مجله علمی پژوهشی دانشگاه علوم پزشکی شهید صدوقی یزد
دوره بیست و هفتم، شماره سه، خرداد ۱۳۹۸
تاریخ دریافت:۱۳۹۷/۳/۲
تاریخ پذیرش:۱۳۹۷/۸/۲۶

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

سید محمدرضا طباطبائی ندوشن[۱]، فاطمه سعادت جو[۲]، مسعود میرزایی*[۳]

**مقاله پژوهشی**

**مقدمه:** بیماری‌های ایسکمیک قلبی یکی از شایع‌ترین بیماری هایی است که آمار مرگ‌ومیر بالایی را در جهان به خود اختصاص می‌دهد. بیماری‌های ایسکمیک قلب به‌دنبال تنگ شدن یا بسته شدن شریان‌های کرونر قلب به‌وجود می آیند که تامین کننده خون قلب هستند، این امر به آهستگی و به مرور زمان رخ می‌دهد. 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 به‌علت داشتن ماهیت نامتوازن ابتدا، توسط روش بوت استراپ (Bootstrap) متوازن‌سازی‌شده ، آنگاه در مرحله آموزش دسته بندها به‌کار برده شده‌اند. 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:** یافته‌ها نشان داد که عمل غربالگری افراد مستعد به بیماری‌های ایسکمیک قلبی با استفاده از تولید نمونه افراد بیمار به روش بوت استراپ و متوازن سازی داده ها امکان‌پذیر است. این روش بیشترین تاثیر را در افزایش حساسیت دسته‌بند کشف زیرگروه CN2 دارد. این دسته‌بند توانایی تشخیص ٪۸۳/۶ از افراد مستعد بیماری را داراست.

**نتیجه‌گیری:** بنابراین می‌توان نتیجه گرفت که روش‌های داده‌کاوی در غربالگری افراد مستعد بیماری ایسکمیک قلبی کارایی مناسبی دارد و به کمک آن می‌توان افراد مستعد این بیماری را نسبت به‌غربالگری سنتی که با مراجعه حضوری افراد به پزشک انجام می‌شود؛ سریع‌تر و با هزینه کمتر شناسایی نمود.

**واژه‌های کلیدی:** داده‌کاوی، پایش سلامت، پیش بینی، بیماری‌های ایسکمیک قلب، متوازن‌سازی داده، القای قواعد CN2-SD

**ارجاع:** طباطبائی سید محمدرضا، سعادت‌جو فاطمه، میرزایی مسعود. **مدل پیش بینی بیماری های عروق کرونر قلب با استفاده از داده‌کاوی داده‌های 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. بیماری‌های ایسکمیک قلب (CHD) Coronary Heart Disease یکی از بزرگ ترین عوامل مرگ ومیر در سراسر جهان به‌شمار می روند (۱).

در علم پزشکی تعیین یک محدوده مشخص برای سالم و یا بیمار بودن شخص وجود ندارد. به‌همین دلیل تصمیم گیری و تشخیص در بسیاری از موارد مبهم و غیرقطعی انجام می شود. در این حرفه، پزشک نمی تواند به‌راحتی 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 اندازه‌گیری و ثبت اطلاعات بیمار به صورت دقیق به وجود آمده و توسعه فناوری اطلاعات و ارتباطات امکان ذخیره‌سازی و آنالیز حجم انبوهی از این داده‌ها را فراهم آورده است. این داده‌ها توسط الگوریتم‌های دسته‌بند و بنا به نیاز کاربران قابل تحلیل است و این تحلیل به ارتقاء سلامت مردم کمک می‌کند. از تحلیل این داده‌ها می توان الگوریتم های تشخیصی استخراج کرد که نمونه ای از آن در این مقاله آورده شده است. این داده‌ها هم‌اکنون در پرونده الکترونیک سلامت ذخیره و نگهداری می‌شود.

این دانش‌ها می‌توانند در زمینه های مختلف مانند پیشگیری، تشخیص، رویکرد درمانی و یا پیش بینی (پروگنوز) بیماری‌ها مفید باشند. طیف بیماری‌هایی که از این دانش‌ها سود می‌برند بسیار گسترده است. یکی از بیماری‌هایی که امید می‌رود با استفاده از این دانش در تشخیص و رویکرد درمانی آن تحول ایجاد شود بیماری‌های ایسکمیک قلبی عروقی است. آنژیوگرافی، آزمایش استاندارد طلایی برای تشخیص بیماری‌های کرونری قلب است. اما این آزمایش تهاجمی بوده و بعضاً دارای عوارض جانبی می‌باشد. لذا رسیدن به‌روش هایی که بتواند با هزینه کمتر و بدون انجام تست‌های تهاجمی امکان تشخیص با حساسیت و ویژگی مطلوب را در دسترس قرار دهد تحولی مهم در این عرصه محسوب می‌شود. با استفاده از داده های جمع‌آوری‌شده در مطالعات بزرگ و تحلیل آن ها با توجه به فاکتورهای خطر بیماری‌های قلبی شناسایی‌شده، (۳) این امکان وجود دارد که با استفاده از داده‌های جمع‌آوری‌شده و استفاده از تکنیک‌های داده‌کاوی و یادگیری ماشین نسبت به پیش‌بینی و یا احتمال وقوع برخی از بیماری های قلبی نظیر CHD به این مهم دست‌یافت.

در مطالعه سلامت مردم یزد (یاس) Yazd Health Study اطلاعات سلامت و بیماری بیش از ده هزار نفر جمع‌آوری و ثبت شده است که شامل فاکتورهای خطر بیماری‌های قلبی نیز بوده و نیاز به تجزیه و تحلیل آن جهت کشف دانش تشخیص بیماری ها و به‌کارگیری آن جهت استفاده در ارتقای سلامت مردم یزد می باشد. سوالات منتخب از پرسش‌نامه طرح یاس مردم یزد برای پیش‌بینی بیماری قلبی در جدول ۱-الف در پیوست آورده شده است. در این مطالعه با استفاده از تکنیک‌های داده‌کاوی و هوش مصنوعی، راه کاری جهت غربالگری و شناسایی افراد مستعد بیماری قلبی در کوتاه‌ترین زمان ارائه شده است. این کار باعث می‌شود 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 روش‌های داده‌کاوی برای پیش‌بینی بیماری CHD با ۱۰۰۰ نمونه که ظرف شش ماه جمع‌آوری‌شده بود، استفاده شدکه ۲۰۲ نفر آن ها فوت‌شده بودند. روش‌های 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 بعدازآن الگوریتم C5 با ۸۹/۶ درصد دقت قرار دارد. برای ارزیابی روش‌های داده‌کاوی 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 دسته‌بندی (۵) از ۱۳۲۲۸ نمونه (۴۰۵۹ نفر فاقد بیماری عروق کرونر و ۹۱۶۹ نفر مبتلا) شامل ۹ فاکتور خطر بیماری قبلی 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 تولیدشده، دقت ۷۴/۱ درصد به‌دست‌آمده است. با حذف ۵ متغیر و انجام آن با ۴ فاکتور خطر بیماری مجدداً همان دقت حاصل‌شده است که نشان می‌دهد 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 داده‌های گم شده با تعداد ۲۹۷ نمونه نسبت به طراحی سیستم خبره با استفاده از شبکه‌های 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 استفاده‌شده 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 ارائه‌شده۶۸/۷ درصد می‌باشد و نسبت به مقالات مقایسه شده که با شبکه عصبی انجام‌شده 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 تصمیم‌گیری Decision Tree به‌کار گرفته شد و این نتیجه به دست آمد که ماشین بردار پشتیبان با دقت ۸۴/۱ درصد بهترین روش برای پیش‌بینی بیماری های قلبی عروقی است (۷). در مقاله بهداد میرزایی و همکاران باعنوان ارائه یک دسته بند مبتنی بر درخت تصمیم و شبکه عصبی و الگوریتم حرکت جمعی پرندگان برای تشخیص بیماری‌های قلبی از داده‌های 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 انجام‌شده است. تعداد ۱۳ پارامتر مؤثر در بیماری به‌صورت تجربی و به‌کمک آزمایش انتخاب شده اند. عمل تشخیص از روی داده‌ها 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 عمل‌آمده این بود که تنها ۹ عامل در بیماری Coronary Artery Disease CAD مؤثر است و در مقایسه باکارهای انجام‌شده مشابه دارای خطای کمتر و دقت بالاتر بوده است (۳). درمقایسه ای بین سه الگوریتم القای قواعد CN2، الگوریتم SOM Self-Organization Map و درخت تصمیم روی‌داده‌های UCI صورت گرفته است. الگوریتم CN2 توانسته است با دقت ۹۳/۷ درصد بیماری را تشخیص دهد (۱۰). درپیش‌بینی بیماری CHD از روی‌داده‌های 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

همان‌طور که قبلاً بیان شد اکثر تحقیقات انجام‌شده در راستای پیش‌بینی بیماران قلبی بوده است. در این مقاله سعی

در غربالگری افراد مستعد بیماری داریم. از آن جا که داده های استفاده شده جهت غربالگری از مجموعه داده‌های مطالعه یاس می باشد لذا در ابتدا روش جمع آوری داده ها بیان می گردد. سپس نحوه پیش‌پردازش داده ها و در انتها نحوه آنالیز و نتایج حاصل از آن آورده شده است. جهت انجام آنالیز روی داده‌ها از روش‌هایی از قبیل شبکه های عصبی مصنوعی، القای قواعد، رگرسیون و آدابوست استفاده شده است. پیاده سازی این روش‌ها در نرم افزارهای RapidMiner Studio (محصول شرکت رپیدماینر واقع در شهر بوستن Boston آمریکا) و Orange (محصول دانشگاه لیوبلیانا**Ljubljana** اسلوونی) صورت گرفته است.

نمونه گیری مطالعه یاس به صورت خوشه‌ای انجام‌شده 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 آینده نگر است که در سطح شهر یزد انجام‌شده است. جامعهٔ آماری مطالعه افراد بالای ۲۰ تا ۶۹ سال ساکن شهرستان یزد بوده اند. روش نمونه‌گیری این مطالعه دومرحله‌ای و بدین شرح بوده است: نمونه‌گیری دسته‌بندی برحسب 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 به‌تکمیل پرسش نامه اقدام شده و خانوارهای بعدی به‌ترتیب انتخاب گردیدند. درصورتی‌که در یک پلاک چند خانوار وجود داشته (مثل مجتمع‌های مسکونی)، از واحد اول شروع و بعد به واحدهای بعدی مراجعه شده است. درصورتی‌که 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.

پرسش گران در زمینه‌های پرسش‌گری، اخذ رضایت آگاهانه و رعایت اصول اخلاق پژوهش آموزش دیده و پس از شرکت در امتحان تئوری (پروتکل مطالعه) و عملی (پرسش گری و اندازه‌گیری فشارخون و شاخص‌های آنتروپومتریک) برای انجام مصاحبه تأیید شده اند. جمع‌آوری اطلاعات موردنیاز از طریق پرسش نامه و به صورت مصاحبه انجام شده است. پرسش نامه دارای پاسخ‌نامه قابل خوانده شدن به روش الکترونیکی بوده و توسط رایانه تصحیح گردیده است. هم چنین در این مطالعه فشارخون، قد و وزن افراد در منازل اندازه‌گیری می‌شود. به افراد دعوت نامه 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 پرسش نامه معتبر در نظرگرفته شده است. جزییات روش مطالعه قبلاً منتشر شده است (۱۲).

از آن جا که تنها داده های مربوط به فاکتورهای خطر بیماری قلبی مد نظر این مقاله است لذا با توجه به پارامترهای انتخاب‌شده از مقاله های سن و کیم (۳،۱۱) به عنوان پارامترهای اصلی در بیماری قلبی و بیماری CHD و هم چنین 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 داده ها مرحله قبل از تحلیل داده ها می‌باشد. داده های به‌دست‌آمده در فرآیند جمع آوری 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 داده کاوی است داده شده است. تعداد داده‌های جمع‌آوری‌شده در این پژوهش ده هزار رکورد بوده که بعد از عملیات پاک سازی به روش حذف داده های گم‌شده Missing Value به تعداد ۸۱۸۸ رکورد تقلیل پیدا کرده است. بعد از آماده سازی داده ها نوبت به طبقه بندی داده ها می‌رسد. با توجه به وجود داده های وضعیت بیماری قلبی در پرسش نامه از روش های طبقه بندی استفاده می‌شود. برای انجام این کار و بالابردن دقت پیش بینی انجام کار از روش‌های گوناگون طبقه بندی استفاده‌شده است. در ادامه هر یک از آن‌ها به‌طور خلاصه معرفی‌شده‌اند. یکی از

روش های دسته‌بندی استفاده از ماشین‌های بردار پشتیبان می‌باشد. این روش عموماً برای مسائلی که در آن‌ها دو دسته داده وجود دارد، استفاده می‌شوند. در این الگوریتم دسته‌بندی، دو صفحه، در مرز دو کلاس داده‌ها قرارگرفته می‌شود و مسئله یافتن مرز حداکثری بین این دو صفحه و درنتیجه بین دو دسته داده می‌باشد. به این صورت که دو صفحه آن‌قدر از هم دور می شوند که به‌داده‌ها برخورد کنند. هدف یافتن دو صفحه‌ای 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 ازجمله 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 دسته‌بندی داده ها است. این الگوریتم بر اساس مجموعه پارامترها و متغیرها امکان پیش‌بینی نتایج بر پایه احتمال را فراهم می‌کند (۱۳).

رگرسیون لوجستیک Logistic Regression یک روش مبتنی بر آمار است که در یادگیری ماشین با نظارت استفاده می‌شود. یادگیری با نظارت به روشی گفته می‌شود که در زمان آموزش، نتیجه‌های هر نمونه از قبل معلوم و دارای برچسب باشد. در این روش تابعی محاسبه می‌شود که مقدار خروجی آن برحسب متغیرهای ورودی است و تابع می‌تواند به‌صورت خطی یا گوسی پیاده‌سازی شود. وقتی خروجی تابع گوسی آن با توزیع برنولی جایگزین شود برای مسائل دسته‌بندی قابل‌استفاده است. رگرسیون لوجستیک حالت خاص پیاده‌سازی 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 می‌توان از طریق شبکه‌های عصبی مصنوعی نظارت کرد که دارای به‌روزرسانی های مداوم بسیاری از متغیرها مانند ضربان قلب، فشارخون و غیره باشند. عمل آموزش در این روش نسبتاً زمان‌گیر و زمان اجرای آن پس از یادگیری بسیار مناسب است (۱۳). دسته‌بند القای قواعد CN2 CN2 Rule Inducer توسط Clark & Niblett درسال ۱۹۸۹ ارائه شد است. این الگوریتم جستجوی شعاعی عمومی به‌خصوصی را توصیف و از آنتروپی 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 می‌شود و بعد از آن اتصال‌هایی افزوده می شود و تالی قواعد بر اساس کلاس اکثریت نمونه هایی که توسط قواعد، پوشش داده می‌شوند 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 زیرمجوعه 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 الگوریتم‌های یادگیری است. روند کلی این الگوریتم بدین‌صورت است که مجموعه ای وزن دار از تعداد زیادی دسته‌بندی کننده ضعیف به‌عنوان دسته‌بندی کننده نهایی انتخاب می‌نماید. به‌عبارت دیگر آدابوست با استفاده از مجموعه ای از دسته‌بندی کننده های ضعیف می تواند منجر به‌یک 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 نمی‌باشد اما برای داده‌های 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 به‌مجموعه داده ای گفته می‌شود 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

دسته بندی این داده ها یافتن نمونه‌های بیماری است. به این ترتیب کلاس اقلیت نسبت به‌سایر نمونه ها دارای اولویت بیشتری است. در اغلب الگوریتم های دسته‌بندی، تمایل در جهت کلاسی است که بیشترین تعداد نمونه ها را دارد. از این رو توانایی کمی در پیش گویی صحیح داده های کلاس اقلیت از خود نشان می دهد. در نتیجه، قوانین دسته بند سبب می شوند تا داده های کلاس اقلیت، نادرست دسته‌بندی شوند. برای حل این مشکل باید از روش های مواجهه با داده های نامتوازن استفاده نمود. تکنیک‌های متنوعی برای حل مسئله در ارتباط با کلاس نامتوازن پیشنهادشده است، که در سه گروه: روش سطح داده، روش سطح الگوریتم و روش حساس به هزینه تقسیم می‌شوند. در ادامه به توصیف هرکدام 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 داده های نامتوازن می شود.

**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 طوری‌که هم در سطح داده و هم در سطح الگوریتم تغییر ایجاد خواهد کرد. مهم ترین نقطه‌ضعف این روش، تعریف هزینه دسته بندی نادرست است که عموماً اطلاعاتی در مورد 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 کارائی 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 می شود. 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 آن ها مثبت بوده و الگوریتم دسته بندی نیز دسته آنها را به درستی مثبت تشخیص داده است.

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

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

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

جدول ۱: ماتریس درهم ریختگی

| | پیش‌بینی | | |
|---|---|---|---|
| واقعی | Positive | Negative | ∑ |
| Positive | TP | FN | P=TP+FN |
| Negative | FP | TN | N=FP+TN |
| Total | TP+FP | FN+TN | |

جدول۲ : 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. جدول۲ نحوه محاسبه برخی از معیارهای ارزیابی بر اساس مقادیر ماتریس در هم ریختگی را نشان می دهد. دقت یا نرخ تشخیص یکی از مهم ترین معیارها برای کارایی الگوریتم یک دسته بند است که نشان‌دهنده میزان پیش بینی صحیح نسبت به‌کل نمونه ها است. با توجه به این که معیار دقت دسته‌بندی، ارزش رکوردهای دسته‌های مختلف را یکسان در نظر می گیرد در مسائلی که با دسته‌هایی با داده های نامتوازن سروکار داریم از معیارهای دیگری استفاده می‌شود. به‌عنوان مثال در پزشکی اغلب تعداد داده های 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 Detection Rate DR و FAR False Alarm Rate اهمیت ویژه ای 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 دسته بندها 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 کننده ای 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 می باشد. در این روش داده ها به دو

قسمت داده‌های آموزشی و داده های آزمون تقسیم بندی می‌شوند. در روش جامع k-Fold Cross Validation داده‌ها به k قسمت مساوی تقسیم‌شده و به تعداد دفعات k، یک قسمت به‌عنوان داده آزمون و بقیه به‌عنوان داده آموزشی مورد استفاده قرار می‌گیرد. میانگین خطای k مرحله، به عنوان خطای دسته بند تعیین می‌گردد. معمول‌ترین مقداری که در متون علمی برای k در نظر گرفته می‌شود برابر با ۱۰ می‌باشد. در روش k-Fold Cross Validation فرض بر آن است که عملیات انتخاب نمونه‌های آموزشی بدون جای گذاری صورت می‌گیرد. به بیان دیگر یک رکورد تنها یک بار در یک فرآیند آموزشی مورد توجه واقع می‌شود. چنانچه هر رکورد در صورت انتخاب شدن برای شرکت در عملیات یادگیری دسته‌بند بتواند مجدداً برای یادگیری مورد استفاده قرار گیرد روش مزبور با نام Bootstrap شناخته می‌شود. (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 نرم‌افزار رپیدماینر انجام شده است. در اکثر کارهای مشابه انجام شده از داده های مربوط به افراد مراجعه کننده به بیمارستان ها استفاده است. به‌عنوان مثال مجموعه داده جمع‌آوری 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. به‌عبارت 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 متوازن‌سازی انجام شده است. در ادامه نتایج حاصل از دسته بندی داده های موجود در طرح یاس مورد بررسی قرار گرفته است.

### 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 نمونه‌گیری، ساکنان شهرستان یزد می‌باشند که به صورت تصادفی خوشه ای نمونه گیری شده‌اند و نسبت افراد سالم به بیماران قلبی ۱۰ به ۱ می‌باشد.

برای برقراری تعادل بین دو کلاس نسبت به متوزان‌سازی در سطح داده و با روش افزایش کلاس اقلیت به روش بوت استراپ 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 دسته‌بند مختلف جهت ارزیابی کارایی الگوریتم های مختلف برای تشخیص افراد مستعد بیماری استفاده شده است. این دسته‌بندها عبارت اند از: بیز، شبکه های عصبی، بردار پشتیبان ماشین، CN2-SD، درخت تصمیم، نزدیک‌ترین همسایه، آدابوست و رگرسیون لجستیک. سپس به دو روش، دسته بندها مورد ارزیابی قرارگرفته اند. ابتدا با روش 10-Fold cross-validation ارزیابی انجام شده است. سپس جهت بررسی کارایی دسته بندها در محیط واقعی کل نمونه ها به‌عنوان داده آزمون استفاده شده اند. نتایج حاصله به‌صورت ماتریس در هم ریختگی در ادامه آورده شده است. از آن‌جا که هدف از دسته بندی شناسایی افراد مستعد بیماری است معیاری که جهت ارزیابی آن استفاده می شود حساسیت Sensivitity است. این معیار نشان‌دهنده 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.

جدول ۳: ماتریس درهم‌ریختگی حاصل از ارزیابی با 10-fold در ساکنین ۲۰-۶۹ سال یزد در سال۱۳۹۴

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

جدول۴ : نتیجه ارزیابی با 10-Fold cross-validation

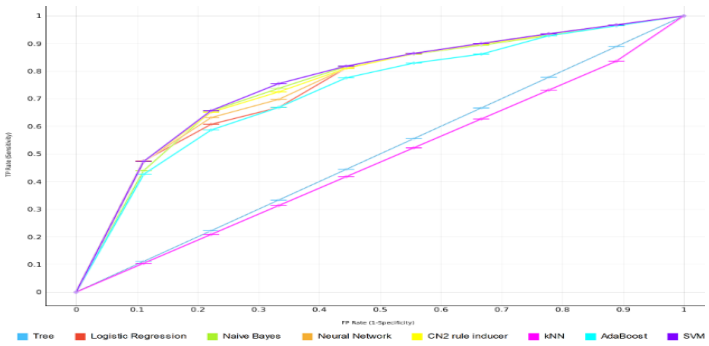
| نام دسته‌بند | سطح زیر نمودار | دقت | حساسیت |
|---|---|---|---|
| آدابوست | ۰/۹۲۵ | ۰/۸۵۵ | ۰/۸۷۱ |
| درخت تصمیم | ۰/۹۱۷ | ۰/۸۴۶ | ۰/۸۷۳ |
| نزدیکترین همسایه | ۰/۹۰۶ | ۰/۸۲۸ | ۰/۸۷۳ |
| ماشین پشتیبان بردار | ۰/۸۵۴ | ۰/۷۹۳ | ۰/۸۰۲ |
| رگرسیون‌لجستیک | ۰/۸۱۹ | ۰/۷۴۵ | ۰/۷۴۱ |
| شبکه عصبی | ۰/۸۱۹ | ۰/۷۴۶ | 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 داده ها

| پیش‌بینی | بردار پشتیبان | | | شبکه عصبی | | |
|---|---|---|---|---|---|---|
| واقعی | 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 | ∑ | بیمار | سالم | ∑ |
| بیمار | ۶۱۷ | ۱۲۱ | ۷۳۸ | ۴۴۹ | ۲۸۹ | ۷۳۸ |
| سالم | ۳۴۸۶ | ۳۹۶۴ | ۷۴۵۰ | ۱۳۹۰ | ۶۰۶۰ | ۷۴۵۰ |
| مجموع | ۴۱۰۳ | ۴۰۸۵ | ۸۱۸۸ | ۱۸۳۹ | ۶۳۴۹ | ۸۱۸۸ |
| پیش‌بینی | رگرسیون‌لجستیک | | | آدابوست | | |
| واقعی | 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 | ∑ | بیمار | سالم | ∑ |
| بیمار | ۰ | ۷۳۸ | ۷۳۸ | ۷۳۸ | ۰ | ۷۳۸ |
| سالم | ۰ | ۷۴۵۰ | ۷۴۵۰ | ۷۴۵۰ | ۰ | ۷۴۵۰ |
| مجموع | ۰ | ۸۱۸۸ | ۸۱۸۸ | ۸۱۸۸ | ۰ | ۸۱۸۸ |

جدول۶: 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 داده‌ها

## بحث

از مقادیر ماتریس درهم‌ریختگی و روابط بیان شده می‌توان 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 بندی‌های 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. همان‌طور که از جدول فوق پیداست، دسته بندهای آدابوست، درخت تصمیم و نزدیک‌ترین همسایه بالاترین دقت را داشته اند و این برتری در معیار سطح زیر نمودار ROC که حاکی از کارایی این دسته بندها است مشهود می‌باشد. 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 داده‌ها قبل از انجام عمل متوازن‌سازی به‌عنوان داده آزمایشی به تمام دسته‌بندها داده شده است. مقادیر ماتریس درهم ریختگی حاصل از انجام عملیات دسته‌بندی روی این داده ها در جدول ۵ آمده است. بر خلاف انتظار سه‌دسته بند مذکور که در ارزیابی به روش 10-Fold cross-validation بهترین کارایی را داشتند با ارزیابی کل داده‌ها به‌عنوان داده آزمون، دارای کمترین کارایی می‌باشند. نتایج این محاسبات در جدول ۶ آمده است. دسته بند ماشین پشتیبان‌بردار بهترین کارایی را در این ارزیابی دارد و بعد از آن دسته بندهای شبکه عصبی، القای قواعد و بیز با سطح زیر نمودار ROC یکسان قرار دارند. از آن‌جا که حساسیت معیار مهمی جهت غربالگری افراد مستعد بیماری می‌باشد در بین این دسته بندها دسته بند القای قواعد با بیشترین حساسیت بهترین عملکرد را دارا می‌باشد. شکل۱ نمودار ROC این ارزیابی را نشان می‌دهد. از آن‌جا 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. به‌دلیل نامتوازن بودن نسبت افراد سالم به بیماران قلبی با تقویت داده های کلاس افراد بیمار به روش بوت استراپ نسبت به متوازن سازی داده‌ها اقدام شده است. نتیجه آموزش و ارزیابی دسته‌بندها به روش 10fold cross validation نشان می‌دهد که روش‌های مختلف ارزیابی تأثیر بسزایی در نتایج این گونه روش متوازن‌سازی دارند و در مقایسه با مقالاتی که از داده‌های UCI استفاده کرده‌اند دقت‌های تقریبا مشابهی را نشان می‌دهند. اما به‌دلیل این که توزیع داده‌های UCI متناسب با

افراد مراجعه‌کننده به بیمارستان است برای تعمیم به کل جامعه جهت استفاده‌هایی هم چون غربالگری افراد مستعد بیماری مناسب نیستند. در مرحله بعد برای بررسی میزان موثربودن دسته بند آموزش‌دیده، ابتدا داده اصلی به‌روش بوت‌استراپ متوازن سپس با استفاده از روش 10fold cross validation آموزش و ارزیابی شده است. داده‌های اصلی (قبل از انجام عمل متوازن‌سازی) نیز توسط دسته بندهای معرفی‌شده در جدول ۶ مورد ارزیابی قرار گرفته است. مقایسه نتایج این دو ارزیابی تاثیر متوازن‌سازی روی دسته بندهای مختلف را نشان می دهد. به‌عنوان مثال در درخت تصمیم نمونه های موجود در کلاس افراد بیمار به دلیل افزایش نمونه با جایگزینی در روش بوت‌استراپ، دارای پراکندگی کمتری هستند.

در ارزیابی دوم نتایج به سمت افراد سالم متمایل شده و مجدداً به‌دلیل 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 دسته‌بند های پارامتریک هستند که رفتار مشابهی از خود به‌روز داده اند. این دسته‌بند ها با توجه به این که کلیه داده‌های آزمون در داده های آموزش وجود دارد مستعد بیش برازش Over Fitting بوده و خطای تعمیم پذیری بالایی دارند. اما در دسته‌بند CN2-SD با توجه به ترکیب الگوریتم القای قواعد با الگوریتم قواعد انجمنی و کشف زیرگروه‌های معنادار، تقویت داده های کلاس افراد بیمار باعث تشکیل قواعد معنادار در کلاس افراد بیمار می‌شود که حساسیت بالای شناسایی افراد بیمار در این دسته بند بیان گر همین موضوع می‌باشد. این روش با توجه به هرس های زیادی که انجام می‌دهد در معرض بیش برازش نمی باشد و دارای قدرت تعمیم پذیری بالاتری است. به‌طوری که تنها با قواعد تولیدشده در نرم افزار دقتی برابر با دسته‌بند 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 نمونه‌های موجود طرح یاس می باشد. برای رفع این مشکل نمونه هایی که فاقد این مقادیر بودند از نمونه های اولیه حذف شده اند.

## سپاس‌گزاری

این مقاله بخشی از پایان‌نامه کارشناسی‌ارشد رشته مهندسی کامپیوتر گرایش نرم‌افزار دانشگاه علم و هنر می‌باشد که بدون حمایت مالی انجام شده است. در پایان از تمامی افرادی که در انجام این پایان‌نامه همکاری نموده‌اند تشکر می‌گردد.

**تعارض در منافع:** وجود ندارد

## 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, Gyeongiu, 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 CN۲ 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]

**Journal of Shahid Sadoughi University of Medical Sciences**
**Vol 27| NO 3 | June 2019**
**Received: 23 May 2018**
**Accepted: 17 Nov 2018**

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

**Seyed Mohammad Reza Tabatabaei Nodoushan[1], Fatemeh Saadatjoo[2], Masoud Mirzaei[*3]**

**Original Article**

**Introdution:** 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

[1]Department of Computer Engineering, Science and Arts University, Yazd, Iran.
[2]Department of Computer Engineering, Science and Arts University, Yazd, Iran.
[3]Department of Epidemiology, Yazd Cardiovascular Research Center, Shahid Sadoughi University of Medical Sciences, Yazd, Iran
***Corresponding author: Tel:** 09134509917, **email:** masoudmirzaei @yahoo.com