www.MatlabKar.com
سایت متلب کار - آموزش متلب - شبیه سازی با متلب

دومین کنفرانس بین‌المللی ترکیبیات، رمزنگاری و محاسبات
I4C 2017
Homepage: http://i4c.iust.ac.ir
دانشگاه علم و صنعت ایران

ارایه یک روش بهبود یافته موازی در پایگاه داده توزیع شده سیستم های کنترل سلامت پزشکی به کمک داده کاوی و درخت تصمیم

مهدی بردبار [1,*]

۱- کارشناس ارشد کامپیوتر، emehdibordbar@gmail.com

چکیده

حوزه پزشکی و سلامت از بخش های مهم در جوامع صنعتی است. استخراج دانایی از میان حجم انبوه داده های مرتبط با سوابق بیماری و پرونده های پزشکی افراد با استفاده از فرایند داده کاوی می تواند منجر به شناسایی قوانین حاکم بر ایجاد، رشد و تسری بیماری ها شده و اطلاعات ارزشمندی را به منظور شناسایی علل رخداد بیماری ها، تشخیص، پیش بینی و درمان بیماری ها با توجه به عوامل محیطی حاکم در اختیار متخصصان و دست انداران حوزه سلامت قرار دهد. هدف این مقاله انتخاب فیلد های تاثیر گذار در پایگاه داده های برنامه های کاربردی که بر روی سیستم ها مختلف بیمارستان ها و مراکز درمانی که به صورت توزیع شده هستند و همچنین طراحی سیستم های اطلاعاتی به منظور پیش بینی با خطای کمتر برخی بیماری ها با استفاده از ویژگی ها و ارتباطات بین ویژگی های مرتبط با بیماری از طریق ترکیب تکنیک های مختلف داده کاوی و درخت تصمیم می باشد. در مقالات پیشین انجام شده در ایران از همه متغیرهای به وجود آورنده بیماری استفاده نشده و براساس حدس و گمان و یا براساس تحقیقات خارجی انجام شده تعدادی متغیر در نظر گرفته شده و کار برروی آنها صورت گرفته است. اما در این مقاله با جمع بندی تمام متغیرهای شناخته شده دارای اهمیت در تحقیقات گذشته داخلی و خارجی و نیز تایید آنها براساس نظر پزشک متخصص مهمترین متغیرها جمع آوری و براساس آن مدلسازی انجام گرفته است. داده ها مورد استفاده در این مقاله با استفاده از متغیرهای شناخته شده در پایگاه تشخیص بیماری قلبی موجود در مرکز UCI از بیمارستانهای سطح کشور جمع آوری شده است که شامل اطلاعات ۹۹۴ بیمار می باشد این اطلاعات برای نمونه در قالب فایل اکسل با ۱۸ ویژگی جمع آوری شده که فیلد آخر نظر پزشک معالج مبنی بر حمله قلبی یا عدم حمله قلبی است. در این مقاله مدل بدست آمده مبتنی بر درخت تصمیم علاوه بر توانایی بالا در تشخیص افراد بیمار و همچنین کارآمد ترین الگوریتم در تحلیل داده ها به صورت موازی که دارای بیشترین دقت در تشخیص این بیماری می باشد.

کلمات کلیدی: بیماری قلبی، سیستمهای اطلاعاتی،درخت تصمیم، داده کاوی، ماتریس اغتشاش،پردازش موازی،توزیع شده

* **Corresponding author:** کارشناس ارشد کامپیوتر
Email: emehdibordbar@gmail.com

www.MatlabKar.com
سایت متلب کار - آموزش متلب - شبیه سازی با متلب

۱. مقدمه

داده‌کاوی عبارت است از اقتباس یا استخراج دانش از مجموعه ای بسیار حجیم از داده ها، به‌بیان‌دیگر، داده‌کاوی فرایندی است که با استفاده از تکنیک‌های هوشمند، دانش را از مجموعه ای از داده ها استخراج می کند که تحلیل‌های ساده آماری قادر به انجام آن نیستند. داده‌کاوی از الگوریتم های بسیار پیچیده ریاضی جهت تقسیم‌بندی داده‌ها و پیشگویی رویدادها استفاده می‌کند (شهرابی،۱۳۹۲).

امروزه مراکز پزشکی با مقاصد گوناگون 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 قابل‌توجه می شود. بنابراین از داده‌کاوی برای غلبه بر این مشکل و به دست آوردن روابط مفید بین عوامل خطرزا در بیماری ها با توجه به شیوع و سهمی که در مرگ‌ومیر انسان ها دارند استفاده می شوند.

در ابتدای قرن بیستم میلادی ۱۰٪ کل مرگ‌ومیرها به علت بیماری‌های قلبی بود. بیماری‌های قلبی علت اصلی مرگ‌ومیر در طول ۱۰ سال گذشته است. سازمان بهداشت جهانی برآورد کرده است که هرساله ۱۲ میلیون نفر در سراسر جهان جان خود را براثر بیماری قلبی از دست می‌دهند (مای شومن[۱]، ۲۰۱۴)

تشخیص بیماری‌های قلبی یک کار قابل‌توجه و خسته‌کننده در علم پزشکی می‌باشد و وظیفه مهم و کار پیچیده‌ای است که باید با دقت و کارآمدی انجام گیرد. با این‌حال ابزارهایی برای تجزیه‌وتحلیل استخراج داده ها وجود دارد که در دسترس بودن این مجموعه عظیم از داده های پزشکی منجر به تجزیه‌وتحلیل درستی در این زمینه گردیده است.

در اکثر مقالات پیشین انجام شده در جهان از همه متغیرهای به وجود آورنده بیماری استفاده نشده و براساس حدس و گمان و یا براساس دیگر مقالات انجام شده و کار بر روی تعدادی متغیر صورت گرفته است که باعث شده در هنگام ارزیابی،نتایج آنها از دقت بسیار کمتری برخوردار باشند. اما در این مقاله با جمع بندی تمام متغیرهای شناخته شده دارای اهمیت در مقالات گذشته داخلی و خارجی و نیز تایید آنها براساس نظر پزشک متخصص مهمترین متغیرها جمع آوری و براساس آن مدلسازی انجام گرفته است که نتایج آن نیز از دقت بسیار بالایی برخوردار میباشد.

تشخیص پزشکی حوزه ای در علم پزشکی است که با رشد زیاد پایگاههای داده های پزشکی، استفاده از دانش داده کاوی برای کشف دانش از این پایگاهها زیاد شده است. از این رو هدف از این مقاله پیش بینی با خطای کمتر بیماری قلبی با استفاده از ویژگی ها و ارتباطات بین ویژگی های مرتبط با بیماری از طریق ترکیب تکنیک داده کاوی می باشد و همچنین تعیین مهمترین فاکتورهای موثر در ایجاد بیماری قلبی با استفاده از تکنیک های داده کاوی می باشد.

۲. پیشینه تحقیق

۱- زن[۲] (۲۰۰۶) تحقیق بر روی ۳۱۳ داده در دو کلاس طبیعی و بیماران قلبی انجام داد. در این تحقیق جهت شناسایی و پیشگویی حملات قلبی از روش‌های داده‌کاوی غیر از خوشه‌بندی داده‌کاوی استفاده گردید. در این تحقیق با توجه به وجود داده‌های بیماران قلبی از ترکیب الگوریتم‌های هوشمند مصنوعی و الگوریتم C5 استفاده گردید که نتایج خوبی را جهت شناسایی بهتر درنتیجه تشخیص و پیشگویی حملات قلبی داشته است.

---
۱ Mai Shouman
۲ Zhan

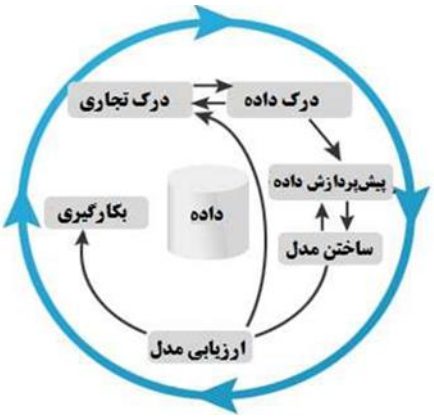
www.MatlabKar.com
سایت متلب کار - آموزش متلب - شبیه سازی با متلب

۲- بلاساندر [۳] (۲۰۱۲) در مقاله تکنیک‌های طبقه‌بندی داده‌کاوی یعنی RIPPER، درخت تصمیم، شبکه‌های عصبی مصنوعی و ماشین بردار پشتیبانی برای پیش‌بینی بیماری‌های قلبی عروقی استفاده‌شده است. این مدل با استفاده از ابزار داده‌کاوی weka نسخه ۳/۶، توسعه داده‌شده است. در آن ۱۴ ویژگی و ۳۱۳ نمونه وجود دارد و درنهایت نتایج به‌دست‌آمده باهم مقایسه شده است. نرخ خطا برای RIPPER، شبکه‌های عصبی مصنوعی، ماشین بردار پشتیبانی و درخت تصمیم‌گیری 0/2756، 0/2248، 0/1588،0/2755 بوده است. دقت RIPPER، شبکه‌های عصبی مصنوعی، ماشین بردار پشتیبانی و درخت تصمیم‌گیری، % 84/08 ، 84/06 %، 84/12% و % 79/05 بود. نویسنده به این نتیجه رسید که ماشین بردار پشتیبانی بهترین روش برای پیش‌بینی بیماری‌های قلبی عروقی است.

۳. فرآیند پیشنهادی براساس استاندارد کریسپ

در چند سال اخیر، استانداردهای گوناگون داده کاوی تحقق‌یافته و امروزه توسط بسیاری از سازندگان برنامه های کاربردی داده کاوی، استفاده می شود. استانداردهای داده کاوی موضوعات گوناگونی را در ارتباط با داده کاوی در برمی‌گیرد. استانداردهای وضع‌شده و به ابعاد مختلف کاربردهای داده کاوی اشاره دارند. یکی از این استانداردها برای فرآیند داده کاوی، استاندارد کریسپ می باشد که در سال ۱۹۹۶ با همکاری شرکت های دایملر کرایسلر (بنز) ، اس پی اس اس و ان سی آر اقدام به تهیه آن شده است. این استاندارد، صنعتی، ابزاری با کاربردهای فراگیر برای تعریف و تصدیق فرآیند داده کاوی تهیه‌شده است.

مدل مرجع داده کاوی در استاندارد کریسپ، یک نمای کلی از چرخه دوام یک پروژه داده کاوی را فراهم می آورد. این مدل شامل مراحل پروژه، وظایف هرکدام و ارتباط بین آنها می باشد. چرخه دوام یک پروژه داده کاوی شامل شش مرحله است که در شکل۱ نشان داده‌شده است. توالی این مراحل انعطاف پذیر می باشد و برگشت به مراحل قبلی و مسیرهای آزاد بین مراحل گاهی موردنیاز است.



شکل ۱ – مراحل انجام یک فرآیندداده‌کاوی

الف) شناخت سیستم:

۳Bala Sundar

www.MatlabKar.com
سایت متلب کار - آموزش متلب - شبیه سازی با متلب

در گام اول با مشورت پزشک متخصص قلب و عروق و نیز با مطالعه برروی بیماری قلبی و تعیین فاکتورهای موثر در ابتلا و همچنین روشهای تشخیصی و درمانی و روشهای پیشگیری از ابتلا به بیماری سعی در شناخت کافی حوزه مورد بررسی داشته است.

ب) مرحله آماده سازی داده ها:

متغیرهای این مطالعه از پایگاه داده های بیماری های قبلی استخراج شد که در این پایگاه، داده هایی از چهار مجموعه داده متفاوت برای تشخیص بیماری های قلبی قرار دارند. متغیرهای فوق از چهار منبع (بنیاد کلنیک کلیولند، انستیتو کاردیولوژی مجارستان، مرکز پزشکی لانگ بیچ کالیفرنیا و بیمارستان دانشگاه زوریخ سویس) بدست آمد. در مجموع در این پایگاهها داده ها با ۷۶ ویژگی یا متغیر مختلف اندازه گیری شده که به طور مستقیم یا غیرمستقیم با بیماریهای قلبی مرتبط بودند. در پیش پردازش داده ها، جهت کم کردن تعداد متغیرها، کاهش زمان پردازش و اجرای مدل های داده کاوی و بیو انفورماتیک، ۱۸ ویژگی مهم انتخاب شدند. متغیر هدف در این مطالعه وجود یا عدم وجود بیماری قلبی بود که در مورد هرکدام از افراد مورد بررسی یکی از این دو حالت ثبت گردید. (مقدار متغیر هدف برابر با یک نشان دهنده وجود بیماری قلبی و صفر، نشان دهنده عدم وجود بیماری می باشد)

داده های استفاده شده در این مطالعه مربوط به پرونده بیماران مراجعه کننده به بیمارستانها می باشد که در سالهای ۱۳۹۳ تا ۱۳۹۴ جمع آوری شده است. تعداد بیماران مراجعه کننده ۲۷۶۸ نفر است که تعداد ۹۹۴ رکورد از مجموع رکوردهای جمع آوری شده قابل استفاده بود و بقیه رکوردها حذف گردید.

ج) مدلسازی:

روشهای داده کاوی متنوعی برای مدل سازی وجود دارد. در این مرحله با استفاده از تکنیکهای داده کاوی به ارائه مدل می پردازیم. مدل سازی با استفاده از نرم افزار کلمنتاین انجام می شود. در این مرحله از الگوریتم درخت تصمیم شامل الگوریتم C ۵/۰ با بکارگیری متغیرهای ورودی و تعیین متغیر هدف استفاده می شود.

د) ارزیابی:

در این مرحله پس از ایجاد مدل بایستی به ارزیابی مدل ایجاد شده بپردازیم. برای صحت مدل داده ها به دو دسته آموزش (۸۰٪) و آزمون (۲۰٪) تقسیم می شوند. داده های بخش آموزش مدل را می سازند و داده های بخش آزمون مدل ایجاد شده را مورد ارزیابی قرار می دهند. جهت ارزیابی مدل ها می توان از شاخص های حساسیت ، ویژگی، دقت ، ارزش اخباری مثبت و ارزش اخباری منفی استفاده کرد. در این تحقیق برای ارزیابی مدل از ماتریس اغتشاش استفاده می شود.

دومین کنفرانس بین‌المللی ترکیبیات، رمزنگاری و محاسبات
Homepage: http://i4c.iust.ac.ir
2017

$$\text{حساسیت} = \frac{\text{تعداد داده های برچسب مثبتی که درست دسته بندی شده اند}}{\text{کل تعداد داده های مثبت}}$$

$$\text{شفافیت} = \frac{\text{تعداد داده های برچسب منفی که درست دسته بندی شده اند}}{\text{کل تعداد داده های منفی}}$$

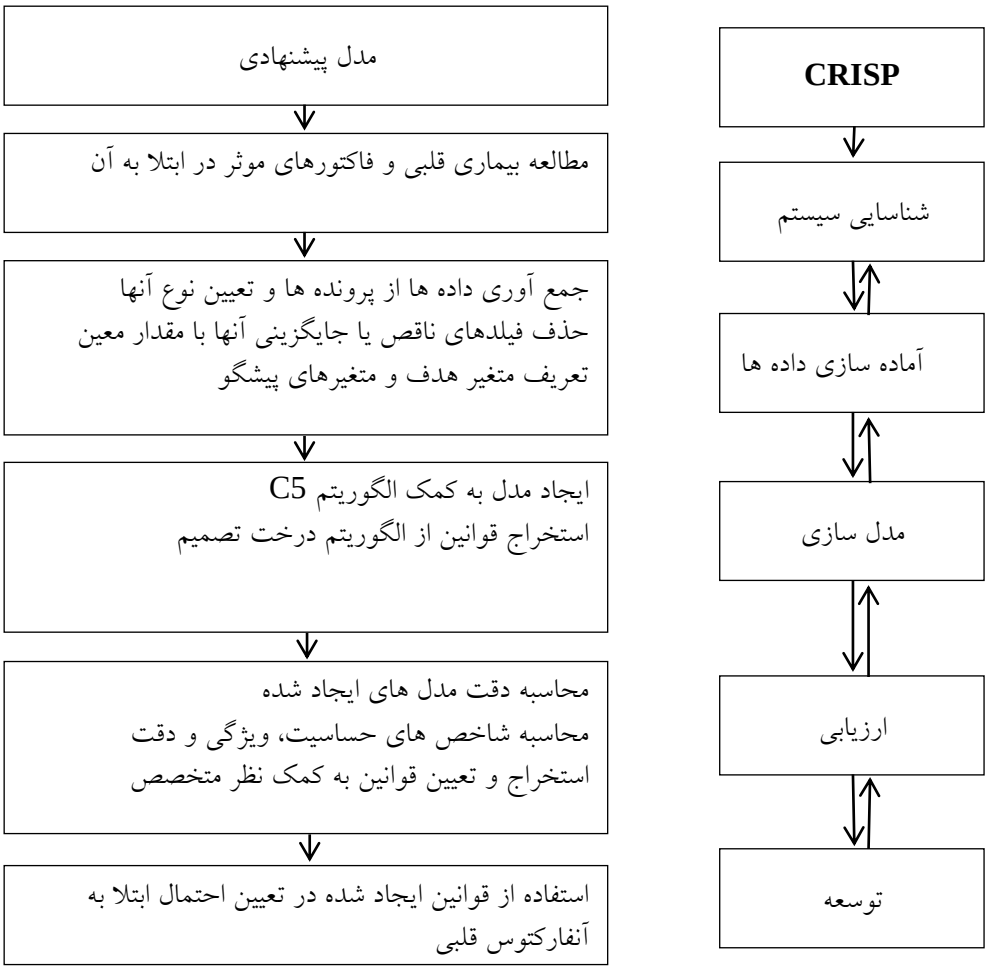
$$\text{دقت} = \frac{\text{تعداد داده های برچسب منفی که درست دسته بندی شده اند}}{\text{تعداد داده های برچسب مثبتی که درست دسته بندی شده اند + تعداد داده های برچسب منفی که به نادرست مثبت دسته بندی شده اند}}$$

$$\text{صحت} = \text{حساسیت} \frac{\text{تعداد داده ها مثبت}}{\text{تعداد کل داده ها}} + \text{شفافیت} \frac{\text{تعداد داده ها منفی}}{\text{تعداد کل داده ها}}$$

هـ) توسعه:

ساخت مدل معمولا پایان پروژه نیست حتی اگر هدف مدل افزایش دانش است، دانش بدست آمده نیاز به سازماندهی دارد و نمایش به طریقی که کاربر بتواند از آن استفاده کند. دانش کشف شده باید سازماندهی شده و به شکل قابل ارائه برای دیگران درآید. ما در این مرحله از ایجاد گزارشات لازم سعی می کنیم توضیح دهیم که براساس مدل ایجاد شده تاثیر گذارترین فاکتورها در ابتلا فرد به آنفارکتوس قلبی کدامند. توجه به اینکه بیماریهای قلبی عروقی از جمله شایع ترین بیماری ها و علل مرگ محسوب می شوند، چنانچه بتوانیم یک مجموعه پرخطر را شناسایی و برنامه های غربالگری را برای آن اجرا کنیم کارایی برنامه بیشتر خواهد شد. علاوه بر این افراد یک جمعیت پرخطر ممکن است بیشتر راغب به شرکت در برنامه های غربالگری باشند به خصوص هنگامی که نتیجه تست (مدل ارائه شده) بر روی آنها مثبت باشد احتمال دارد بیشتر به توصیه های پزشکی گوش فرا دهند. پس از اجرای الگوریتم ها قوانین بدست آمده نتایج به متخصص مورد نظر ارائه شده و قوانینی که از نظر بالینی معتبر باشند به عنوان قوانین نهایی ارائه می گردد.

www.MatlabKar.com
سایت متلب کار - آموزش متلب - شبیه سازی با متلب



شکل ۲: متدولوژی کریسپ و چارچوب استفاده شده در این مقاله

۴. درخت تصمیم C ۵/۰

الگوریتم C ۵/۰ یک نوع درخت تصمیم گیری تک متغیره و بهبود یافته الگوریتم C ۴/۵ است که توسط محقق استرالیایی کوئین لن در سال ۱۹۹۳ طراحی شد. این الگوریتم مشابه الگوریتم CART ابتدا درختی کامل پر ایجاد می کند ولی استراتژی هرس آن کاملا متفاوت است. این الگوریتم کلاسه بندی را با تقسیم داده ها به زیرمجموعه هایی که شامل رکوردهای همگن تر از والد خود هستند انجام می دهد. در C ۵/۰ تقسیم کردن نمونه ها براساس فیلدی که

www.MatlabKar.com
سایت متلب کار - آموزش متلب - شبیه سازی با متلب

بیشترین بهره اطلاعات را دارد صورت می گیرد. هرزیر نمونه توسط اولین انشعاب تعیین می شود. سپس معمولا براساس فیلدی جدید مجددا تقسیم بندی انجام می گیرد و این فرایند به دفعات تکرار می شود تا اینکه زیر نمونه ها قابلیت تقسیم شدن را نداشته باشند. سرانجام انشعابها پایین ترین سطح از نوآزموده می شوند و آن انشعابهایی که ارزش چندانی ندارند از مدل حذف می شوند. استنتاج قانون از طریق الگوریتم C ۵/۰ براساس درخت تصمیم گیری است. این احتمال وجود دارد که بیش از یک قانون به ازای هر رکورد خاص صدق کند و یا هیچ قانونی به کار نرود. اگر چندین قانون برای یک رکورد مناسب باشند هر قانون مبتنی بر اطمینان مربوط به هر قانون، وزنی تحت عنوان vote می گیرد در اینصورت براساس ترکیب وزن همه قوانین مناسب برای رکورد، پیش بینی نهایی تعیین می شود و اگر هیچ قانونی مناسب نباشند یک پیش گویی پیش فرض به آن رکورد نسبت داده می شود.

۵. پایگاه داده توزیع شده:

یک پایگاه داده توزیع شده به طور کامل در یک محل فیزیکی مجرد ذخیره نشده است . درعوض در یک شبکه از کامپیوترها که از لحاظ جغرافیایی پراکنده شده اند و از طریق لینک های اتصالی به هم متصل اند پراکنده شده اند. داده های مربوطه به این تحقیق براساس داده های ذخیره شده در پایگاه داده های مختلف کلینیک ها و بیمارستانها که باید به صورت توزیع شده پس از اجرای فازهای آماده سازی داده ها جمع آوری می شوند که در اینجا به صورت نمونه این داده ها شامل اطلاعات ۹۹۴ بیمار می باشد که این اطلاعات در قالب فایل اکسل با ۱۸ ویژگی جمع آوری شده که فیلد آخر نظر پزشک معالج مبنی بر حمله قلبی یا عدم حمله قلبی این بیماران می باشد.

۶. متغیرهای تحقیق

پایگاه تشخیص بیماری قلبی موجود در مرکز UCI نتیجه تلاشهای انجام شده درخصوص بیماری قلبی مربوط به منطقه کلیوند (بنیاد کلنیک کلیولند، انستیتو کاردیولوژی مجارستان، مرکز پزشکی لانگ بیچ کالیفرنیا و بیمارستان دانشگاه زوریخ سویس) بوده است. این پایگاه دارای ۷۶ متغیر می باشد که همه این ۷۶ متغیر مورد استفاده نخواهد بود و فقط ۱۸ مورد از آنها مفید است که در ادامه ذکر خواهد شد.

۱. سن بیمار :

۲.جنسیت بیمار: ۱-مرد ۰-زن

۳.نوع درد قفسه سینه: ۱- آنژین معمولی ۲- آنژین غیرمعمولی ۳- درد غیرآنژینی ۴- بدون علامت

۴.فشار خون در حالت استراحت: در زمان ورود به بیمارستان برحسب میلیمتر جیوه

۵.کلسترول سرم خون : کلسترول موجود در سرم خون

۶.قند خون ناشتا: ۱- بالاتر از ml/dl ۱۲۰ ۰ -پایین تر از ml/dl ۱۲۰

۷.نتایج الکتروکاردیوگرافی: ۰- نرمال ۱- موج ST-T غیرمعمول (موج T وارونه باشد و یا ST بالاتر از ۰/۰۵ باشد) ۲ - کوچک شدن بطن چپ

۸حداکثر ضربان قلب: بیشترین میزان ثبت شده برای ضربان قلب

۹.آنژین القا شده: ۱- بله ۰- خیر

۱۰.افسردگی ST القا شده (ST: پایین آمدن موج ST در ورزش نسبت به حالت استراحت) :صفر تا ۳

۱۱.شیب قله موجST (شیب پیک بخشST) :۱ -شیب بالا ۲- صاف ۳- شیب پایین

دومین کنفرانس بین‌المللی ترکیبیات، رمزنگاری و محاسبات
Homepage: http://i4c.iust.ac.ir
2017

۱۲.تعداد عروق بزرگ مشخص شده توسط فلوروسکوپی(تعداد رگهای اصلی): صفر تا ۳

۱۳.نتیجه آنژیوگرافی: ۳-نرمال ۶- نقص ثابت ۷- نقص برگشت پذیر

۱۴.سابقه مصرف سیگار: ۰-ندارد ۱-دارد

۱۵.سابقه خانوادگی بیماری قلبی: ۰-ندارد ۱-دارد

۱۶.سابقه اعتیاد: ۰-ندارد ۱-دارد

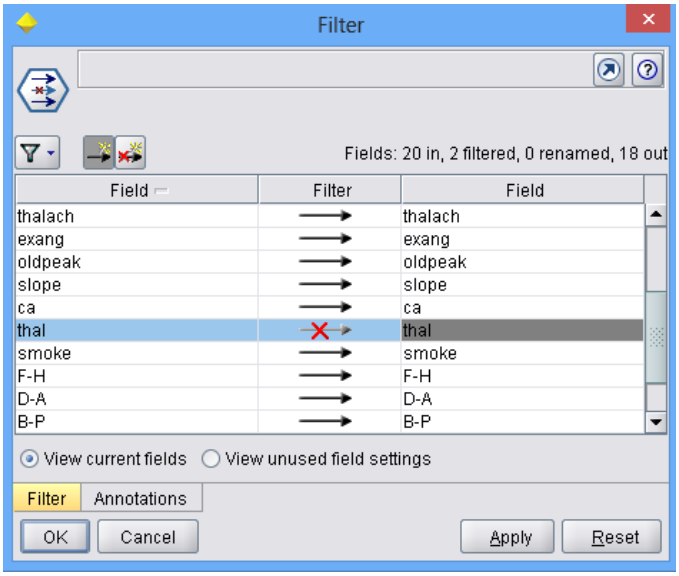
۱۷.سابقه فشار خون بالا: ۰-ندارد ۱-دارد

۱۸.سابقه بیماری دیابت: ۰-ندارد ۱-دارد

۱۹.تشخیص بیماری قلبی: ۰- باریک شدن قطر بیشتر از ۵۰٪ (عدم حمله قلبی) ۱- باریک شدن قطر کمتر از ۵۰٪ (خطرحمله قلبی)

۷. پیاده سازی

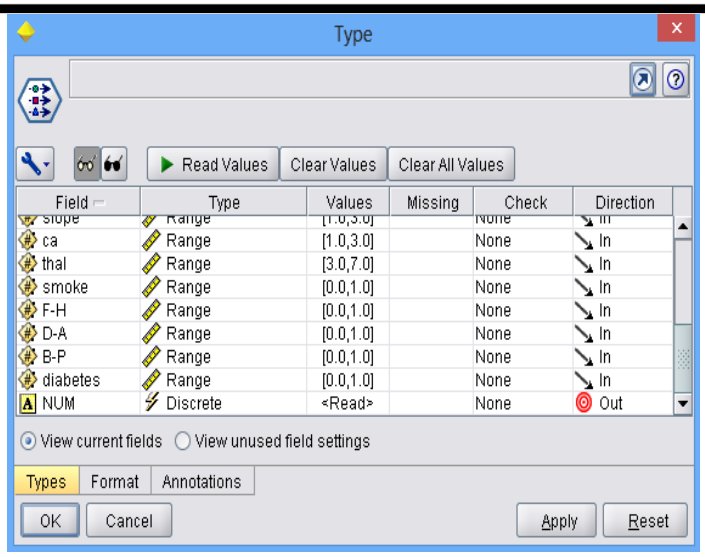
ابتدا فایل اکسل پایگاه داده ای مورد نظر را فراخوانی می کنیم. سپس ما می توانیم در مورد داده های ورودی پیش پردازش انجام دهیم. این پیش‌پردازش‌ها در مرحله اول شامل حذف داده‌های نامربوط، انتخاب و فیلتر کردن بعضی از داده‌ها و فیلدها، پرکردن داده‌های خالی و فیلدهای از بین رفته و نرمال‌سازی داده‌ها می‌باشد که در شکل زیر نمونه‌ای از کار صورت گرفته با استفاده از نرم‌افزار کلمنتاین نمایش داده شده است.



شکل 3: آماده سازی داده ها و حذف داده های کم اهمیت تر

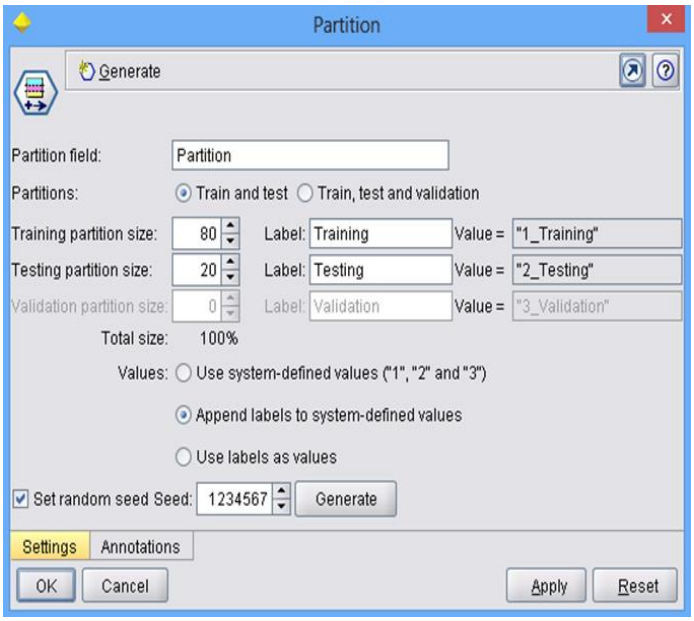
در این قسمت فیلدهایی که از اهمیت کمتری برخوردار است حذف می گردد. ازجمله این فیلدها thal (نتیجه آنژیوگرافی) اشاره کرد.

دومین کنفرانس بین‌المللی ترکیبیات، رمزنگاری و محاسبات
Homepage: http://i4c.iust.ac.ir
I4C 2017
دانشگاه علم و صنعت ایران



شکل 4: تعیین داده های ورودی و خروجی

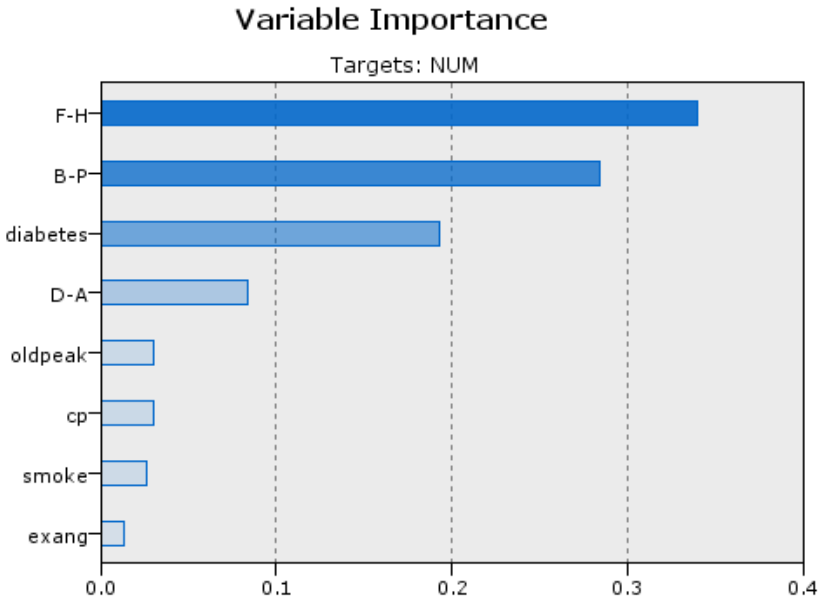
تمام متغیرها به عنوان ورودی و فیلد NUM که مشخص کننده بیماری یا عدم بیماری است به عنوان خروجی در نظر گرفته می شود.



شکل5: تعیین میزان داده ها جهت یادگیری و تست

همانطور که مشخص است ۸۰ درصد از داده ها جهت یادگیری و ۲۰ درصد به عنوان داده های تست در نظر گرفته شده است.

۸. توضیح تکنیک درخت تصمیم C ۵/۰



شکل6: مشخصه های مهم پس از اعمال تکنیک *C5/0*

جدول۱: نتایج تحلیل داده‌ها براساس تکنیک C5/0

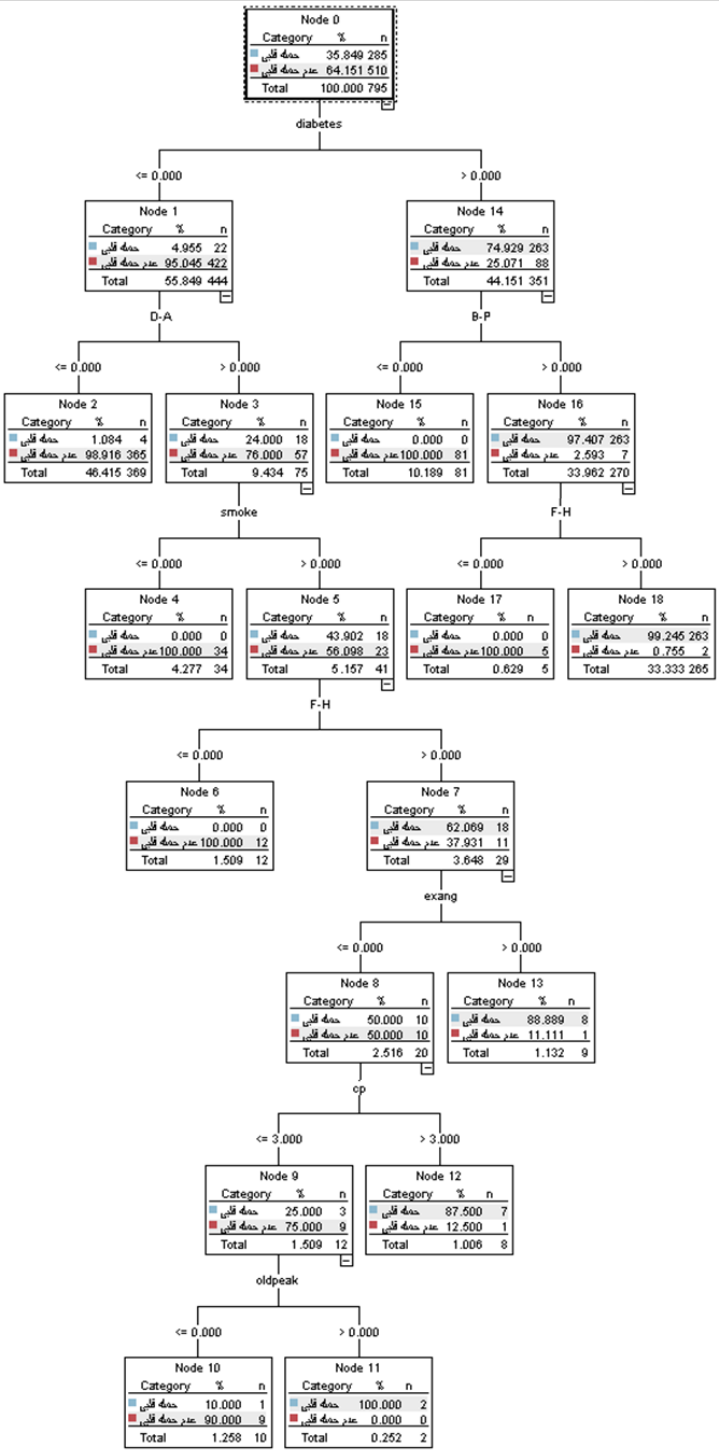
رتبه	نام فیلد	توضیحات
۱	F-H	سابقه خانوادگی بیماری قلبی
۲	B-P	سابقه فشار خون بالا
۳	diabetes	سابقه بیماری دیابت
۴	D-A	سابقه اعتیاد
۵	old peak	القا شدهSTافسردگی
۶	CP	نوع درد قفسه سینه

نتایجC5/0 نشان می دهد که فیلد F-H سابقه خانوادگی بیماری قلبی بیشترین اهمیت و تاثیر گذاری را دارا می باشد .

نمونه ای از قانون ایجاد شده توسط این درخت به صورت زیر است:

اگر شخص سابقه خانوادگی بیماری قلبی داشته باشد و فشار خون بالا و دیابت داشته باشد و همچنین سابقه اعتیاد نیز داشته باشد آنگاه احتمال گرفتاری به بیماری قلبی بالاست.

دومین کنفرانس بین‌المللی ترکیبیات، رمزنگاری و محاسبات
Homepage: http://i4c.iust.ac.ir
دانشگاه علم و صنعت ایران
I4C 2017



شکل7: ساختار درختی ایجاد شده پس از اعمال تکنیک C5/0

دومین کنفرانس بین‌المللی ترکیبیات، رمزنگاری و محاسبات
Homepage: http://i4c.iust.ac.ir
دانشگاه علم و صنعت ایران
I4C 2017

۹. ارزیابی مدل پیش‌بینی

ابزارهای مختلفی برای ارزیابی مدل‌ها می‌توان در نظر گرفت که در ادامه به معیار ماتریس اعتشاش (ماتریس رخداد) اشاره خواهد شد:

1-9. معیار ماتریس اغتشاش

ماتریس اغتشاش یا ماتریس رخداد، یک ابزار بصری جهت نمایش دقت کلاسه‌بندی می‌باشد که جهت نمایش رابطه بین نتایج وکلاس‌های پیش‌بینی استفاده می‌گردد.

جدول ۲: ماتریس اغتشاش

		کلاس پیش بینی[۱]	
کلاس واقعی[۲]		Class0.0	Class1.0
	Class0.0	A (TP)	B (FN)
	Class1.0	C (FP)	D (TN)

که درآن:

تعداد پیش‌بینی‌های صحیح درکلاس A : TP

تعداد پیش‌بینی‌های نادرست درکلاس B : FN

تعداد پیش‌بینی‌های نادرست درکلاس C : FP

تعداد پیش‌بینی‌های صحیح درکلاس D : TN

برمبنای جدول فوق مقادیر دقت[۳]، صحت[۴]، نرخ خطا وحساسیت[۵] که از ملاک‌های ارزیابی الگوریتم‌ها تلقی می‌شوند، طبق فرمول‌های زیر حاصل می‌گردد.

دقت: عبارت است از تعداد نمونه هایی که به درستی تشخیص داده می شوند نسبت به کل نمونه ها

$$Accuracy = \frac{a+d}{a+b+c+d} = \frac{TP+TN}{TP+TN+FP+FN} \quad (۱)$$

[۱] Predicted Class
[۲] Actual Class
[۳] Precision
[۴] Accuracy
[۵] Sensivity

دومین کنفرانس بین‌المللی ترکیبیات، رمزنگاری و محاسبات
Homepage: http://i4c.iust.ac.ir
2017
دانشگاه علم و صنعت ایران

حساسیت: احتمال پیش بینی درست عود توسط الگوریتم ها (مثبت واقعی تقسیم بر منفی کاذب+ مثبت واقعی)

$$Sensivity = \frac{a}{a+d} = \frac{TP}{TP+FN} \qquad (۲)$$

ویژگی: احتمال پیش بینی درست عدم عود توسط الگوریتم ها (منفی واقعی تقسیم برمثبت کاذب+ منفی واقعی)

$$Specificity = \frac{d}{b+c} = \frac{TN}{TN+FP} \qquad (۳)$$

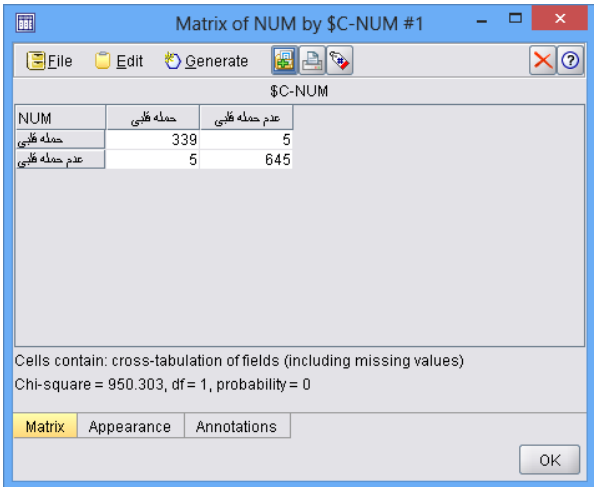
نرخ خطا: عبارت است از تعداد نمونه هایی که به درستی تشخیص داده نشده نسبت به کل نمونه ها

$$Error = \frac{b+c}{a+b+c+d} = \frac{FN+FP}{TP+TN+FP+FN} \qquad (٤)$$

به عنوان نمونه در اینجا با استفاده از نرم‌افزار کلمنتاین ابتدا ماتریس اغتشاش مربوط به الگوریتم پیشنهادی رابه دست آورده و سپس مقادیرصحت و خطا را محاسبه خواهیم کرد.
در تمامی مدل‌ها ۸۰ درصد از داده‌ها برای آموزش و ۲۰ درصد برای تست در نظر گرفته شده‌اند.

2-9. نتایج پیاده‌سازی

بعد از پیاده‌سازی پیشنهادی در نرم افزار کلمنتاین اشکال و جداول زیر که بیانگر دقت و نتیجه کار می‌باشد را بیان می کنیم .



شکل 8: ماتریس اغتشاش پس از اعمال تکنیک C5/0

$$Accuracy = \frac{۳۳۹ + ۶۴۵}{۳۳۹ + ۵ + ۵ + ۶۴۵} = ۹۸.۹۹ \ \%$$

$$Error = \frac{۵ + ۵}{۳۳۹ + ۵ + ۵ + ۶۴۵} = ۱.۰۱ \ \%$$

در شکل فوق نتایج ماحصل از پیاده سازی برای داده های آموزش و تست به صورت جداگانه نمایش داده شده است و نیز همچنین ابتدا عملکرد درخت تصمیم و شبکه عصبی به صورت جداگانه در نظر گرفته شده است.

Results for output field NUM
Comparing $C-NUM with NUM

'Partition'	1_Training		2_Testing	
Correct	786	98.87%	198	99.5%
Wrong	9	1.13%	1	0.5%
Total	795		199	

شکل 9: نتایج پیاده‌سازی اعمال تکنیک *C5/0*

10. مزایای روش‌های پیشنهادی

۱- به طور کلی می توان گفت که مزایای اصلی روش‌های پیشنهادی بالا بودن دقت تشخیص بیماری و کاربردی بودن پایگاه داده مورد استفاده دانست.

۲ - استفاده از الگوریتم های داده کاوی می تواند سیستم های نوین و با صرفه تر در نظام سلامت و درمان ارائه کد و با دقت بالایی قادر به تشخیص بیماری می باشد.

۳ - همچنین مشابه نبودن پایگاه داده این مقاله در مقایسه با سایر مقالات

۴ -استفاده از داده های واقعی برای رسیدن به نتیجه مطلوب تر

۵- استفاده از تمام متغیرهای مربوط به بیماری برای اجرای مدل و رسیدن به نتیجه مطلوب تر

۶- یکی از مشکلات اساسی مربوط به این بیماری عدم تشخیص بموقع و صحیح آن می باشد. امروزه پزشکان بیش از هر چیز با تکیه بر تجربیات و دانسته های خود، آزمایشات پیچیده و وقت گیر به این بیماری پی می برند. با این وجود خطاهای انسانی اجتناب ناپذیر است.

۷- جمع آوری اطلاعات از مراکز درمانی مختلف.

۸-استفاده از الگوریتم C5/0 برای کارایی بیشتر در پردازش داده‌ها.

۹- دستیابی بهتر به داده‌ها در پایگاه داده های مختلف بیمارستانها و مراکز داده.

11. پیشنهادات برای تحقیقات آتی

۱- پیشنهاد به دانشگاههای علوم پزشکی برای ملزم کردن برنامه نویسان نرم افزار اطلاعات بیمارستانی و نیز پرسنل بیمارستانها برای وارد کردن اطلاعات کامل پرونده بیماران در نرم افزار. با انجام این عمل اطلاعات کامل تری از گذشته بیمار در دسترس خواهد بود و دیگر نیازی به مراجعه به پرونده نیست.

۲- می توان با استفاده از نتایج به دست آمده از این پروژه و مشخص شدن ویژگی های پر اهمیت در تشخیص بیماری، نرم افزار کاربردی و ساده ای برای استفاده افراد عادی جامعه طراحی گردد تا هر فرد با وارد کردن نتایج آزمایشهای

www.MatlabKar.com سایت متلب کار - آموزش متلب - شبیه سازی با متلب

پزشکی خود بتواند به سرعت و بدون نیاز پزشک تشخیص دهد که به این بیماری دچار خواهد شد یا خیر و احتمال ابتلا چه قدر خواهد بود.

۳- استفاده از پایگاه داده سایر کشورهای همسایه برای پی بردن به یک نتیجه واحد.

۴- اشتراک داده ها به کمک شبکه کردن سیستم های مختلف اطلاعاتی درمانگاه ها و مراکز درمانی.

۵- سازگاری و هماهنگی داده های سازمانهای پزشکی و مراکز درمانی.

12. نتیجه گیری

امروزه در دانش پزشکی جمع آوری داده های فراوان در مورد بیماری های مختلف از اهمیت فراوانی برخوردار است. 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. تشخیص بیماری‌های یک کار قابل‌توجه و خسته‌کننده در علم پزشکی می‌باشد و وظیفه مهم، اما کار پیچیده‌ای است که باید با دقت و کارآمدی انجام گیرد. بااین‌حال ابزارهایی برای تجزیه‌وتحلیل استخراج داده ها وجود دارد که در دسترس بودن این مجموعه عظیم از داده های پزشکی منجر به تجزیه‌وتحلیل درستی در این زمینه گردیده است.داده کاوی در مدیریت مراقبت های بهداشتی باتوجه به این دلیل که اطلاعات موجود در این رشته طبیعتاً ناهمگن هستند و مجموعه ای از محدودیت های اخلاقی، حقوقی واجتماعی برای محرمانه نگه داشتن اطلاعات پزشکی اعمال می شود، مشابه با زمینه های دیگر نیست. اخیرا، استفاده از دانش و تجربه متخصصان مختلف و داده های آزمایش های بالینی بیماران جمع آوری شده دریک پایگاه داده در تمامی روش های تشخیص در همه جا به رسمیت شناخته شده است.در این مقاله، یک روش داده کاوی برای پیش بینی بیماری قلبی ارائه شد. با بهره گیری از تکنیک درخت تصمیم C5/0با دقت پیش بینی ۹۸,۹۹ درصد توانستیم این بیماری با بالاترین دقت در مقایسه با کارهای انجام شده قبلی را پیش بینی کنیم.

۱۳. مراجع

۱. سعیدی، وحید، یوسفیان، علیرضا، شهرابی، جمال، (۱۳۹۲)، تشخیص بیماری قلبی با استفاده از تکنیکهای داده کاوی، پنجمین کنفرانس داده کاوی ایران، دانشگاه امیرکبیر.

2. Mai Shouman, Tim Turner, Rob Stock,(2014)." Integrating Decision Tree and K-Means Clustering with Different Initial Centroid Selection Methods in the Diagnosis of Heart Disease Patients ", Conference: International Conference of Data Mining & Knowledge Management Process (CKDP), At Dubai, UAE, Aug 06, 2014.Atluri, S.N. and Shen, S. (2002),"*The Meshless Local Petrov–Galerkin (MLPG) Method",* Tech Science Press, USA.

3. Zhan.Y, Chen.H and Zhang.G.C,(2006). " An optimization Algorithm of K-NN classification " , Proceedings of the fifth International conference on Machin Learning and Cybernetics, Dalian, 13-16 August 2006.

4. T. D. Bala Sundar V, N SARAVANAN,(2012). "Development of a Data Clustering Algorithm for Predicting Heart," International Journal of Computer Applications (0975 - 888),vol. 48– No.7, pp. 8-13.