق:

هدف داده کاوی کشف انگاره های معتبر، جدید و قابل ردیابی در حجم عظیمی از داده ها با استفاده از ابزارهای آماری و هوش مصنوعی است. تاریخچه کشف دانش در پایگاه های اطلاعاتی که امروزه به داده کاوی مشهور است قدمت چندانی ندارد. پژوهشی در خصوص بیماران مبتلا به سرطان پستان که حداقل هرکدام به مدت دو سال تحت پیگیری بوده اند، انجام دادند. اطلاعات این بیماران در مرکز تحقیقات سرطان پستان جهاد دانشگاهی برای پیگیری اقدامات درمانی ثبت و بیماران حداقل به مدت دو سال پس از تشخیص، تحت نظر این مرکز بوده و پیگیری های بعدی برای آنها انجام شد. بررسیهای صورت گرفته نشان داد که دقت در سه الگوریتم داده کاوی، یعنی درخت تصمیم گیری، ANN، SVM، به ترتیب ۰/۹۳۶، ۰/۹۴۷، ۰/۹۵۷ بوده است(۷).

محمودی و همکاران(۱۳۹۲) در تحقیقی به شناسایی مدل پیش بینی بیماری عروق کرونر با استفاده از شبکه های عصبی و گزینش متغیر مبتنی بر درخت رگرسیون و طبقه بندی پرداخته اند. در این مطالعه مدل بدست آمده مبتنی بر شبکه های عصبی، علاوه بر توانایی بالا در تشخیص افراد بیمار، تعداد قابل قبولی از افرادی که فاقد بیماری عروق کرونر بودند را نیز شناسایی کرد.

[DOI: 10.29252/sjimu.25.1.20]

همچنین، بکار گیری تکنیک های گزینش متغیر در این مطالعه نیز نتایج خوبی در زمینه کاهش پیچیدگی مدل به همراه داشت و منجر به تولید مدلی متشکل از تنها چهار ریسک فاکتور سن، جنس، دیابت و فشارخون بالا گردید(۸).

در بیمارستانBrigham and Women تحقیقی با فرضیه ای مبنی بر این که" می توان عناصر اطلاعاتی موجود در پرونده ی الکترونیک سلامت را با تکنیک داده کاوی شناسایی و روابط بالینی معنی دار و درستی را کشف کرد " انجام شد؛ در این مطالعه با اجرای داده کاوی بر روی۱۰۰۰۰۰پرونده و تمرکز بر مجموعه داد ه های تشخیصی (۲۷۲۷۴۹ مورد)، دارویی (۴۴۲۶۵۸ مورد) و نتایج آزمایشگاهی (۱۱۸۰۱۰۶۸مورد) روابط علمی مهمی بین مشکلات بیمار، داروها و نتایج آزمایشگاهی آن ها کشف و توصیف شد(۹).

در ایالت آلاما نوعی سیستم نظارتی وجود دارد که از تکنیکهای داده کاوی استفاده می کند. این سیستم با استفاده از قوانین و روابط داده کاوی بر روی کشت خون بیمار و داده های بالینی به دست آمده از سیستم اطلاعات الگوهای (Laboratory information system) آزمایشگاه جدید و جالب توجهی را مشخص می سازد و ماهانه الگوهایی که توسط کارشناسان کنترل عفونت مورد بررسی قرار م یگیرد را تهیه می کند. سازندگان این سیستم دریافته اند که ارتقای کنترل عفونت با سیستم داده کاوی حساس تر از سیستم کنترل عفونت سنتی عمل می کند(۱۰).

در تحقیقی که درسال ۲۰۰۷ توسط گروهی از محققین دانشگاه توکیو انجام گردید از شبکه های عصبی مصنوعی در این خصوص استفاده شد. از ۳۷۲۵۶ بیماری که اطلاعات آنها در یک دوره ۵ ساله از پایگاه داده استخراج شده بود ۸۱ متغیر از قبیل سن، جنسیت، مرحله بیماری و.... انتخاب گردیدند. نتایج مطالعات میزان دقت این مدل را در پیش بینی بقا بیماران ۸۴/۵٪ نشان داد. یعنی مدل داده کاوی به کار رفته بر مبنای شبکه های عصبی به میزان ۸۴/۵٪ توانست میزان بقای بیماران مبتلا به سرطان پستان را به درستی پیش بینی کند(۱۱).

کاظمی و همکاران(۱۳۹۲) تحقیق را با عنوان تشخیص بیماری هپاتیت با ترکیب روشهای داده کاوی انجام داده اند. در این تحقیق مدلی ترکیبی از داده کاوی ارائه شده تا ضمن تعیین ویژگی های مهم بیماری هپاتیت بالاترین دقت در پیش بینی قابلیت زندگی افراد مبتلا به هپاتیت را ارائه دهد. نتایج تحقیق بدست آمده نشان می دهد که این مدل با کاهش ۶۸٪ ویژگی ها (حذف ۱۳ ویژگی از ۱۹ ویژگی) با دقت ۹۷,۴۲٪ پیش بینی را انجام داده است(۱۲).

در تحقیق انجام شده توسط ژو و دیگران(۲۰۱۰) بر روی ۲۰۰۰۰ بیمار بستری طب سنتی چینی و ۲۰۰۰۰ بیمار سرپایی از روشهای خوشه بندی استفاده گردید که از میان ۵ روش خوشه بندی نظیر SVM ، شبکه عصبی و درخت تصمیم، برای سندروم های متفاوت در ۱۰۶۹ مورد اپیدمیولوژی بالینی، روش SVM نتایج پیشگویانه بهتری را نشان داد. همچنین از قوانین وابستگی و تحلیل شبکه پیچیده (CAN) جهت یافتن نقاط مفید برای انجام عمل طب سوزنی و الگوهای ترکیب گیاهان دارویی، از تکنیک آموزش ماشین چندگانه(SVM) جهت پیشگویی شروع نفروپاتی و از شبکه های عصبی برای پیش بینی نقش اطلاعات تشخیصی در اثر درمان آرتوئید روماتوئید بهره جسته شد(۱۳).

کلاهی و رافع(۱۳۹۲) تحقیق را با عنوان ارائه راه حلی برای تشخیص بیماری به کمک تکنیک های داده کاوی به انجام رسانیده اند. در این تحقیق از ترکیب یکی از تکنیک های داده کاوی به نام روش نزدیکترین همسایه مجاور به همراه انتخاب ویژگی جهت تشخیص و پیش بینی سرطان سینه استفاده شده و از تکنیک F-score برای انتخاب مهمترین ویژگی ها استفاده نموده و همچنین معیارهای شباهت مختلف را برای تعیین یک معیار شباهت مناسب مورد بررسی قرار داده و مقدار مناسبk را براساس معیار شباهت انتخاب شده، برای روشKNN تعیین کرده اند. این روش ترکیبی را بر روی مجموعه داده سرطان سینه Wisconsin از مخزن داده UCI پیاده سازی نموده اند(۱۴).

هنرور عراقی(۱۳۸۸) پایان نامه خود را با عنوان آنالیز بیماریهای قلبی به روش داده کاوی با تکنیک استخراج

[DOI: 10.29252/sjimu.25.1.20]

دانش مبتنی بر شهود انجام داده است. در این پایان‌نامه روشی برای انجام تشخیص پزشکی از روی پایگاه داده ی پزشکی پیشنهاد شده است که شامل ایجاد فاکتور اطمینان از روی اپیزودهای پزشکی با توجه به داده‌های بالینی- تشخیصی، تعیین توابع وزنی از روی این فاکتور با دو روش نگاشت و فازی، ترکیب توابع وزنی از اپیزودهای مختلف و در نهایت تشخیص نهایی است. روش ارائه شده با صحت قابل قبولی توانایی تشخیص آنفارکتوس میوکارد و محل آنرا فراهم کرد. همچنین این روش توانایی بکارگیری در سایر تشخیص های پزشکی را نیز دارا می‌باشد(۱۵).

بنابراین از داده کاوی برای غلبه بر این مشکل استفاده می شود و با توجه به شیوع و سهمی که بیماری های قلبی در مرگ و میر انسان ها دارند، در این پژوهش، اطلاعات مربوط به بیماری های قلبی، با استفاده از تکنیک های داده کاوی، مورد تحلیل قرار گرفته است تا بتوان با استفاده از این اطلاعات، به مدل و الگوی دقیق تری جهت پیش بینی بیماری قلبی دست یافت.

تعاریف مفهومی متغیرهای پژوهش:

- داده کاوی: داده کاوی به بررسی و تجزیه و تحلیل مقادیر عظیمی از داده ها به منظور کشف الگوها و قوانین معنی دار اطلاق میشود که عمدتا" از طریق ساختن مدل ها و الگوریتم ها، ورودی ها را با هدف یا مقصد خاصی مرتبط می نماید.
- شبکه های عصبی: شبکه های عصبی با الهام گیری از الگوی مغز انسان ضمن فرایند آموزش، اطلاعات مربوطه را درون شبکه ذخیره می نمایند . این شبکه امکان یادگیری داشته و همانند شبکه های زیستی میتواند با توجه به اطلاعات اولیه چیزی را بیاموزد و یا بر اساس آموخته های خود تصمیم گیری نماید(۱۶).

مواد و روش ها

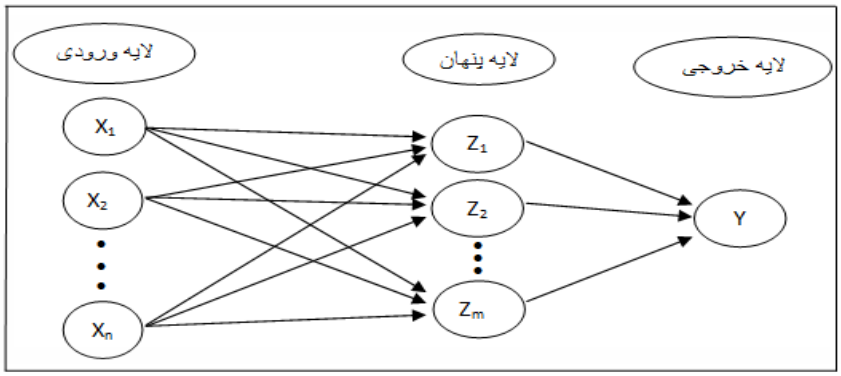
تحقیق حاضر دارای ۱۴ متغیر می‌باشد که از این میان ۱۳ متغیر بعنوان متغیر توضیحی (ورودی) و یک متغیر بعنوان متغیر پاسخ در نظر گرفته می‌شود.

دراین تحقیق از آمار توصیفی برای توصیف مشخصات دموگرافیک استفاده شده و برای بررسی سوالات تحقیق از ابزارهای داده‌کاوی درنرم افزار Clementine استفاده شده است. ابزارهای مختلفی به منظور کلاس-بندی وجود دارند که از جمله پرکاربردترین آنها می‌توان به شبکه عصبی(NeuralNetwork)، کارت، جنگل تصادفی(RandomForest)، ماشین بردار پشتیبان (SupportVectorMachine) و رگرسیون لجستیک(LogisticRegression) اشاره کرد. در این پژوهش با استفاده از مدل شبکه عصبی، عمل کلاس-بندی برای پیش‌بینی بیماری قلبی انجام شده است. در زیر به بررسی این مدل می پردازیم.

مدل شبکه عصبی

شبکه‌های عصبی مصنوعی، رده‌ای از روشهای یادگیری است که در زمینه‌های آماری و هوش مصنوعی رشد و توسعه یافته است و دارای قابلیت رگرسیونی و کلاس‌بندی می‌باشد. در مدل شبکه عصبی، متغیرهای توضیحی و متغیر پاسخ می‌توانند از نوع کمی یا کیفی باشند و همچنین متغیر پاسخ، رابطه-ای غیرخطی و غیر مستقیم با متغیر های توضیحی دارد. چارچوب شبکه‌های عصبی از چند لایه به نام‌های لایه ورودی، لایه (های) پنهان و لایه خروجی تشکیل شده است. شکل۱-۲، ساختار شماتیکی یک شبکه را با یک لایه پنهان نشان می‌دهد که در آن متغیرهای توضیحی و متغیر پاسخ به ترتیب در لایه ورودی و لایه خروجی واقع شده و اجزای لایه پنهان، Z_i ها، نورون یا گره(Node) نامیده می‌شوند.

[DOI: 10.29252/sjimu.25.1.20]



شکل شماره ۱- شمایی از ساختار شبکه عصبی

شبکه‌های عصبی را از لحاظ نحوه ی اتصال گره ها به دو نوع تقسیم می‌کنند.

۱) شبکه‌های عصبی پیشرو: کـه در آن گـره‌هـای هـر لایه فقط به گره‌های لایه بعدی متصل می‌شوند (شکل ۱).

۲) شبکه‌های عصبی پسرو: که در آن گره‌های هر لایه به گره‌های لایه‌های بعدی یا به خودشان متصل مـی-شوند.

در این مطالعه از شبکه‌های عصبی پیشرو استفاده شـده است. یکی از متداول‌ترین سبک‌های معمـاری شـبکه-های عصبی که در آن الگـوی ارتبـاطی بـین لایـه‌هـا مشـــخص مـــی‌شـــود، پرســـپترون چنـــد لایه(MultilayerPerceptron) است. در این سـاختار، هر نورون، تابعی از ترکیب خطی متغیرهای توضیحی، و همچنین متغیر پاسخ نیز تابعی از ترکیب خطی نورون‌ها است. یعنی:

$$Z_i = \sigma\left(\alpha_{0,i} + \sum_{j=1}^{n} \beta_{j,i} X_j\right);\ i = 1,2,\dots,m \quad , \quad y = f(\gamma + \sum_{i=1}^{m} \theta_i Z_i)$$

که در آن σ تابع فعال سازی نامیده شده و نمودار ایـن تابع غیـر خطـی، S- شـکل(Sigmoid Activation Function) است که از جمله آنها می‌توان بـه تانژانـت هایپربولیـک(HyperbolicTangent)، آرک تانژانـت (ArcTangent) و سینوس(Sin) اشاره کرد. همچنـین تابع f، خطی یا غیر خطی بوده و در صورت غیر خطی بودن می‌تواند همانند σ دارای یکی از فرم‌هـای مـذکور باشد. در برآورد پارامترهای مدل ابتدا مقادیر اولیـه ایـن پارامترها به تصادف انتخاب شـده و سـپس طـی یـک فرآیند یادگیری، این مقادیر دائما" بـروز مـی‌شـوند تـا زمانی‌که میزان خطای کـلاس‌بنـدی بـه حالـت نسـبتاً پایداری برسد(۱۷). نهایتـاً شـبکه عصـبی بـا محاسـبه احتمال شرطی قـرار گـرفتن مشـاهده ی جدیـد x_0 در کلاس 0 یا ۱ که به آن احتمال پسین گوینـد، بیمـاری قلبـی را پـیش‌بینـی مـی‌کنـد، بـه اینصـورت کـه اگـر $P(Y=1|x_0) > 0.5$، آنگاه مشاهده‌ی جدید x_0 به کلاس۱ و در غیر اینصورت به کلاس 0 تخصیص می‌یابد(۱۷).

معرفی داده‌ها و متغیرها:

داده‌هـای مـورد اسـتفاده در ایـن مطالعـه، مربـوط بـه اطلاعات۲۷۰ بیمار است که از انبـار داده سـایت UCI استخراج شده و شامل ۲۷۰ مشاهده و ۱۴ متغیـر مـی-باشد. در جدول۲ اطلاعات مربوط بـه متغیرهـای مـورد استفاده بطور خلاصه بیان شده است.

[DOI: 10.29252/sjimu.25.1.20]

جدول شماره۱- متغیرهای مورد استفاده در پژوهش

سن	Age	کمی
جنسیت	Sex	کیفی
درد قفسه سینه	Chest-pain	کیفی
فشار خون	Blood-press	کمی
کلسترول سرم	Serum-chol	کمی
قند خون	Blood-sugar	کیفی
ایستایی نوار قلب	Res-elec	کیفی
ضربان قلب	Heart-rate	کمی
آنژین	Angina	کیفی
کاهش ST	ST-dep	کمی
اوج ST	Peak-ST	کیفی
تعداد عروق بزرگ	Nbr-ves	کمی
نقص	Defect	کیفی
بیماری قلبی	Heart-dis	کیفی

در جدول۱، متغیر "بیماری قلبی" به عنوان متغیر پاسخ (پیش‌بینی شونده) و بقیه متغیرها به عنوان متغیر توضیحی (پیش‌بینی کننده) در نظر گرفته می‌شوند.

آمار توصیفی

در این بخش با استفاده از جداول فراوانی، اطلاعات مربوط به متغیرهای جمعیت شناسی تحقیق را توصیف می‌کنیم (جدول شماره ۲ و ۳).

جدول۲- آمار توصیفی متغیر "سن"

کمترین	بیشترین	میانگین	انحراف معیار
۲۹	۷۷	۵۴,۴۳	۹,۱۰۹

جدول۳- آمار توصیفی متغیر "جنسیت"

جنسیت	فراوانی	درصد فراوانی
زن	۸۷	۳۲,۲
مرد	۱۸۳	۶۷,۸
مجموع		۱۰۰

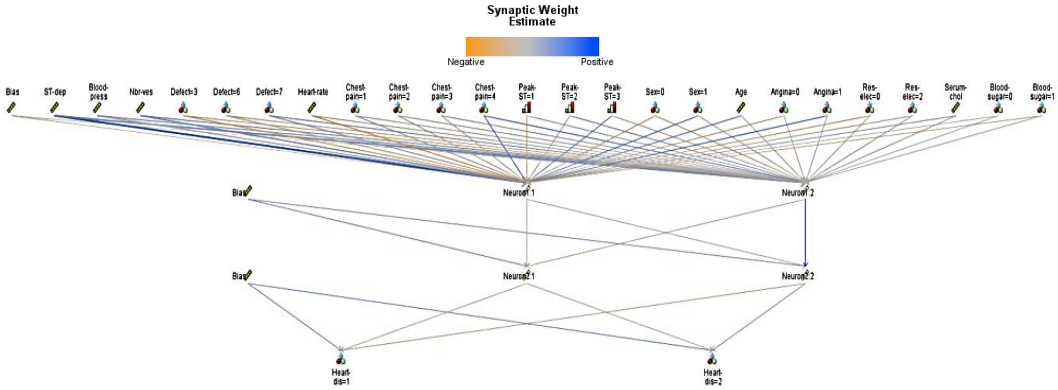
در این بخش با استفاده از مدل‌شبکه عصبی سعی در پیش‌بینی ابتلا به بیماری قلبی مراجعه کنندگان به بیمارستان بر اساس اطلاعات مربوط به متغیرهای توضیحی می‌شود.

در ابتدا مجموعه اصلی داده‌ها که شامل ۲۷۰مراجعه کننده است به دو قسمت مجموعه داده مدل‌ساز (۸۰٪ کل داده‌ها، ۲۰۴ مراجعه کننده) و مجموعه داده آزمون (۲۰٪ کل داده‌ها، ۶۶ مراجعه کننده) تقسیم می‌شود. با استفاده از مجموعه داده مدل‌ساز عمل مدل‌سازی مدل مربوطه انجام می‌شود. در نهایت با استفاده از مجموعه داده آزمون دقت مدل برازش داده شده مورد ارزیابی قرار می‌گیرد. کلیه مراحل ذکر شده برای برازش مدل در محیط نرم‌افزار Clementine14.2 انجام شده است.

[DOI: 10.29252/sjimu.25.1.20]

یافته های پژوهش

نتایج برازش مدل شبکه عصبی با ساختار پرسپترون چندلایه پس از برازش مدل شبکه عصبی ایجاد شده در شکل ۲ نشان داده شده است. بهترین مدل شبکه عصبی با ساختار پرسپترون چندلایه برازش داده شده، مدلی دارای دو لایه پنهان و در هر لایه پنهان ۲ نرون است.



شکل۲- نمودار مدل شبکه عصبی با دو لایه پنهان و در هر لایه پنهان دو نرون

ضرایب مربوط به اتصال گره‌ها در لایه‌های ورودی، پنهان و خروجی در جدول۴ نشان داده شده است.

مثبت یا منفی بودن هر یک از ضرایب از روی شکل ۲ و رنگ اتصال‌دهنده‌ها مشخص می‌شود. با توجه به شکل۲ و جدول۴، می‌توان مقادیر هر یک از گره‌ها را در لایه‌های پنهان و لایه خروجی محاسبه کرد. بعنوان مثال، مقادیر گره اول در لایه پنهان اول (Neuron 1.1)، گره اول در لایه پنهان دوم (Neuron 2.1) و گره‌های موجود در لایه خروجی به صورت زیر محاسبه می‌شود.

جدول٤- ضرایب اتصال گره‌ها در لایه‌های ورودی، پنهان و خروجی

نام گره (نود)	ضریب گره در لایه بعد	لایه متغیر	نوع گره
Bias	۲۳,۰۷	اول	**scale**
Peak-ST=1	۱۲	اول	**order_set**
Peak-ST=2	۱۱,۰۷۶	اول	**order_set**
Peak-ST=3	۱۰,۱۵	اول	**order_set**
ST-dep	۲۲,۱۵	اول	**scale**
Blood-press	۲۱,۲۳	اول	**scale**
Nbr-ves	۲۰,۳۰	اول	**scale**
Heart-rate	۱۶,۶۱	اول	**scale**
Age	۷,۳۸	اول	**scale**
Serum-chol	۲,۷۶	اول	**scale**
Defect=3	۱۹,۳۸	اول	**set**
Defect=6	۱۸,۴۶	اول	**set**
Defect=7	۱۷,۵۳	اول	**set**
Chest-pain=1	۱۵,۶۹	اول	**set**
Chest-pain=2	۱۴,۷۶	اول	**set**
Chest-pain=3	۱۳,۸۴	اول	**set**
Chest-pain=4	۱۲,۹۲	اول	**set**
Sex=0	۹,۲۳	اول	**set**
Sex=1	۸,۳۰	اول	**set**

[DOI: 10.29252/sjimu.25.1.20]

set	اول	۶،۴۶	Angina=0
set	اول	۵،۵۳	Angina=1
set	اول	۴،۶۱	Res-elec=0
set	اول	۳،۶۹	Res-elec=2
set	اول	۱،۸۴	Blood-sugar=0
set	اول	۰،۹۲	Blood-sugar=1
scale	دوم	۱۸	Bias
scale	دوم	۱۲	Neuron 1.1
scale	دوم	۶	Neuron 1.2
scale	سوم	۱۸	Bias
scale	سوم	۱۲	Neuron 2.1
scale	سوم	۶	Neuron 2.2

$$\{Neuron\ 1.1\} = Tanh\big(-۲۳،۰۷ - (۱۲ * \{\text{Peak} - \text{ST} = 1\}) + (۱۱،۰۷۶ * \{\text{Peak} - \text{ST} = 2\}) + (۱۰،۱۵ * \{\text{Peak} - \text{ST} = 3\}) - (۲۰،۳۰ * \{Nbr - ves\}) - (۲،۷۶ * \{\text{Serum} - \text{chol}\}) + (۲۱،۲۳ * \{\text{Blood} - \text{press}\}) + (۷،۳۸ * \{\text{Age}\}) - (۱۶،۶۱ * \{\text{Heart} - \text{rate}\}) + (۲۲،۱۵ * \{\text{ST} - \text{dep}\}) + (۱۹،۳۸ * \{\text{Defect} = 3\}) + (۱۸،۴۶ * \{\text{Defect} = 6\}) - (۱۷،۵۳ * \{\text{Defect} = 7\}) + (۱۵،۶۹ * \{\text{Chest} - \text{pain} = 1\}) - (۱۴،۷۶ * \{\text{Chest} - \text{pain} = 2\}) - (۱۳،۸۴ * \{\text{Chest} - \text{pain} = 3\}) + (۱۲،۹۲ * \{\text{Chest} - \text{pain} = 4\}) - (۹،۲۳ * \{\text{Sex} = 0\}) + (۸،۳۰ * \{\text{Sex} = 1\}) - (۶،۴۶ * \{\text{Angina} = 0\}) + (۵،۵۳ * \{\text{Angina} = 1\}) - (۴،۶۱ * \{\text{Res} - \text{elec} = 0\}) - (۳،۶۹ * \{\text{Res} - \text{elec} = 2\}) - (۱،۸۴ * \{\text{Blood} - \text{sugar} = 0\}) - (۰،۹۲ * \{\text{Blood} - \text{sugar} = 1\})\big)$$

$$\{Neuron\ 2.1\} = Tanh\big(۱۸ - (۱۲ * Neuron\ 1.1) - (۶ * Neuron\ 1.2)\big)$$

$$\{Heart - Dis = 1\} = ۱۸ + (۱۲ * Neuron\ 2.1) - (۶ * Neuron\ 2.2)$$

$$\{Heart - Dis = 2\} = ۱۸ - (۱۲ * Neuron\ 2.1) + (۶ * Neuron\ 2.2)$$

با توجه به روابط بالا اگر برای یک مشاهده جدید، مقدار برآورد شده $\{Heart - Dis = 1\}$ بزرگتر از $\{Heart - Dis = 2\}$ شود، آنگاه فرد مورد نظر به بیماری قلبی مبتلا نیست و بالعکس اگر مقدار برآورد شده $\{Heart - Dis = 1\}$ کوچکتر از $\{Heart - Dis = 2\}$ شود، آنگاه فرد مورد نظر به بیماری قلبی مبتلا است.

در نهایت با استفاده از شبکه عصبی برازش داده شده، عمل کلاس‌بندی برای مجموعه مشاهدات مدل‌ساز و آزمون انجام شده که نتایج آن به ترتیب در جدول۵ و ۶ بیان شده است.

[DOI: 10.29252/sjimu.25.1.20]

جدول ۵- کلاس‌بندی متغیر "بیماری قلبی" با استفاده از مدل "شبکه عصبی با ساختار پرسپترون چند لایه" برای مجموعه داده مدلساز

		پیش‌بینی شده			درصد صحیح
		Heart-dis =1	Heart-dis =2	غیرقابل پیش‌بینی	
مشاهده شده	Heart-dis =1	۱۰۸	۹	۱	٪۹۱,۵۲
	Heart-dis =2	۱۴	۷۱	۱	٪۸۲,۵۵
کل					٪۸۷,۷۵

با توجه به جدول۷ می‌بینیم که مدل "شبکه عصبی با ساختار پرسپترون چند لایه"با دقت ۸۷,۷۵ درصد عمل کلاس‌بندی را برای مجموعه داده مدل‌ساز انجام داده است. با توجه به جدول می‌بینیم که بر اساس مدل، یک مشاهده دارای کلاس متغیر $Heart-dis = 1$ و یک مشاهده دارای کلاس متغیر پاسخ $Heart-dis = 2$ ، قابل پیش‌بینی نبوده‌اند. علت این امر بدست آمدن مقدار ۰,۵ برای مقدار $Heart-dis = 1$ و $Heart-dis = 2$ برای این دو مشاهده است و مدل نمی‌تواند برای کلاس متغیر پاسخ این دو مشاهده تصمیم‌گیری کند.

جدول ٦- کلاس‌بندی متغیر "بیماری قلبی" با استفاده از مدل "شبکه عصبی با ساختار پرسپترون چند لایه"برای مجموعه داده آزمون

		پیش‌بینی شده		درصد صحیح
		Heart-dis =1	Heart-dis =2	
مشاهده شده	Heart-dis =1	۲۹	۳	٪۸۷,۵
	Heart-dis =2	۸	۲۶	٪۷۶,۴۷
کل				٪۸۳,۳۳

با توجه به جدول۸ می‌بینیم که مدل "شبکه عصبی با ساختار پرسپترون چند لایه"با دقت ۸۳,۳۳ درصد عمل کلاس‌بندی را برای مجموعه داده آزمون انجام داده است.

بطور کلی نحوه کلاس‌بندی مدل شبکه عصبی با ساختار پرسپترون چند لایه را برای دو مجموعه داده مدل‌ساز و آزمون در جدول ۷می‌توان مشاهده کرد.

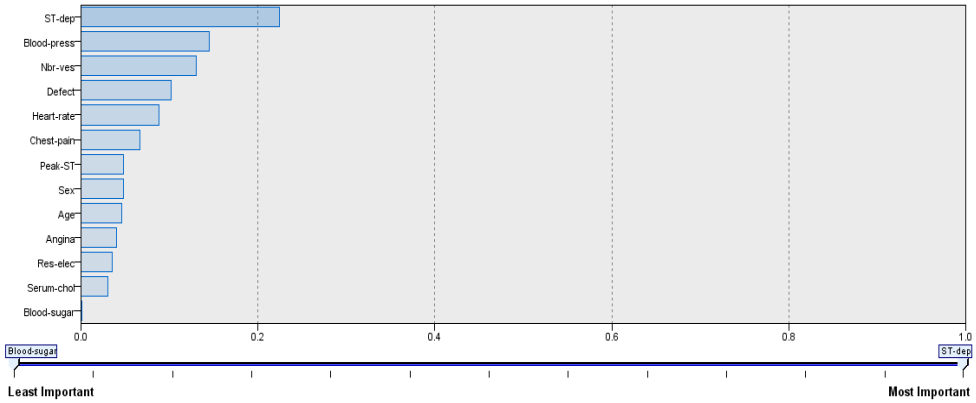
جدول ۷- کلاس‌بندی متغیر "بیماری قلبی" بر اساس مدل "شبکه عصبی با ساختار پرسپترون چند لایه"

نحوه کلاس‌بندی متغیر وابسته	داده‌های مدل‌ساز		داده‌های آزمون	
	تعداد	درصد	تعداد	درصد
صحیح	۱۷۹	۸۷,۷۵	۵۵	۸۳,۳۳
اشتباه	۲۵	۱۲,۲۵	۱۱	۱۶,۶۷
مجموع	۲۰۴	۱۰۰	۶۶	۱۰۰

رتبه‌بندی اهمیت متغیرها:

در مدل شبکه عصبی با ساختار پرسپترون چند لایه، ترتیب اهمیت متغیرهای توضیحی برای پیش‌بینی متغیر پاسخ به صورت شکل۱۰ می‌باشد.

[DOI: 10.29252/sjimu.25.1.20]



شکل۳- نمودار اهمیت متغیرها برای پیش‌بینی متغیر پاسخ در مدل شبکه عصبی با ساختار پرسپترون چند لایه

با توجه به شکل۳ می‌بینیم که متغیرهای ST-dep(کاهش ST)، Blood-press(فشار خون)، Nbr-ves(تعداد عروق بزرگ)، Defect (نقص)، Heart-rate (ضربان قلب)، Chest-pain (درد قفسه سینه)، Peak-ST(اوج ST)، Sex (جنسیت)، Age(سن)،Angina(آنژین)،Res-elec(ایستایی نوار قلب)، Serum-chol(کلسترول سرم) و Blood-sugar(قند خون) به ترتیب بیشترین اهمیت را از لحاظ مدل "شبکه عصبی با ساختار پرسپترون چند لایه" در پیش‌بینی متغیر پاسخ Heart-dis دارند

بحث و نتیجه‌گیری

امروزه بخش سلامت بیش ترین نیاز را به داده کاوی پیدا کرده است و حرکت از پزشکی سنتی به سمت پزشکی مبتنی بر شواهد از جمله مواردی است که می تواند مؤکد این امر باشد(۲).تعداد و اندازه پایگاه داده های پزشکی به سرعت در حال افزایش است و مدلهای توسعه یافته تکنیک داده کاوی می توانند برای پزشکان جهت کمک در تصمیم گیری موثر و کاربردی باشند(۷). بیمارستان بازوی مهم ارائه خدمات بهداشتی درمانی و اولین سطح ارجاع با قلمرو مسئولیت های مشخص است و به این لحاظ مهمترین سازمان بهداشتی درمانی به شمار می آید. از یک سو به دلیل اهمیت شیوه های تصمیم گیری مدیران بیمارستان ها در پیشبرد اهداف بیمارستان و اهمیت قدرت پیش بینی آنان در حل مشکلات درمانی بیماران، و از سوی دیگر به دلیل عدم انجام مطالعات مشابه در این زمینه، این مطالعه با هدف استفاده مدیران بیمارستان ها از نتایج حاصل از داده کاوی سیستم های اطلاعات بیمارستانی (HIS) جهت پیش بینی دقیق تر و تصمیم گیری بهتر و مؤثرتر برای درمان بیماران صورت گرفته است.

با توجه به برآورد وزن‌های مدل، نهایتا مدل با استفاده از رابطه زیر عمل کلاس‌بندی را برای رکوردها انجام می‌دهد.

$$\{\boldsymbol{Heart-Dis}=\mathbf{1}\} = ۱۸ + \left(۱۲ * Neuron\ 2.1\right) - \left(۶ * Neuron\ 2.2\right)$$

$$\{\boldsymbol{Heart-Dis}=\mathbf{2}\} = ۱۸ - \left(۱۲ * Neuron\ 2.1\right) + \left(۶ * Neuron\ 2.2\right)$$

با توجه به روابط بالا اگر برای یک مشاهده جدید، مقدار برآورد شده $\{Heart-Dis=1\}$ بزرگتر از $\{Heart-Dis=2\}$ شود، آنگاه فرد مورد نظر به بیماری قلبی مبتلا نیست و بالعکس اگر مقدار برآورد شده $\{Heart-Dis=1\}$ کوچکتر از $\{Heart-Dis=2\}$ شود، آنگاه فرد مورد نظر به بیماری قلبی مبتلا است.

به همین صورت عمل کلاس‌بندی برای تمام رکوردها (مجموعه رکوردهای مدل‌ساز و آزمون) محاسبه شده است. نتایج نشان داد که دقت مدل در کلاس‌بندی رکوردها از لحاظ متغیر پاسخ $\{Heart-dis\}$ برای مجموعه رکوردهای مدل‌ساز ۸۷٫۷۵٪ و برای مجموعه رکوردهای آزمون ۸۳٫۳۳٪ می‌باشد. که این میزان دقت با یافته های محمودی و همکاران(۱۳۹۲) همخوانی دارد.

همچنین متغیرهای تعداد عروق بزرگ(*Nbr-ves*)، کاهش استرس(*ST-dep*)، نقص(*Defect*)، درد قفسه سینه(*Chest-pain*)، اوج استرس(*Peak-ST*)، ضربان قلب(*Heart-rate*)، آنژین(*Angina*)،جنسیت(*Sex*)،سن(*Age*)،ایستایی نوار قلب(*Res-elec*)، فشار

[DOI: 10.29252/sjimu.25.1.20]

خون(*Blood-press*)، قند خون(*Blood-sugar*) و کلسترول سرم(*Serum-chol*) به ترتیب بیشترین اهمیت را از لحاظ مدل "شبکه عصبی با ساختار پرسپترون چند لایه" در پیش‌بینی متغیر پاسخ بیماری قلبی(*Heart-dis*) دارند.

References

1. Rezaeian A. Principles of organization and management.Tehran Samt Publication. 2012;P.81-6.

2.Amereh M. [Survey data mining algorithms and compare them on a case study]. Ecommerce 2014; 71:40-2.(Persian)

3.Moghadasi H, Hosseine A, Asadi F, Jahanbakhsh M. [Data mining and its application in health]. Health Inform Manage 2013; 9:297-304. (Persian)

4. Koh HC, Tan G. Data mining applications in healthcare. J Health Inform Manage 2005;19:64-72.

5. Anderson KM. Cardiovascular disease risk profiles. American Heart J1991;121:293-8.

6. Knight J, Wells S, Marshall R, Exeter D, Jackson R. Developing a synthetic national population to investigate the impact of different cardiovascular disease risk management strategies a derivation and validation study. PLos One2017;12: 170-3.

7.Tolueiashlaghi A, Purebrahimi A, Ebrahimi M, Ghasemahmad L. [Predict breast cancer recurrence by three data mining techniques]. Iranian J Breast Dis 2013;5: 24-34.(Persian)

8.Mahmoodi A, Asgarimoghadam R, Moazam M, Saeghian S. [Identify models predict coronary artery disease using Neural networks and variable selection based on classification and regression tree]. J Shahrekord Med Sci Uni2014; 15:22-7. (Persian)

9.Tavakoli N, Jahanbakhsh M. [Opportunities and challenges of EHR implementation in Isfahan]. J Manage Uni Isfahan 2010;3:41-7. (Persian)

10. Obenshain MK. Application of data mining techniques to healthcare data. Infect Control Hosp Epidemiol 2004; 25:690-5.

11. Endo A, Shibata T, Tanaka H. Comparison of seven algorithms to predict breast cancer survival. Biomed Soft Comput Hum Sci2008;13:11-6.

12.Kazemi A, Yousof zadeh A, Azimi P. Detect of Hepatitis by combining data mining methods. The first national conference application of intelligent systems in science and technology. 2014.

13. Zhou X. Development of traditional Chinese medicine clinical data warehouse for medical knowledge discovery and decision support. Art Intell Med J2010;48:139-52.

14.Kolahi S, Rafe W. Presentation a solution to diagnosis using data mining techniques. The first national conference electrical and computer southern Iran. 2014.

15.Henroraraghi B. [Heart disease analysis by data mining techniques to extract knowledge based on intuition]. Msc Thesis Azad Uni2010.(Persian)

16.Bahramizanvar M. [Data mining discover the hidden data]. Research office and risk control of Bank Sepah 2011. (Persian)

17. Hastie T, Tibshirani R .The Elements of Statistical Learning data mining inference and prediction. 2th ed. Springer Series Statistics Publication. 2009;P.184.

[DOI: 10.29252/sjimu.25.1.20]

Heart Disease Forecast Using Neural Network Data Mining Technique

Kazemi M[1], Mehdizadeh M[2], Shiri A[3]*

(Received: April 5, 2015 Accepted: February 27, 2016)

Abstract

Introduction: Data mining refers to the study and analysis of large amounts of data for discovering meaningful patterns and rules. Mainly through the models and algorithms, data mining puts the inputs in a specific order. Data mining techniques sometimes lead to the identification of meaningful algorithms which can use available and low-cost data in order to provide us with areas of infection, prevention, and treatment of diseases and help the physicians in timely and accurate diagnosis.

Materials & methods: The present paper aimed to study the use of the results of data mining of hospital information systems by hospital managers for more accurate prediction and more effective decision-making about treatment of patients. The data used in this study included the information of 270 patients (14 variables) extracted from the database of UCI website. A "neural networks" model was used for the prediction of affliction with heart disease and its accuracy was measured and compared.

Findings: According to the results, it can be observed that Multilayer Perceptron Neural Networks Model has classified the set of test observations with an accuracy of 83.33%.

*Discussion & conclusion*s: The results showed that the accuracy of "neural networks" model in classification of records in terms of heart disease response is 87.75% for the set of modeling records and 83.33% for the set of test records. In addition, the findings revealed that the variables of the number of large vessels (Nbr-ves), stress reduction (ST-dep), defect, chest pain, stress peak (Peak-ST), heart rate, angina, gender, age, static ECG (Res-elec), blood pressure (Blood-press), blood sugar, and serum cholesterol (Serum-chol), respectively, have the highest importance in "Multilayer Perceptron Neural Networks" model for the prediction of heart disease response.

Keywords: Data mining, Neural etwork, Heart-disease

1. Ilam University of Medical Sciences ,Ilam, Iran
2. Dept of Entrepreneurship and Rural Development, Faculty of Agriculture, Ilam University,Ilam, Iran
3. Dept of Management, Faculty of Humanities, Ilam University,Ilam, Iran
** Correspondin author Email: Hossein.mahdizadeh@ilam.ac.ir*

[DOI: 10.29252/sjimu.25.1.20]