

مدل پیش بینی بیماری های عروق کرونر قلب با استفاده از داده کاوی داده های مطالعه سلامت مردم یزد (یاس)

سید محمدرضا طباطبائی ندوشن^۱، فاطمه سعادت جو^۲، مسعود میرزایی^{۳*}

مقاله پژوهشی

مقدمه: بیماری های ایسکمیک قلبی یکی از شایع ترین بیماری هایی است که آمار مرگ و میر بالایی را در جهان به خود اختصاص می دهد. بیماری های ایسکمیک قلب به دنبال تنگ شدن یا بسته شدن شریان های کرونر قلب به وجود می آیند که تامین کننده خون قلب هستند، این امر به آهستگی و به مرور زمان رخ می دهد. شناسایی افراد مستعد به بیماری و تغییر در الگوی زندگی آن ها می تواند سبب کاهش مرگ و میر و باعث افزایش طول عمر گردد.

روش بررسی: مطالعه سلامت مردم یزد (یاس) به بررسی سلامت و بیماری های نمونه تصادفی ۱۰۰۰۰ نفر از مردم یزد در سال های ۹۴-۱۳۹۳ پرداخته است. این داده ها به علت داشتن ماهیت نامتوازن ابتدا، توسط روش بوت استرپ (Bootstrap) متوازن سازی شده، آنگاه در مرحله آموزش دسته بندها به کار برده شده اند. از دسته بندهای متفاوتی از قبیل شبکه عصبی مصنوعی (Artificial Neural Network)، القای قواعد (Rule Inducer)، رگرسیون (Regression) و آدابوست (Adaboost) جهت ارزیابی روش پیشنهادی با دو سناریو استفاده شده است.

نتایج: یافته ها نشان داد که عمل غربالگری افراد مستعد به بیماری های ایسکمیک قلبی با استفاده از تولید نمونه افراد بیمار به روش بوت استرپ و متوازن سازی داده ها امکان پذیر است. این روش بیشترین تاثیر را در افزایش حساسیت دسته بند کشف زیرگروه CN2 دارد. این دسته بند توانایی تشخیص ۸۳/۶٪ از افراد مستعد بیماری را داراست.

نتیجه گیری: بنابراین می توان نتیجه گرفت که روش های داده کاوی در غربالگری افراد مستعد بیماری ایسکمیک قلبی کارایی مناسبی دارد و به کمک آن می توان افراد مستعد این بیماری را نسبت به غربالگری سنتی که با مراجعه حضوری افراد به پزشک انجام می شود؛ سریع تر و با هزینه کمتر شناسایی نمود.

واژه های کلیدی: داده کاوی، پایش سلامت، پیش بینی، بیماری های ایسکمیک قلب، متوازن سازی داده، القای قواعد CN2-SD

ارجاع: طباطبائی سید محمدرضا، سعادت جو فاطمه، میرزایی مسعود. مدل پیش بینی بیماری های عروق کرونر قلب با استفاده از داده کاوی داده های مطالعه سلامت مردم یزد (یاس). مجله علمی پژوهشی دانشگاه علوم پزشکی شهید صدوقی یزد ۱۳۹۸؛ ۲۷ (۳): ۶۰-۱۳۴۶.

۱- گروه مهندسی کامپیوتر، دانشگاه علم و هنر، یزد، ایران.

۲- گروه مهندسی کامپیوتر، دانشگاه علم و هنر، یزد، ایران.

۳- مرکز تحقیقات قلب و عروق دانشگاه علوم پزشکی خدمات بهداشتی درمانی شهید صدوقی یزد، یزد، ایران.

* (نویسنده مسئول): تلفن: ۰۹۱۳۴۵۰۹۹۱۷، پست الکترونیکی: masoudmirzaei@yahoo.com، صندوق پستی: ۸۹۱۶۹۷۸۴۷۷

مقدمه

قلب بخشی از بدن است که ادامه حیات انسان به کارایی آن بستگی دارد و در صورت عملکرد نامناسب قلب، بخش های دیگر بدن مانند مغز و کلیه و غیره تحت تأثیر قرار می گیرند. در این حالت گردش خون برای بدن کافی نیست و نارسایی خون به مغز باعث سکته مغزی Stroke و نارسایی خون به قلب باعث حمله قلبی Heart Attack یا سایر بیماری های ایسکمیک قلب می شود. بیماری های ایسکمیک قلب (CHD) Coronary Heart Disease یکی از بزرگ ترین عوامل مرگ و میر در سراسر جهان به شمار می روند (۱).

در علم پزشکی تعیین یک محدوده مشخص برای سالم و یا بیمار بودن شخص وجود ندارد. به همین دلیل تصمیم گیری و تشخیص در بسیاری از موارد مبهم و غیرقطعی انجام می شود. در این حرفه، پزشک نمی تواند به راحتی تمام فاکتورهای مؤثر در بیماری را در نظر بگیرد. بنابراین پزشک به ابزار مناسبی نیاز دارد که همه این فاکتورها را در کنار هم در نظر گرفته و در شرایط غیرقطعی و مبهم بیمار، تشخیص قطعی را پیشنهاد دهد (۲). با پیشرفت ابزارهای پزشکی امکان اندازه گیری و ثبت اطلاعات بیمار به صورت دقیق به وجود آمده و توسعه فناوری اطلاعات و ارتباطات امکان ذخیره سازی و آنالیز حجم انبوهی از این داده ها را فراهم آورده است. این داده ها توسط الگوریتم های دسته بند و بنا به نیاز کاربران قابل تحلیل است و این تحلیل به ارتقاء سلامت مردم کمک می کند. از تحلیل این داده ها می توان الگوریتم های تشخیصی استخراج کرد که نمونه ای از آن در این مقاله آورده شده است. این داده ها هم اکنون در پرونده الکترونیک سلامت ذخیره و نگهداری می شود.

این دانش ها می توانند در زمینه های مختلف مانند پیشگیری، تشخیص، رویکرد درمانی و یا پیش بینی (پروگنوز) بیماری ها مفید باشند. طیف بیماری هایی که از این دانش ها سود می برند بسیار گسترده است. یکی از بیماری هایی که امید می رود با استفاده از این دانش در تشخیص و رویکرد درمانی آن تحول ایجاد شود بیماری های ایسکمیک قلبی عروقی

است. آنژیوگرافی، آزمایش استاندارد طلایی برای تشخیص بیماری های کرونری قلب است. اما این آزمایش تهاجمی بوده و بعضاً دارای عوارض جانبی می باشد. لذا رسیدن به روش هایی که بتواند با هزینه کمتر و بدون انجام تست های تهاجمی امکان تشخیص با حساسیت و ویژگی مطلوب را در دسترس قرار دهد تحولی مهم در این عرصه محسوب می شود. با استفاده از داده های جمع آوری شده در مطالعات بزرگ و تحلیل آن ها با توجه به فاکتورهای خطر بیماری های قلبی شناسایی شده، (۳) این امکان وجود دارد که با استفاده از داده های جمع آوری شده و استفاده از تکنیک های داده کاوی و یادگیری ماشین نسبت به پیش بینی و یا احتمال وقوع برخی از بیماری های قلبی نظیر CHD به این مهم دست یافت.

در مطالعه سلامت مردم یزد (یاس) Yazd Health Study اطلاعات سلامت و بیماری بیش از ده هزار نفر جمع آوری و ثبت شده است که شامل فاکتورهای خطر بیماری های قلبی نیز بوده و نیاز به تجزیه و تحلیل آن جهت کشف دانش تشخیص بیماری ها و به کارگیری آن جهت استفاده در ارتقای سلامت مردم یزد می باشد. سوالات منتخب از پرسش نامه طرح یاس مردم یزد برای پیش بینی بیماری قلبی در جدول ۱- الف در پیوست آورده شده است. در این مطالعه با استفاده از تکنیک های داده کاوی و هوش مصنوعی، راه کاری جهت غربالگری و شناسایی افراد مستعد بیماری قلبی در کوتاه ترین زمان ارائه شده است. این کار باعث می شود تا با آموزش بعدی و تغییر سبک زندگی و حذف فاکتورهای خطر، در مسیر افزایش طول عمر و امید به زندگی افراد مستعد به بیماری گام مؤثر برداشته شود. در مطالعه ای از روش های داده کاوی برای پیش بینی بیماری CHD با ۱۰۰۰ نمونه که ظرف شش ماه جمع آوری شده بود، استفاده شد که ۲۰۲ نفر آن ها فوت شده بودند. روش های بکار رفته شده Support Vector Machine، SVM، MPL Multi-Layer Perceptron و C5 بود. در این بررسی روش SVM با دقت ۹۲/۱ درصد بهترین روش و در

رتبه دوم روش شبکه عصبی چند لایه (MLP) با دقت ۹۱ درصد و بعد از آن الگوریتم C5 با ۸۹/۶ درصد دقت قرار دارد. برای ارزیابی روش های داده کاوی از روش 10-Fold cross-validation استفاده شده است (۴).

در مدل پیش بینی بیماری عروق کرونر به کمک شبکه های عصبی و گزینش متغیر مبتنی بر درخت رگرسیون و دسته بندی (۵) از ۱۳۲۲۸ نمونه ۴۰۵۹ نفر فاقد بیماری عروق کرونر و ۹۱۶۹ نفر مبتلا شامل ۹ فاکتور خطر بیماری قلبی استفاده شده است. این نمونه ها از مرکز قلب تهران و با استفاده از آنژیوگرافی به دست آمده است. بعد از هفت بار مدل سازی و مقایسه مدل های تولید شده، دقت ۷۴/۱ درصد به دست آمده است. با حذف ۵ متغیر و انجام آن با ۴ فاکتور خطر بیماری مجدداً همان دقت حاصل شده است که نشان می دهد تکنیک های گزینش متغیر به کاهش پیچیدگی مدل کمک می کند.

در مقاله راسل داس و همکاران از داده های بیماری قلبی University of California, Irvine UCI جهت طراحی سیستم خبره استفاده شده بود. در آن مقاله ۱۳ متغیر از ۷۶ متغیر انتخاب شده و بعد از حذف داده های گم شده با تعداد ۲۹۷ نمونه نسبت به طراحی سیستم خبره با استفاده از شبکه های عصبی مصنوعی Artificial Neural Network اقدام گردیده است که بهترین دقت آن ۸۹/۱ درصد بوده است (۲). در طراحی یک سیستم خبره برای تشخیص بیماری قلبی از ۱۹ صفت مجموعه داده UCI استفاده شده است. آن مقاله یک سیستم خبره مبتنی بر منطق فازی را ارائه داده است و آن را با سیستم های پشتیبان تصمیم بالینی مبتنی بر شبکه عصبی مصنوعی مقایسه کرده است. دقت سیستم فازی Mamdani ارائه شده ۶۸/۷ درصد می باشد و نسبت به مقالات مقایسه شده که با شبکه عصبی انجام شده بود دارای دقت بالاتری است (۶). در مقاله میلان کومار و همکاران از ۱۴ صفت، ۳۱۳ نمونه از داده های UCI و الگوریتم های شبکه های عصبی مصنوعی، ماشین بردار پشتیبان (SVM) و درخت تصمیم گیری

Decision Tree به کار گرفته شد و این نتیجه به دست آمد که ماشین بردار پشتیبان با دقت ۸۴/۱ درصد بهترین روش برای پیش بینی بیماری های قلبی عروقی است (۷). در مقاله بهداد میرزایی و همکاران با عنوان ارائه یک دسته بند مبتنی بر درخت تصمیم و شبکه عصبی و الگوریتم حرکت جمعی پرندگان برای تشخیص بیماری های قلبی از داده های UCI با ۱۴ ویژگی استفاده شده است. بهترین دقت برای درخت تصمیم با مقدار ۸۹/۳ درصد بوده است (۸). در خصوص تشخیص خودکار بیماری قلبی با استفاده از شبکه عصبی و سیستم عصبی فازی تحقیقی انجام شده است. تعداد ۱۳ پارامتر مؤثر در بیماری به صورت تجربی و به کمک آزمایش انتخاب شده اند. عمل تشخیص از روی داده ها توسط شبکه عصبی فازی و شبکه عصبی مصنوعی به ترتیب با دقت ۸۷ و ۷۵/۹ درصد صورت پذیرفت (۹).

کومار سن و همکاران با استفاده از روش شبکه عصبی فازی تجمیعی دوسطحی Neuro-Fuzzy Integrated Approach Two Level نسبت به آنالیز مجموعه داده UCI اقدام کردند. نتیجه بررسی به عمل آمده این بود که تنها ۹ عامل در بیماری Coronary Artery Disease CAD مؤثر است و در مقایسه با کارهای انجام شده مشابه دارای خطای کمتر و دقت بالاتر بوده است (۳). در مقایسه ای بین سه الگوریتم القای قواعد CN2، الگوریتم SOM Self-Organization Map و درخت تصمیم روی داده های UCI صورت گرفته است. الگوریتم CN2 توانسته است با دقت ۹۳/۷ درصد بیماری را تشخیص دهد (۱۰). در پیش بینی بیماری CHD از روی داده های طرح سلامت و تغذیه کشور کره انجام شده است. در آن مقاله روش ترکیبی از درخت تصمیم و منطق فازی ارائه گردیده که منجر به افزایش دقت و حساسیت در تشخیص شده است. دقت پیش بینی در روش ترکیبی فوق برابر ۶۹/۵ درصد بوده است (۱۱).

روش بررسی

همان طور که قبلاً بیان شد اکثر تحقیقات انجام شده در راستای پیش بینی بیماران قلبی بوده است. در این مقاله سعی

مصاحبه تأیید شده‌اند. جمع‌آوری اطلاعات مورد نیاز از طریق پرسش‌نامه و به صورت مصاحبه انجام شده است. پرسش‌نامه دارای پاسخ‌نامه قابل خوانده شدن به روش الکترونیکی بوده و توسط رایانه تصحیح گردیده است. هم‌چنین در این مطالعه فشارخون، قد و وزن افراد در منازل اندازه‌گیری می‌شود. به افراد دعوت‌نامه جهت حضور در آزمایشگاه مرکزی و تحویل نمونه خون داده شده است تا اطلاعات بیشتری جمع‌آوری گردد. اعتبار صوری پرسش‌نامه مورد بررسی قرار گرفته و پرسش‌نامه روی ۵۰ شرکت‌کننده پایلوت شده است. آلفای کرونباخ Cronbach's Alpha برابر ۰/۸۱ بوده بنابراین پرسش‌نامه معتبر در نظر گرفته شده است. جزییات روش مطالعه قبلاً منتشر شده است (۱۲).

از آن‌جا که تنها داده‌های مربوط به فاکتورهای خطر بیماری قلبی مد نظر این مقاله است لذا با توجه به پارامترهای انتخاب‌شده از مقاله‌های سن و کیم (۳،۱۱) به عنوان پارامترهای اصلی در بیماری قلبی و بیماری CHD و هم‌چنین با استفاده از نظر خبرگان تنها ده سؤال مرتبط از این پرسش‌نامه انتخاب شده است. پیوست الف-۱ سؤالات انتخاب شده را نشان می‌دهد. پاک‌سازی داده‌ها مرحله قبل از تحلیل داده‌ها می‌باشد. داده‌های به‌دست‌آمده در فرآیند جمع‌آوری اطلاعات باید قبل از به‌کارگیری توسط الگوریتم پیشنهادی و سایر الگوریتم‌ها آماده‌سازی شوند (۱۳). برای آماده‌سازی، داده‌های ورودی در قالب فایل صفحه گسترده به نرم‌افزار رپیدماینر که یکی از نرم‌افزارهای داده‌کاوی است داده شده است. تعداد داده‌های جمع‌آوری‌شده در این پژوهش ده هزار رکورد بوده که بعد از عملیات پاک‌سازی به روش حذف داده‌های گم‌شده Missing Value به تعداد ۸۱۸۸ رکورد تقلیل پیدا کرده است. بعد از آماده‌سازی داده‌ها نوبت به طبقه‌بندی داده‌ها می‌رسد. با توجه به وجود داده‌های وضعیت بیماری قلبی در پرسش‌نامه از روش‌های طبقه‌بندی استفاده می‌شود. برای انجام این کار و بالابردن دقت پیش‌بینی انجام کار از روش‌های گوناگون طبقه‌بندی استفاده شده است. در ادامه هر یک از آن‌ها به‌طور خلاصه معرفی شده‌اند. یکی از

در غربالگری افراد مستعد بیماری داریم. از آن‌جا که داده‌های استفاده شده جهت غربالگری از مجموعه داده‌های مطالعه یاس می‌باشد لذا در ابتدا روش جمع‌آوری داده‌ها بیان می‌گردد. سپس نحوه پیش‌پردازش داده‌ها و در انتها نحوه آنالیز و نتایج حاصل از آن آورده شده است. جهت انجام آنالیز روی داده‌ها از روش‌هایی از قبیل شبکه‌های عصبی مصنوعی، القای قواعد، رگرسیون و آدابوست استفاده شده است. پیاده‌سازی این روش‌ها در نرم‌افزارهای RapidMiner Studio (محصول شرکت رپیدماینر واقع در شهر بوستن Boston آمریکا) و Orange (محصول دانشگاه لیوبلیانا Ljubljana اسلونی) صورت گرفته است.

نمونه‌گیری مطالعه یاس به صورت خوشه‌ای انجام شده است. در مرحله اول، نمونه‌گیری دسته‌بندی بر حسب کد پستی و خوشه‌های شهری صورت گرفته است. سپس ۲۰۰ خوشه ۵۰ نفری انتخاب شده‌اند. این طرح یک مطالعه آینده‌نگر است که در سطح شهر یزد انجام شده است. جامعه آماری مطالعه افراد بالای ۲۰ تا ۶۹ سال ساکن شهرستان یزد بوده‌اند. روش نمونه‌گیری این مطالعه دومرحله‌ای و بدین شرح بوده است: نمونه‌گیری دسته‌بندی بر حسب کد پستی و خوشه‌های شهری است. در هر بلوک، بر اساس لیست فهرست برداری خانوار سال ۱۳۹۲، سرخوشه‌ها انتخاب و با حرکت از سمت راست نسبت به تکمیل پرسش‌نامه اقدام شده و خانوارهای بعدی به ترتیب انتخاب گردیدند. در صورتی که در یک پلاک چند خانوار وجود داشته (مثل مجتمع‌های مسکونی)، از واحد اول شروع و بعد به واحدهای بعدی مراجعه شده است. در صورتی که بیش از یک نفر واجد شرایط در محل بوده با همه افراد ۲۰ تا ۶۹ سال مصاحبه صورت گرفته است (ولی در هر گروه سنی ده ساله فقط یک نفر از هر آدرس) تا امکان بررسی تجمعات فامیلی فراهم شود.

پرسش‌گران در زمینه‌های پرسش‌گری، اخذ رضایت آگاهانه و رعایت اصول اخلاق پژوهش آموزش دیده و پس از شرکت در امتحان تئوری (پروتکل مطالعه) و عملی (پرسش‌گری و اندازه‌گیری فشارخون و شاخص‌های آنتروپومتریک) برای انجام

روش های دسته بندی استفاده از ماشین های بردار پشتیبان می باشد. این روش عموماً برای مسائلی که در آن ها دو دسته داده وجود دارد، استفاده می شوند. در این الگوریتم دسته بندی، دو صفحه، در مرز دو کلاس داده ها قرار گرفته می شود و مسئله یافتن مرز حداکثری بین این دو صفحه و در نتیجه بین دو دسته داده می باشد. به این صورت که دو صفحه آن قدر از هم دور می شوند که به داده ها برخورد نکنند. هدف یافتن دو صفحه ای است که بیشترین فاصله را دارد. دومین دسته بند استفاده شده در این مقاله الگوریتم بیز Naïve Bayes می باشد. این الگوریتم تحت نام های مختلف از جمله بیز ساده و مستقل، شبکه بیزی و شبکه باور شناخته شده است. همه این نام ها برگرفته از قضیه بیز در قاعده تصمیم گیری جهت دسته بندی داده ها است. این الگوریتم بر اساس مجموعه پارامترها و متغیرها امکان پیش بینی نتایج بر پایه احتمال را فراهم می کند (۱۳).

رگرسیون لجستیک Logistic Regression یک روش مبتنی بر آمار است که در یادگیری ماشین با نظارت استفاده می شود. یادگیری با نظارت به روشی گفته می شود که در زمان آموزش، نتیجه های هر نمونه از قبل معلوم و دارای برچسب باشد. در این روش تابعی محاسبه می شود که مقدار خروجی آن بر حسب متغیرهای ورودی است و تابع می تواند به صورت خطی یا گوسی پیاده سازی شود. وقتی خروجی تابع گوسی آن با توزیع برنولی جایگزین شود برای مسائل دسته بندی قابل استفاده است. رگرسیون لجستیک حالت خاص پیاده سازی رگرسیون با توزیع برنولی هست که خروجی آن دو حالت دارد مانند سلامت یا بیمار (۱۴). شبکه عصبی مصنوعی بر اساس مغز انسان یا حیوانات مدل می شود و به صورت فرم ساده ای از ورودی ها و خروجی ها است. تجهیزات پزشکی را می توان از طریق شبکه های عصبی مصنوعی نظارت کرد که دارای به روزرسانی های مداوم بسیاری از متغیرها مانند ضربان قلب، فشارخون و غیره باشند. عمل آموزش در این روش نسبتاً زمان گیر و زمان اجرای آن پس از یادگیری بسیار مناسب است (۱۳). دسته بند القای قواعد CN2 Rule Inducer توسط

Clark & Niblett در سال ۱۹۸۹ ارائه شد است. این الگوریتم جستجوی شعاعی عمومی به خصوصی را توصیف و از آن ترویج جهت ایجاد قواعد ترتیبی Ordered Rule استفاده می کند. مجموعه قوانین ایجاد شده مستقل هستند و لزوماً یک درخت را ایجاد نمی کنند. این الگوریتم ابتدا از یک اتصال تهی شروع می شود و بعد از آن اتصال هایی افزوده می شود و تالی قواعد بر اساس کلاس اکثریت نمونه هایی که توسط قواعد، پوشش داده می شوند تعیین می گردد (۱۵). در سال ۱۹۹۱ با استفاده تغییر در تابع اکتشافی و تخمین خطای لاپلاس امکان ایجاد قواعد غیر ترتیبی توسط این دسته بند ایجاد و القای قواعد CN2 نام گرفت. در سال ۲۰۰۴ با تطبیق دسته بند یادگیری، القای قواعد CN2 با کشف زیرمجموعه ای که هدف آن پیدا کردن قواعد معنادار از زیرگروه هایی که به اندازه کافی بزرگ و از لحاظ آماری غیرمعمول بودند توسعه داده شد که CN2-SD CN2 Subgroup Discovery نام گرفت و در این مقاله از این روش استفاده شده است. الگوریتم آدابوست Adaboost روشی برای بهبود دقت الگوریتم های یادگیری است. روند کلی این الگوریتم بدین صورت است که مجموعه ای وزن دار از تعداد زیادی دسته بندی کننده ضعیف به عنوان دسته بندی کننده نهایی انتخاب می نماید. به عبارت دیگر آدابوست با استفاده از مجموعه ای از دسته بندی کننده های ضعیف می تواند منجر به یک دسته بندی کننده قدرتمند گردد.

از آن جا که این روش می تواند دسته بندهای ضعیف را تقویت نماید برای دسته بندی داده های نویزدار مناسب نمی باشد اما برای داده های نامتوازن عملکرد بهتری دارد (۱۶). یکی از مسائل مهم در زمینه داده کاوی، مسئله دسته بندی مجموعه داده های نامتوازن است. اصطلاح مجموعه داده نامتوازن، عموماً به مجموعه داده ای گفته می شود که تعداد نمونه ها در کلاس های گوناگون اختلاف بسیاری داشته باشند (۴). الگوریتم های معمول داده کاوی در مواجهه با مشکل کلاس های نامتوازن معمولاً عملکرد ضعیفی دارند. در یک مسئله تشخیص پزشکی نمونه های بیماری در مقایسه با نمونه های معمول در اقلیت قرار دارند. ولی هدف از

این هزینه، در مجموعه داده‌ها وجود ندارد (۱۸). در این مقاله از روش سطح داده جهت متوازن‌سازی داده‌ها استفاده شده است. دانشی که در مرحله یادگیری دسته‌بند تولید می‌شود؛ می‌بایست در مرحله ارزیابی مورد تحلیل قرار گیرد تا به توان ارزش آن را تعیین نمود و در پی آن کارایی الگوریتم یادگیرنده دسته‌بند را نیز مشخص کرد. این معیارها را می‌توان هم برای مجموعه داده‌های آموزشی در مرحله یادگیری و هم برای مجموعه رکوردهای آزمایشی در مرحله ارزیابی محاسبه نمود. در ابتدا مفهوم ماتریس درهم ریختگی Confusion Matrix بیان می‌شود. این ماتریس چگونگی عملکرد الگوریتم دسته‌بندی را با توجه به مجموعه داده ورودی به تفکیک انواع دسته‌های مساله دسته‌بندی مطابق جدول ۱ نمایش می‌دهد.

TP: بیانگر تعداد رکوردهایی است که دسته واقعی آن‌ها مثبت بوده و الگوریتم دسته‌بندی نیز دسته آن‌ها را به درستی مثبت تشخیص داده است.

TN: بیانگر تعداد رکوردهایی است که دسته واقعی آن‌ها منفی بوده و الگوریتم دسته‌بندی نیز دسته آن‌ها را به درستی منفی تشخیص داده است.

FP: بیانگر تعداد رکوردهایی است که دسته واقعی آن‌ها منفی بوده و الگوریتم دسته‌بندی دسته آن‌ها را به اشتباه مثبت تشخیص داده است.

FN: بیانگر تعداد رکوردهایی است که دسته واقعی آن‌ها مثبت بوده و الگوریتم دسته‌بندی دسته آن‌ها را به اشتباه منفی تشخیص داده است.

جدول ۱: ماتریس درهم ریختگی

واقعی	پیش‌بینی		Σ
	Positive	Negative	
Positive	TP	FN	$P=TP+FN$
Negative	FP	TN	$N=FP+TN$
Total	$TP+FP$	$FN+TN$	

دسته‌بندی این داده‌ها یافتن نمونه‌های بیماری است. به این ترتیب کلاس اقلیت نسبت به سایر نمونه‌ها دارای اولویت بیشتری است. در اغلب الگوریتم‌های دسته‌بندی، تمایل در جهت کلاسی است که بیشترین تعداد نمونه‌ها را دارد. از این رو توانایی کمی در پیش‌گویی صحیح داده‌های کلاس اقلیت از خود نشان می‌دهد. در نتیجه، قوانین دسته‌بند سبب می‌شوند تا داده‌های کلاس اقلیت، نادرست دسته‌بندی شوند. برای حل این مشکل باید از روش‌های مواجهه با داده‌های نامتوازن استفاده نمود. تکنیک‌های متنوعی برای حل مسئله در ارتباط با کلاس نامتوازن پیشنهاد شده است، که در سه گروه: روش سطح داده، روش سطح الگوریتم و روش حساس به هزینه تقسیم می‌شوند. در ادامه به توصیف هر کدام از این روش‌ها خواهیم پرداخت.

۱. روش سطح داده

در این روش با اضافه کردن مرحله پیش‌پردازش Preprocessing قبل از دسته‌بندی، موجب متوازن شدن مجموعه داده‌های نامتوازن می‌شود.

۲. روش سطح الگوریتم

این روش که یکی از روش‌های بازبینی در سطح الگوریتم است با تغییر در الگوریتم دسته‌بندی به نوعی مسئله عدم توازن مرتفع می‌شود. این روش فرآیند یادگیری را به سمت کلاس اقلیت سوق می‌دهد.

۳. روش حساس به هزینه

این روش مابین روش سطح داده و الگوریتمی قرار دارد، به طوری که هم در سطح داده و هم در سطح الگوریتم تغییر ایجاد خواهد کرد. مهم‌ترین نقطه‌ضعف این روش، تعریف هزینه دسته‌بندی نادرست است که عموماً اطلاعاتی در مورد

جدول ۲: نحوه محاسبه برخی از معیارهای ارزیابی

Measure	Formula
accuracy, recognition rate	$\frac{TP + TN}{P + N}$
error rate, misclassification rate	$\frac{FP + FN}{P + N}$
sensitivity, true positive rate, recall	$\frac{TP}{P}$
specificity, true negative rate	$\frac{TN}{N}$
precision	$\frac{TP}{TP + FP}$
detection rate	$\frac{TP}{FN + TP}$
false alarm rate	$\frac{FP}{TN + FP}$

از مقادیر ماتریس درهم‌ریختگی جهت ارزیابی دسته بند استفاده می‌شود. جدول ۲ نحوه محاسبه برخی از معیارهای ارزیابی بر اساس مقادیر ماتریس در هم ریختگی را نشان می‌دهد. دقت یا نرخ تشخیص یکی از مهم ترین معیارها برای کارایی الگوریتم یک دسته بند است که نشان‌دهنده میزان پیش بینی صحیح نسبت به کل نمونه ها است. با توجه به این که معیار دقت دسته‌بندی، ارزش رکوردهای دسته‌های مختلف را یکسان در نظر می‌گیرد در مسائلی که با دسته‌هایی با داده‌های نامتوازن سروکار داریم از معیارهای دیگری استفاده می‌شود. به‌عنوان مثال در پزشکی اغلب تعداد داده‌های افراد بیمار به نسبت کل جامعه دارای تعداد کمی است و پیش‌بینی آن از اهمیت بالایی برخوردار است لذا معیار دقت نمی‌تواند برای ارزیابی آن مناسب باشد. در این گونه مسائل، معیارهای دیگری نظیر DR Detection Rate و FAR False Alarm Rate اهمیت ویژه‌ای دارند. این معیارها توجه بیشتری به دسته بند مثبت نشان می‌دهند. معیار DR نشان می‌دهد که دقت تشخیص دسته مثبت چه مقدار است و معیار FAR نرخ هشدار غلط را با توجه به دسته منفی بیان می‌کند. معیار مهم دیگری که برای تعیین میزان کارایی یک دسته بند

استفاده می‌شود معیار Area Under Curve AUC است و نشان دهنده سطح زیر نمودار ROC Receiver Operating Characteristic می‌باشد. هر چه مقدار این عدد مربوط به یک دسته بند بزرگتر باشد کارایی نهایی دسته بند مطلوب‌تر ارزیابی می‌شود. نمودار ROC روشی برای بررسی کارایی دسته‌بندها است. در واقع منحنی‌های ROC منحنی‌های دو بعدی هستند که DR یا همان نرخ تشخیص صحیح دسته مثبت روی محور Y و FAR یا همان نرخ هشدار غلط روی محور X رسم می‌شوند. به بیان دیگر یک منحنی ROC مصالحه نسبی میان سودها و هزینه‌ها را نشان می‌دهد. به همین دلیل معیار AUC که سطح زیر نمودار ROC را نشان می‌دهد می‌تواند نقش تعیین کننده‌ای در معرفی دسته بند برتر ایفا کند. مقدار AUC برای یک دسته بند که به طور تصادفی، دسته نمونه مورد بررسی را تعیین می‌کند برابر ۰.۵ است. هم چنین بیشترین مقدار این معیار برابر یک بوده و برای وضعیتی رخ می‌دهد که دسته‌بند ایده آل بوده و به تواند کلیه نمونه‌های مثبت را بدون هرگونه هشدار غلطی تشخیص دهد. یک روش ارزیابی میزان خطا و بررسی قابلیت تعمیم‌پذیری دسته بند، روش Cross Validation می‌باشد. در این روش داده‌ها به دو

ملاحظات اخلاقی

پروپوزال این تحقیق توسط دانشگاه علم و هنر تایید شده است (کد اخلاق IR.SSU.REC.1397.246837)

نتایج

در مطالعه یاس که از داده های آن برای این پژوهش استفاده شده، جامعه نمونه گیری، ساکنان شهرستان یزد می باشند که به صورت تصادفی خوشه ای نمونه گیری شده اند و نسبت افراد سالم به بیماران قلبی ۱۰ به ۱ می باشد.

برای برقراری تعادل بین دو کلاس نسبت به متوازن سازی در سطح داده و با روش افزایش کلاس اقلیت به روش بوت استرپ اقدام شده است. با این کار تعداد نمونه های افراد بیمار از ۷۳۸ به ۷۳۸۰ نمونه افزایش یافته و تعداد نمونه های افراد سالم همان ۷۴۵۰ نمونه باقی مانده است. بدین صورت مجموع نمونه ها از ۸۱۸۸ به ۱۴۸۳۰ نمونه افزایش یافته است. از چندین دسته بند مختلف جهت ارزیابی کارایی الگوریتم های مختلف برای تشخیص افراد مستعد بیماری استفاده شده است. این دسته بندها عبارت اند از: بیز، شبکه های عصبی، بردار پشتیبان ماشین، CN2-SD، درخت تصمیم، نزدیک ترین همسایه، آدابوست و رگرسیون لجستیک. سپس به دو روش، دسته بندها مورد ارزیابی قرار گرفته اند. ابتدا با روش 10-Fold cross-validation از ارزیابی انجام شده است. سپس جهت بررسی کارایی دسته بندها در محیط واقعی کل نمونه ها به عنوان داده آزمون استفاده شده اند. نتایج حاصله به صورت ماتریس در هم ریختگی در ادامه آورده شده است. از آن جا که هدف از دسته بندی شناسایی افراد مستعد بیماری است معیاری که جهت ارزیابی آن استفاده می شود حساسیت Sensitivity است. این معیار نشان دهنده نسبت افراد مستعد بیماری که توسط دسته بند درست تشخیص داده شده به کل افراد مستعد بیماری است. در این ارزیابی الگوریتم های آدابوست، درخت تصمیم و نزدیک ترین همسایه با حساسیت ۸۷ درصدی در تشخیص افراد مستعد بیماری بهترین عملکرد را داشتند.

قسمت داده های آموزشی و داده های آزمون تقسیم بندی می شوند. در روش جامع k-Fold Cross Validation داده ها به k قسمت مساوی تقسیم شده و به تعداد دفعات k، یک قسمت به عنوان داده آزمون و بقیه به عنوان داده آموزشی مورد استفاده قرار می گیرد. میانگین خطای k مرحله، به عنوان خطای دسته بند تعیین می گردد. معمول ترین مقداری که در متون علمی برای k در نظر گرفته می شود برابر با ۱۰ می باشد. در روش k-Fold Cross Validation فرض بر آن است که عملیات انتخاب نمونه های آموزشی بدون جای گذاری صورت می گیرد. به بیان دیگر یک رکورد تنها یک بار در یک فرآیند آموزشی مورد توجه واقع می شود. چنانچه هر رکورد در صورت انتخاب شدن برای شرکت در عملیات یادگیری دسته بند بتواند مجدداً برای یادگیری مورد استفاده قرار گیرد روش مزبور با نام Bootstrap شناخته می شود. (۱۷).

در این مقاله با توجه به کم بودن نمونه های افراد مبتلا به بیماری قلبی، با استفاده از روش Bootstrap و جای گذاری نمونه های کلاس افراد بیمار، تعداد نمونه موجود در این کلاس ده برابر افزایش یافته است. لازم به ذکر است میزان افزایش تعداد نمونه کلاس با توجه به تعداد نمونه سالم و در راستای متوازن سازی داده ها انجام می شود و قابل تنظیم می باشد. کلیه عملیات جای گذاری نمونه ها در نرم افزار رپیدماینر انجام شده است. در اکثر کارهای مشابه انجام شده از داده های مربوط به افراد مراجعه کننده به بیمارستان ها استفاده است. به عنوان مثال مجموعه داده جمع آوری شده در (۵) جامعه نمونه گیری، محدود به افرادی است که نسبت به آنژیوگرافی اقدام کرده اند و در آن افراد بیمار بیش از دو برابر افراد سالم می باشد. به عبارت دیگر دارای جامعه آماری محدودی می باشند. در این مقاله با توجه به نوع جامعه آماری استفاده شده، تعداد افراد سالم بسیار بیشتر از افراد بیمار بوده و با استفاده از روش Bootstrap عمل متوازن سازی انجام شده است. در ادامه نتایج حاصل از دسته بندی داده های موجود در طرح یاس مورد بررسی قرار گرفته است.

جدول ۳: ماتریس درهم‌ریختگی حاصل از ارزیابی با 10-fold در ساکنین ۲۰-۶۹ سال یزد در سال ۱۳۹۴

پیش‌بینی واقعی	آدابوست		درخت تصمیم		Σ
	بیمار	سالم	بیمار	سالم	
بیمار	۶۴۳۰	۹۵۰	۶۴۴۶	۹۳۴	۷۳۸۰
سالم	۱۱۹۳	۶۲۵۷	۱۳۴۵	۶۱۰۵	۷۴۵۰
مجموع	۷۶۲۳	۷۲۰۷	۷۷۹۱	۷۰۳۹	۱۴۸۳۰
پیش‌بینی	نزدیک‌ترین همسایه		بردار پشتیبان		Σ
	بیمار	سالم	بیمار	سالم	
بیمار	۶۴۴۶	۹۳۴	۵۹۲۰	۱۴۶۰	۷۳۸۰
سالم	۱۶۲۴	۵۸۲۶	۱۶۰۴	۵۸۴۶	۷۴۵۰
مجموع	۸۰۷۰	۶۷۶۰	۷۵۲۴	۷۳۰۶	۱۴۸۳۰
پیش‌بینی واقعی	رگرسیون لجستیک		شبکه عصبی		Σ
	بیمار	سالم	بیمار	سالم	
بیمار	۵۴۷۰	۱۹۱۰	۵۴۳۳	۱۹۴۷	۷۳۸۰
سالم	۱۸۷۳	۵۵۷۷	۱۸۴۱	۵۶۰۹	۷۴۵۰
مجموع	۷۳۴۳	۷۴۸۷	۷۲۷۴	۷۵۵۶	۱۴۸۳۰
پیش‌بینی	بیز		CN2		Σ
	بیمار	سالم	بیمار	سالم	
بیمار	۵۳۱۷	۲۰۶۳	۵۲۲۳	۲۱۵۷	۷۳۸۰
سالم	۱۸۲۲	۵۶۲۸	۱۷۲۷	۵۷۲۳	۷۴۵۰
مجموع	۷۱۳۹	۷۶۹۱	۶۹۵۰	۷۸۸۰	۱۴۸۳۰

جدول ۴: نتیجه ارزیابی با 10-Fold cross-validation

نام دسته‌بند	سطح زیر نمودار	دقت	حساسیت
آدابوست	۰/۹۲۵	۰/۸۵۵	۰/۸۷۱
درخت تصمیم	۰/۹۱۷	۰/۸۴۶	۰/۸۷۳
نزدیک‌ترین همسایه	۰/۹۰۶	۰/۸۲۸	۰/۸۷۳
ماشین پشتیبان بردار	۰/۸۵۴	۰/۷۹۳	۰/۸۰۲
رگرسیون لجستیک	۰/۸۱۹	۰/۷۴۵	۰/۷۴۱
شبکه عصبی	۰/۸۱۹	۰/۷۴۶	۰/۷۳۶
بیز	۰/۸۰۷	۰/۷۳۸	۰/۷۲
القای قواعد	۰/۸۰۷	۰/۷۳۸	۰/۷۰۸

جدول ۵: ماتریس درهم‌ریختگی حاصل از ارزیابی با کل داده‌ها

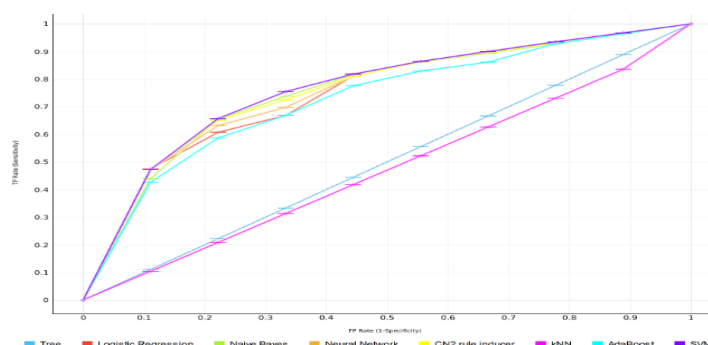
پیش‌بینی واقعی	بردار پشتیبان		شبکه عصبی		بیمار	سالم	Σ
	بیمار	سالم	بیمار	سالم			
بیمار	۵۶۲	۱۷۶	۷۳۸	۵۶۷	۱۷۱	۷۳۸	
سالم	۲۵۲۹	۴۹۲۱	۷۴۵۰	۲۹۶۹	۴۴۸۱	۷۴۵۰	
مجموع	۳۰۹۱	۵۰۹۷	۸۱۸۸	۳۵۳۶	۴۶۵۲	۸۱۸۸	
پیش‌بینی CN2							
بیز							
واقعی	بیمار	سالم	Σ	بیمار	سالم	Σ	
بیمار	۶۱۷	۱۲۱	۷۳۸	۴۴۹	۲۸۹	۷۳۸	
سالم	۳۴۸۶	۳۹۶۴	۷۴۵۰	۱۳۹۰	۶۰۶۰	۷۴۵۰	
مجموع	۴۱۰۳	۴۰۸۵	۸۱۸۸	۱۸۳۹	۶۳۴۹	۸۱۸۸	

پیش‌بینی رگرسیون لجستیک آدابوست

واقعی	بیمار		سالم		Σ		بیمار	سالم	Σ
	بیمار	سالم	بیمار	سالم	بیمار	سالم			
بیمار	۴۰۱	۳۳۷	۷۳۸	۲۴۱	۴۹۷	۷۳۸			
سالم	۱۰۸۳	۶۳۶۷	۷۴۵۰	۴۹۰	۶۹۶۰	۷۴۵۰			
مجموع	۱۴۸۴	۶۷۰۴	۸۱۸۸	۷۳۱	۷۴۵۷	۸۱۸۸			
پیش‌بینی درخت تصمیم نزدیک‌ترین همسایه									
واقعی	بیمار	سالم	Σ	بیمار	سالم	Σ	بیمار	سالم	Σ
بیمار	۰	۷۳۸	۷۳۸	۷۳۸	۰	۷۳۸			
سالم	۰	۷۴۵۰	۷۴۵۰	۷۴۵۰	۰	۷۴۵۰			
مجموع	۰	۸۱۸۸	۸۱۸۸	۸۱۸۸	۰	۸۱۸۸			

جدول ۶: نتیجه ارزیابی با کل داده‌ها

دسته‌بند	سطح زیر نمودار	دقت	حساسیت
ماشین پشتیبان بردار	۰/۷۷۲	۰/۶۷	۰/۷۶۲
شبکه عصبی	۰/۷۶۴	۰/۶۱۷	۰/۷۶۸
القای قواعد	۰/۷۶۴	۰/۵۵۹	۰/۸۳۶
بیز	۰/۷۶۳	۰/۷۹۵	۰/۶۰۸
رگرسیون لجستیک	۰/۷۵۹	۰/۸۲۷	۰/۵۴۳
آدابوست	۰/۷۳۳	۰/۸۷۹	۰/۳۲۷
درخت تصمیم	۰/۵	۰/۹۱	۰
نزدیک‌ترین همسایه	۰/۴۷۵	۰/۰۹	۱



شکل ۱: نمودار ROC ارزیابی با کل داده ها

دسته بندهای شبکه عصبی، القای قواعد و بیز با سطح زیر نمودار ROC یکسان قرار دارند. از آن جا که حساسیت معیار مهمی جهت غربالگری افراد مستعد بیماری می باشد در بین این دسته بندها دسته بند القای قواعد با بیشترین حساسیت بهترین عملکرد را دارا می باشد. شکل ۱ نمودار ROC این ارزیابی را نشان می دهد. از آن جا داده های موجود در دنیای واقعی مثل جامعه آماری طرح یاس، نامتوازن می باشند لذا پیشنهاد می شود جهت غربالگری افراد مستعد بیماری قلبی از روش القای قواعد استفاده شود. این روش با داشتن نرخ بالای حساسیت نسبت به سایر دسته بندهای معرفی شده در این مقاله می تواند عملکرد بهتری را روی این دسته از داده ها داشته باشد.

نتیجه گیری

جهت انجام عمل غربالگری بیماران مستعد بیماری کرونر قلبی از داده های طرح یاس استفاده شده است. به دلیل نامتوازن بودن نسبت افراد سالم به بیماران قلبی با تقویت داده های کلاس افراد بیمار به روش بوت استرپ نسبت به متوازن سازی داده ها اقدام شده است. نتیجه آموزش و ارزیابی دسته بندها به روش 10fold cross validation نشان می دهد که روش های مختلف ارزیابی تأثیر بسزایی در نتایج این گونه روش متوازن سازی دارند و در مقایسه با مقالاتی که از داده های UCI استفاده کرده اند دقت های تقریباً مشابهی را نشان می دهند. اما به دلیل این که توزیع داده های UCI متناسب با

بحث

از مقادیر ماتریس درهم ریختگی و روابط بیان شده می توان مقادیر معیارهای ارزیابی را محاسبه کرد. در این جا از سه معیار دقت، حساسیت و مقادیر سطح زیر نمودار ROC استفاده شده است. این مقادیر برای دسته بندی های استفاده شده در جدول ۴ آورده شده است. معیار دقت، عملکرد الگوریتم را هم در تشخیص افراد سالم و هم در تشخیص افراد مستعد بیماری با وزن یکسان نشان می دهد. همان طور که از جدول فوق پیداست، دسته بندهای آدابوست، درخت تصمیم و نزدیک ترین همسایه بالاترین دقت را داشته اند و این برتری در معیار سطح زیر نمودار ROC که حاکی از کارایی این دسته بندها است مشهود می باشد. برای این که نشان دهیم آیا این سه دسته بند در محیط واقعی یعنی محیطی که نسبت افراد سالم به افراد مستعد بیماری قلبی در آن ۱۰ به ۱ است دارای عملکرد مشابهی هستند یا خیر، کل داده ها قبل از انجام عمل متوازن سازی به عنوان داده آزمایشی به تمام دسته بندها داده شده است. مقادیر ماتریس درهم ریختگی حاصل از انجام عملیات دسته بندی روی این داده ها در جدول ۵ آمده است. بر خلاف انتظار سه دسته بند مذکور که در ارزیابی به روش 10-Fold cross-validation بهترین کارایی را داشتند با ارزیابی کل داده ها به عنوان داده آزمون، دارای کمترین کارایی می باشند. نتایج این محاسبات در جدول ۶ آمده است. دسته بند ماشین پشتیبان بردار بهترین کارایی را در این ارزیابی دارد و بعد از آن

داده های کلاس افراد بیمار باعث تشکیل قواعد معنادار در کلاس افراد بیمار می شود که حساسیت بالای شناسایی افراد بیمار در این دسته بند بیان گر همین موضوع می باشد. این روش با توجه به هرس های زیادی که انجام می دهد در معرض بیش برآزش نمی باشد و دارای قدرت تعمیم پذیری بالاتری است. به طوری که تنها با قواعد تولید شده در نرم افزار دقتی برابر با دسته بند شبکه عصبی دارد و با حساسیت ۸۳/۶ درصد بهترین عملکرد را در شناسایی افراد مستعد بیماری دارد. از محدودیت های این مقاله، می توان عدم وجود اطلاعات مربوط به مصرف الکل در پرسش نامه که از فاکتورهای خطر ابتلا به اسکیمیک قلبی است؛ را نام برد. محدودیت دیگر تحقیق عدم وجود داده های مرتبط با فاکتورهای خطر در نمونه های موجود طرح یاس می باشد. برای رفع این مشکل نمونه هایی که فاقد این مقادیر بودند از نمونه های اولیه حذف شده اند.

سپاس گذاری

این مقاله بخشی از پایان نامه کارشناسی ارشد رشته مهندسی کامپیوتر گرایش نرم افزار دانشگاه علم و هنر می باشد که بدون حمایت مالی انجام شده است. در پایان از تمامی افرادی که در انجام این پایان نامه همکاری نموده اند تشکر می گردد.

تعارض در منافع: وجود ندارد

افراد مراجعه کننده به بیمارستان است برای تعمیم به کل جامعه جهت استفاده هایی هم چون غربالگری افراد مستعد بیماری مناسب نیستند. در مرحله بعد برای بررسی میزان موثر بودن دسته بند آموزش دیده، ابتدا داده اصلی به روش بوت استرپ متوازن سپس با استفاده از روش 10fold cross validation آموزش و ارزیابی شده است. داده های اصلی (قبل از انجام عمل متوازن سازی) نیز توسط دسته بندهای معرفی شده در جدول ۶ مورد ارزیابی قرار گرفته است. مقایسه نتایج این دو ارزیابی تاثیر متوازن سازی روی دسته بندهای مختلف را نشان می دهد. به عنوان مثال در درخت تصمیم نمونه های موجود در کلاس افراد بیمار به دلیل افزایش نمونه با جایگزینی در روش بوت استرپ، دارای پراکندگی کمتری هستند.

در ارزیابی دوم نتایج به سمت افراد سالم متمایل شده و مجدداً به دلیل وجود نمونه های مشابه در دسته بند نزدیک ترین همسایه، نتایج به سمت کلاس بیمار متمایل شده است. دسته بند های بردار پشتیبان و شبکه عصبی از جمله دسته بند های پارامتریک هستند که رفتار مشابهی از خود به روز داده اند. این دسته بند ها با توجه به این که کلیه داده های آزمون در داده های آموزش وجود دارد مستعد بیش برآزش Over Fitting بوده و خطای تعمیم پذیری بالایی دارند. اما در دسته بند CN2-SD با توجه به ترکیب الگوریتم القای قواعد با الگوریتم قواعد انجمنی و کشف زیرگروه های معنادار، تقویت

References:

- 1-Purusothaman G, Krishnakumari P. *A Survey of Data Mining Techniques on Risk Prediction: Heart Disease*. Indian J Sci Techno 2015; 8(12): 1-5.
- 2- Das R, Turkoglu I, Sengur A. *Effective Diagnosis of Heart Disease through Neural Networks Ensembles*. Expert Systems with Applications 2009; 36(4): 7675-680.
- 3-Sen AK, Patel SB, Shukla DP. *A Data Mining Technique For Prediction Of Coronary Heart Disease Using Neuro-Fuzzy Integrated Approach Two Levels*. International J Engineer and Computer Sci 2013; 2(9): 1663-671.
- 4-Xing Y, Wang J, Zhao Z. *Combination data mining methods with new medical data to predicting*

- outcome of coronary heart disease*. IEEE International Conference on Convergence Information Technology, Gyeongju, South Korea, 21-23 Nov 2007: 868-72.
- 5-Mahmoudi E, Asgari Moghadam R, Moazzam MH, Sadeghian S. *Prediction Model For Coronary Artery Disease Using Neural Networks And Feature Selection Based On Classification And Regression Tree*. J Shahrekord Uni Med Sci 2013; 15(5): 47-56. [Persain]
- 6-Hedayati E. S, *Designing an Expert System for Heart Disease*. The 9th Symposium on Advances in Science & Technology, 4 December 2014, Mashhad, 868-72.
- 7- Kumari M, Godara S. *Comparative Study of Data Mining Classification Methods in Cardiovascular Disease Prediction*. International J Computer Sci Trends Techno 2011; 2(2): 304-8.
- 8-Mirzaei, B, Reza R. *A Decision Based Modeling and Neural Network Based Modeling Model and Bird Motion Detection Algorithm for Heart Disease*. 2015, 2nd IEEE National Conference of Technology, Energy and Data on Electrical and Computer Engineering, Kermanshah, 2016. [Persain]
- 9-Abushariah MA, Alqudah AA, Adwan OY, Yousef RM. *Automatic Heart Disease Diagnosis System Based On Artificial Neural Network (ANN) and Adaptive Neuro-Fuzzy Inference Systems (ANFIS) Approaches*. J Software Engineering Applications 2014; 7(12): 1055-64.
- 10- Ramaraj M, Selvadoss TA. *A Comparative Study Of CN₂ Rule And SVM Algorithm And Prediction Of Heart Disease Datasets Using Clustering Algorithms*. Network Complex Systems 2013; 3(10): 1-6.
- 11- Kim J, Lee J, Lee Y. *Data-Mining-Based Coronary Heart Disease Risk Prediction Model Using Fuzzy Logic and Decision Tree*. Healthc Inform Res 2015; 21(3):167-74.
- 12- Mirzaei M, Salehi-Abargouei A, Mirzaei M, Mohsenpour MA. *Cohort Profile: The Yazd Health Study (YaHS): a population-based study of adults aged 20–70 years (study design and baseline population data)*. International journal of epidemiology. 2017; 47(3):697-8h.
- 13- Bell J. *Machine Learning: Hands-On for Developers and Technical Professionals*. John Wiley & Sons. 2014.
- 14- Murphy P. K. *Machine Learning a Probabilistic Perspective*, Book, Publisher: The MIT Press; Published 2012, Published Online 23 Apr 2014:62-63, from: <https://doi.org/10.1080/09332480.2014.914768>.
- 15- Clark P, Niblett T. *The CN₂ induction algorithm*. Machine learning 1989; 3(4): 261-83.
- 16- Molahosseini A, Amirkhani H, Rahmati M, *Using Adaboost Classifier Composition And Genetic Algorithms In Cryogenic Analysis Methods*. 15th

- International Annual Conference of the Computer Society of Iran, Tehran, 2009.
- 17- Han J, Pei J, Kamber M. *Data mining: concepts and techniques*. Elsevier, 2011.
- 18- Pirmohammadi A, *Classification of unbalanced data using a combination of classifiers and support vector descriptors*. Master dissertation, University of Basic Science Zanjan, 2014. [Persain]

The prediction model for cardiovascular disease using Yazd's health study data (YaHS)

Seyed Mohammad Reza Tabatabaei Nodoushan¹, Fatemeh Saadatjoo², Masoud Mirzaei^{*3}

Original Article

Introduction: Ischemic heart disease is one of the most common diseases, which has led to high mortality rates all over the world. This disease is caused by narrowing or blockage of coronary arteries, which are the provider of blood to the heart. Identifying the people susceptible to this disease and bringing changes in their lifestyles has been said to reduce the related mortality rates and increase the patient's longevity.

Methods: Yazd people Health Study (YaHS) was conducted on a random sample of 10,000 people living in the city of Yazd, Iran in the years 2014-15 for a general health and disease survey. These data were first balanced by bootstrapping technique due to their unbalanced nature. Next, classification methods were used in the training phase. Various classifiers, such as artificial neural network, rule inducer, regression, and AdaBoost were used in order to evaluate the proposed method with two scenarios.

Results: The results showed that the screening of the people susceptible to ischemic heart disease had the most significant effect on increasing the sensitivity of the discovery classifier of CN2 subgroup through using balanced data by bootstrapping method followed by their analysis for the purpose of producing a sample of the patients. This classifier proved to have the potential for detecting 83.6% of the people susceptible to this disease.

Conclusion: Therefore, it can be concluded that data mining methods are effective in screening for susceptible people with ischemic heart disease. This method can be compared with other traditional screening methods in that it is more cost-effective and faster.

Keywords: Data mining, Health monitoring, Prediction, Ischemic heart disease, Data balancing, Rule induction CN2-SD.

Citation: Tabatabaei MR, Saadatjoo F, Mirzaei M. **The prediction model for cardiovascular disease using Yazd's health study data (YaHS).** J Shahid Sadoughi Uni Med Sci 2019; 27(3): 1346-60

¹Department of Computer Engineering, Science and Arts University, Yazd, Iran.

²Department of Computer Engineering, Science and Arts University, Yazd, Iran.

³Department of Epidemiology, Yazd Cardiovascular Research Center, Shahid Sadoughi University of Medical Sciences, Yazd, Iran

***Corresponding author: Tel:** 09134509917, **email:** masoudmirzaei@yahoo.com