

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 پذیرش: ۹۴/۱۲/۸

تاریخ دریافت: ۹۴/۱/۱۶

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 عمدتاً 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 تشخیص 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. داده های مورد استفاده در این مطالعه، مربوط به اطلاعات ۲۷۰ 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 پیش بینی آن مورد بررسی و مقایسه قرار گرفته است.

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

بحث و نتیجه گیری: نتایج نشان داد که دقت مدل در کلاس بندی رکوردها از لحاظ متغیر پاسخ بیماری قلبی (*Heart-dis*) برای مجموعه رکوردهای مدل ساز ۸۷,۷۵٪ و برای مجموعه رکوردهای آزمون ۸۳,۳۳٪ می باشد. 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 پیش بینی 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

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 در آن 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 داده 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 زمینه ی 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

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 سال ۲۰۰۵ در 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 ۱۰ درصد 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 تشخیص 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 هر کدام 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.

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 مناسب k را براساس معیار شباهت انتخاب شده، برای روش KNN 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

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 بیمارستان Brigham 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 روی ۱۰۰۰۰ پرونده و تمرکز بر مجموعه داده های تشخیصی (۲۷۲۷۴۹ مورد)، دارویی (۴۴۲۶۵۸ مورد) و نتایج آزمایشگاهی (۱۸۰۱۰۶۸ مورد) روابط علمی مهمی بین مشکلات بیمار، داروها و نتایج آزمایشگاهی آن ها کشف و توصیف شد (۹).

در ایالت آلباما 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 در سال 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).

مواد و روش ها

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

در این تحقیق از آمار توصیفی برای توصیف مشخصات دموگرافیک استفاده شده و برای بررسی سوالات تحقیق از ابزارهای داده کاوی در نرم افزار Clementine استفاده شده است. ابزارهای مختلفی به منظور کلاس بندی وجود دارند که از جمله پرکاربردترین آنها می توان 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 پیش بینی بیماری قلبی انجام شده است. در زیر به بررسی این مدل می پردازیم.

مدل شبکه عصبی

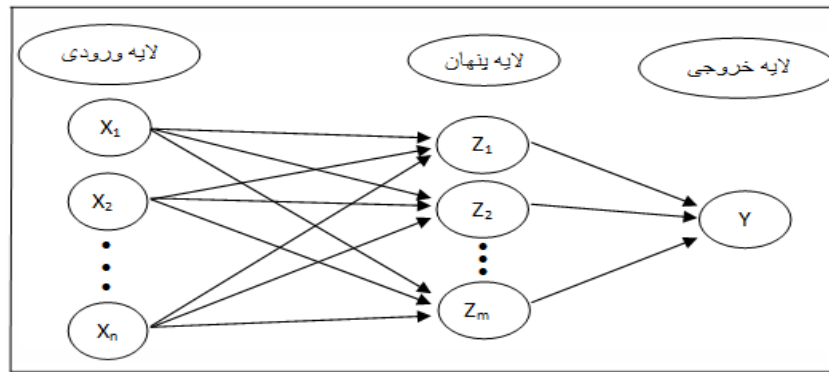
شبکه های عصبی مصنوعی، رده ای از روشهای یادگیری است که در زمینه های آماری و هوش مصنوعی رشد و توسعه یافته است و دارای قابلیت رگرسیونی و کلاس بندی می باشد. در مدل شبکه عصبی، متغیرهای توضیحی و متغیر پاسخ می توانند از نوع کمی یا کیفی باشند و همچنین متغیر پاسخ، رابطه ای غیرخطی و غیر مستقیم با متغیرهای توضیحی دارد. چارچوب شبکه های عصبی از چند لایه به نام های لایه ورودی، لایه (های) پنهان و لایه خروجی تشکیل شده است. شکل ۱-۲، ساختار شماتیکی یک شبکه را با یک لایه پنهان نشان می دهد 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 می شوند.

دانش مبتنی بر شهود انجام داده است. در این پایان نامه 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 داده های 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 آنفارتکوس میوکارد و محل آنرا فراهم کرد. همچنین این روش توانایی بکارگیری در سایر تشخیص های پزشکی را نیز دارا می باشد (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 عمدتاً 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).



شکل شماره ۱- شمایی از ساختار شبکه عصبی

هایپربولیک (HyperbolicTangent)، آرک تانژانت (ArcTangent) و سینوس (Sin) اشاره کرد. همچنین تابع f ، خطی یا غیر خطی بوده و در صورت غیر خطی بودن می تواند همانند σ دارای یکی از فرمهای مذکور باشد. در برآورد پارامترهای مدل ابتدا مقادیر اولیه این پارامترها به تصادف انتخاب شده و سپس طی یک فرآیند یادگیری، این مقادیر دائماً "بروز می شوند تا زمانی که میزان خطای کلاس بندی به حالت نسبتاً پایداری برسد (۱۷). نهایتاً شبکه عصبی با محاسبه احتمال شرطی قرار گرفتن مشاهده ی جدید x_0 در کلاس 0 یا 1 که به آن احتمال پسین گویند، بیماری قلبی را پیش بینی می کند، به اینصورت که اگر $P(Y = 1|x_0) > 0.5$ ، آنگاه مشاهده ی جدید x_0 به کلاس 1 و در غیر اینصورت به کلاس 0 تخصیص می یابد (۱۷).

معرفی داده ها و متغیرها:

داده های مورد استفاده در این مطالعه، مربوط به اطلاعات ۲۷۰ بیمار است که از انبار داده سایت UCI استخراج شده و شامل ۲۷۰ مشاهده و ۱۴ متغیر می باشد. در جدول ۲ اطلاعات مربوط به متغیرهای مورد استفاده بطور خلاصه بیان شده است.

شبکه های عصبی را از لحاظ نحوه ی اتصال گره ها به دو نوع تقسیم می کنند.

(۱) شبکه های عصبی پیشرو: که در آن گره های هر لایه فقط به گره های لایه بعدی متصل می شوند (شکل ۱).

(۲) شبکه های عصبی پسرو: که در آن گره های هر لایه به گره های لایه های بعدی یا به خودشان متصل می شوند.

در این مطالعه از شبکه های عصبی پیشرو استفاده شده است. یکی از متداول ترین سبک های معماری شبکه های عصبی که در آن الگوی ارتباطی بین لایه ها مشخص می شود، پرسپترون چند لایه (MultilayerPerceptron) است. در این ساختار، هر نورون، تابعی از ترکیب خطی متغیرهای توضیحی، و همچنین متغیر پاسخ نیز تابعی از ترکیب خطی نورون ها است. یعنی:

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

که در آن σ تابع فعال سازی نامیده شده و نمودار این تابع غیر خطی، S- شکل (Sigmoid Activation Function) است که از جمله آنها می توان به تانژانت

جدول شماره ۱- 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 جدول ۱، متغیر "بیماری قلبی" به عنوان متغیر پاسخ (پیش بینی شونده) و بقیه متغیرها به عنوان متغیر توضیحی (پیش بینی کننده) در نظر گرفته می شوند.

آمار توصیفی

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

جدول ۲- آمار توصیفی متغیر "سن"

| کمترین | بیشترین | میانگین | انحراف معیار |
|--------|---------|---------|--------------|
| ۲۹ | ۷۷ | ۵۴٫۴۳ | ۹٫۱۰۹ |

جدول ۳- آمار توصیفی متغیر "جنسیت"

| جنسیت | فراوانی | درصد فراوانی |
|-------|---------|--------------|
| زن | ۸۷ | ۳۲٫۲ |
| مرد | ۱۸۳ | ۶۷٫۸ |
| مجموع | | ۱۰۰ |

(۲۰٪ کل داده‌ها، ۶۶ مراجعه کننده) تقسیم می شود. با استفاده از مجموعه داده مدل ساز عمل مدل سازی مدل مربوطه انجام می شود. در نهایت با استفاده از مجموعه داده آزمون دقت مدل برازش داده شده مورد ارزیابی قرار می گیرد. کلیه مراحل ذکر شده برای برازش مدل در محیط نرم افزار Clementine 14.2 انجام شده است.

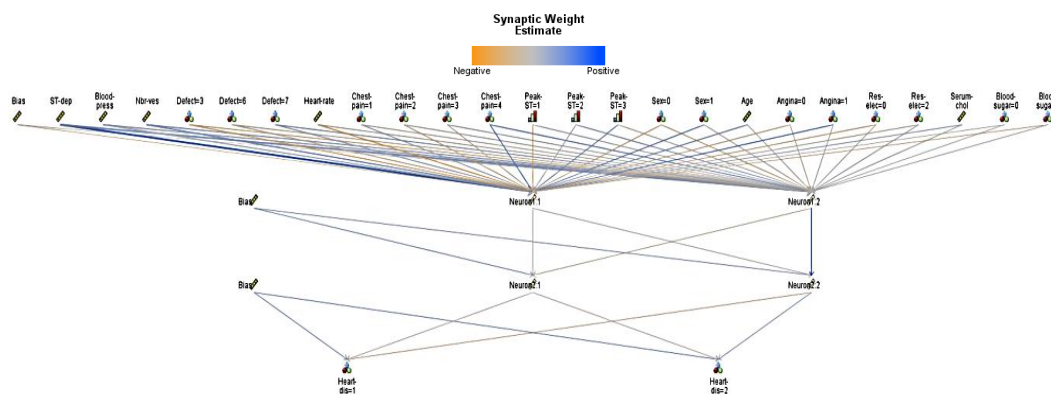
در این بخش با استفاده از مدل شبکه عصبی سعی در پیش بینی ابتلا به بیماری قلبی مراجعه کنندگان به بیمارستان بر اساس اطلاعات مربوط به متغیرهای توضیحی می شود.

در ابتدا مجموعه اصلی داده‌ها که شامل ۲۷۰ مراجعه کننده است به دو قسمت مجموعه داده مدل ساز (۸۰٪ کل داده‌ها، ۲۰۴ مراجعه کننده) و مجموعه داده آزمون

یافته های پژوهش

عصبی با ساختار پرسپترون چندلایه برآزش داده شده، مدلی دارای دو لایه پنهان و در هر لایه پنهان ۲ نرون است.

نتایج برآزش مدل شبکه عصبی با ساختار پرسپترون چندلایه پس از برآزش مدل شبکه عصبی ایجاد شده در شکل ۲ نشان داده شده است. بهترین مدل شبکه



شکل ۲- نمودار مدل شبکه عصبی با دو لایه پنهان و در هر لایه پنهان دو نرون

در لایه های 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 گره های موجود در لایه خروجی به صورت زیر محاسبه می شود.

ضرایب مربوط به اتصال گره ها در لایه های ورودی، پنهان و خروجی در جدول ۴ نشان داده شده است. مثبت یا منفی بودن هر یک از ضرایب از روی شکل ۲ و رنگ اتصال دهنده ها مشخص می شود. با توجه به شکل ۲ و جدول ۴، می توان مقادیر هر یک از گره ها را

جدول ۴- ضرایب اتصال گره ها در لایه های 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 |

| | | | |
|-------|-----|------|---------------|
| 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 |

$$\begin{aligned} \{Neuron\ 1.1\} = & Tanh(-23.07 - (12 * \{Peak - ST = 1\}) + (11.076 * \{Peak - ST = 2\}) \\ & + (10.15 * \{Peak - ST = 3\}) - (20.30 * \{Nbr - ves\}) - (2.76 * \{Serum - chol\}) \\ & + (21.23 * \{Blood - press\}) + (7.38 * \{Age\}) - (16.61 * \{Heart - rate\}) \\ & + (22.15 * \{ST - dep\}) + (19.38 * \{Defect = 3\}) + (18.46 * \{Defect = 6\}) \\ & - (17.52 * \{Defect = 7\}) + (15.69 * \{Chest - pain = 1\}) - (14.76 * \{Chest - pain = 2\}) \\ & - (13.84 * \{Chest - pain = 3\}) + (12.92 * \{Chest - pain = 4\}) - (9.23 * \{Sex = 0\}) \\ & + (8.30 * \{Sex = 1\}) - (6.46 * \{Angina = 0\}) + (5.53 * \{Angina = 1\}) \\ & - (4.61 * \{Res - elec = 0\}) - (3.69 * \{Res - elec = 2\}) - (1.84 * \{Blood - sugar = 0\}) \\ & - (0.92 * \{Blood - sugar = 1\}) \end{aligned}$$

$$\{Neuron\ 2.1\} = Tanh(18 - (12 * Neuron\ 1.1) - (6 * Neuron\ 1.2))$$

$$\{Heart - Dis = 1\} = 18 + (12 * Neuron\ 2.1) - (6 * Neuron\ 2.2)$$

$$\{Heart - Dis = 2\} = 18 - (12 * Neuron\ 2.1) + (6 * Neuron\ 2.2)$$

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 است.

بیماری قلبی مبتلا است. در نهایت با استفاده از شبکه عصبی برآزش داده شده، عمل کلاس بندی برای مجموعه مشاهدات مدل ساز و آزمون انجام شده که نتایج آن به ترتیب در جدول ۵ و ۶ بیان شده است.

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

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

| | | پیش بینی شده | | | درصد صحیح |
|------------|--------------|--------------|--------------|-------------------|-----------|
| | | Heart-dis =1 | Heart-dis =2 | غیر قابل پیش بینی | |
| مشاهده شده | Heart-dis =1 | ۱۰۸ | ۹ | ۱ | %۹۱,۵۲ |
| | Heart-dis =2 | ۱۴ | ۷۱ | ۱ | %۸۲,۵۵ |
| کل | | | | | %۸۷,۷۵ |

با توجه به جدول ۷ می بینیم که مدل "شبکه عصبی با ساختار پرسپترون چند لایه" با دقت ۸۷,۷۵ درصد عمل کلاس بندی را برای مجموعه داده مدل ساز انجام داده است. با توجه به جدول می بینیم که بر اساس مدل، یک مشاهده دارای کلاس متغیر $Heart - dis = 1$ و یک مشاهده دارای کلاس متغیر پاسخ

با توجه به جدول ۷ می بینیم که مدل "شبکه عصبی با ساختار پرسپترون چند لایه" با دقت ۸۷,۷۵ درصد عمل کلاس بندی را برای مجموعه داده مدل ساز انجام داده است. با توجه به جدول می بینیم که بر اساس مدل، یک مشاهده دارای کلاس متغیر $Heart - dis = 1$ و یک مشاهده دارای کلاس متغیر پاسخ

جدول ۶- کلاس بندی متغیر "بیماری قلبی" با استفاده از مدل "شبکه عصبی با ساختار پرسپترون چند لایه" برای مجموعه داده آزمون

| | | پیش بینی 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 |

بطور کلی نحوه کلاس بندی مدل شبکه عصبی با ساختار پرسپترون چند لایه را برای دو مجموعه داده مدل ساز و آزمون در جدول ۷ می توان مشاهده کرد.

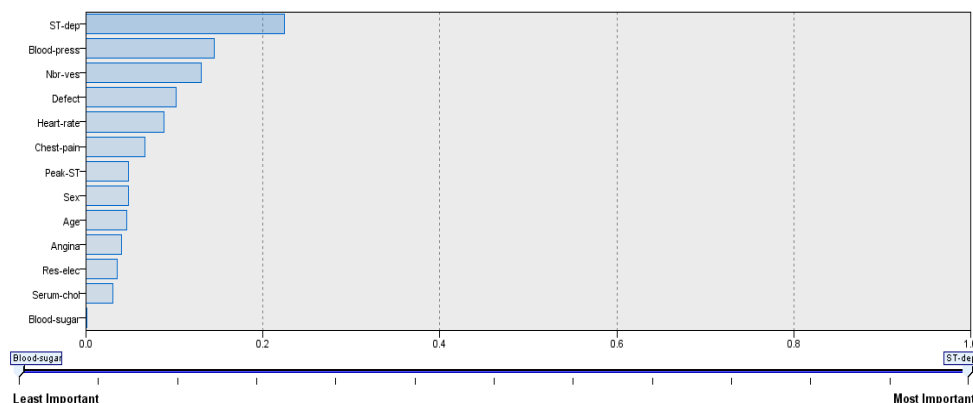
با توجه به جدول ۸ می بینیم که مدل "شبکه عصبی با ساختار پرسپترون چند لایه" با دقت ۸۳,۳۳ درصد عمل کلاس بندی را برای مجموعه داده آزمون انجام داده است.

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

| داده های آزمون | | داده های مدل ساز | | نحوه کلاس بندی متغیر وابسته |
|----------------|-------|------------------|-------|-----------------------------|
| تعداد | درصد | تعداد | درصد | |
| ۵۵ | ۸۳,۳۳ | ۱۷۹ | ۸۷,۷۵ | صحیح |
| ۱۱ | ۱۶,۶۷ | ۲۵ | ۱۲,۲۵ | اشتباه |
| ۶۶ | ۱۰۰ | ۲۰۴ | ۱۰۰ | مجموع |

رتبه بندی اهمیت متغیرها:

در مدل شبکه عصبی با ساختار پرسپترون چند لایه، ترتیب اهمیت متغیرهای توضیحی برای پیش بینی متغیر پاسخ به صورت شکل ۱۰ می باشد.



شکل ۳- نمودار اهمیت متغیرها برای پیش‌بینی متغیر پاسخ در مدل شبکه عصبی با ساختار پرسپترون چند لایه

کاوی سیستم‌های اطلاعات بیمارستانی (HIS) جهت پیش‌بینی دقیق‌تر و تصمیم‌گیری بهتر و مؤثرتر برای درمان بیماران صورت گرفته است. با توجه به برآورد وزن‌های مدل، نهایتاً مدل با استفاده از رابطه زیر عمل کلاس‌بندی را برای رکوردها انجام می‌دهد.

$$\{Heart - Dis = 1\} = 18 + (12 * Neuron 2.1) - (6 * Neuron 2.2)$$

$$\{Heart - Dis = 2\} = 18 - (12 * Neuron 2.1) + (6 * Neuron 2.2)$$

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.

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 یافته‌های 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

با توجه به شکل ۳ 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 بیش‌ترین نیاز را به داده‌کاوی پیدا کرده است و حرکت از پزشکی سنتی به سمت پزشکی مبتنی بر شواهد از جمله مواردی است که می‌تواند مؤکد این امر باشد (۲). تعداد و اندازه پایگاه داده‌های پزشکی به سرعت در حال افزایش است و مدل‌های توسعه یافته تکنیک داده‌کاوی می‌توانند برای پزشکان جهت کمک در تصمیم‌گیری مؤثر و کاربردی باشند (۷). بیمارستان بازوی مهم ارائه خدمات بهداشتی درمانی و اولین سطح ارجاع با قلمرو مسئولیت‌های مشخص است و به این لحاظ مهمترین سازمان بهداشتی درمانی به شمار می‌آید. از یک سو به دلیل اهمیت شیوه‌های تصمیم‌گیری مدیران بیمارستان‌ها در پیشبرد اهداف بیمارستان و اهمیت قدرت پیش‌بینی 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 بیمارستان‌ها از نتایج حاصل از داده

پرسپترون چند لایه" در پیش بینی متغیر پاسخ بیماری قلبی (*Heart-dis*) دارند.

خون (*Blood-press*)، قند خون (*Blood-sugar*) و کلسترول سرم (*Serum-chol*) به ترتیب بیشترین اهمیت را از لحاظ مدل "شبکه عصبی با ساختار

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 J* 1991;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 One* 2017;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 Uni* 2014; 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 Sci* 2008;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 J* 2010;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 Uni 2010. (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.

Heart Disease Forecast Using Neural Network Data Mining Technique

Kazemi M¹, Mehdizadeh M^{2*}, Shiri A³

(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 & conclusions: 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 network, 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.mahdzadeh@ilam.ac.ir