مجله علمی پژوهشی دانشگاه علوم پزشکی شهید صدوقی یزد
دوره سی و یک، شماره هفت، مهر ۱۴۰۲
تاریخ دریافت:۱۴۰۱/۰۷/۳۰
تاریخ پذیرش:۱۴۰۱/۱۲/۲۴

مقایسه عملکرد الگوریتم‌های داده‌کاوی در پیش‌بینی بیماری‌های عروق کرونر قلبی با استفاده از داده‌های مطالعه سلامت مردم یزد (یاس)

اعظم برزگری[۱]، سیده فاطمه نورانی[*۲]، مسعود میرزائی[۳]

مقاله پژوهشی

مقدمه: بیماری‌های قلبی - عروقی اصلی‌ترین عامل مرگ و میر در سراسر جهان بوده و یکی از ده دلیل اول مرگ در ۱۵ سال اخیر می‌باشد. بیماری‌های ایسکمیک قلبی نوعی بیماری قلبی است که به دلیل تنگ شدن شریان‌های تغذیه‌کننده بافت قلب (عروق کرونر) ایجاد می‌شود. هدف از این پژوهش مقایسه الگوریتم‌های داده کاوی در پیش‌بینی زودهنگام بیماری قلبی با توجه به علائم اولیه بیمار می‌باشد.

روش بررسی: در این پژوهش از داده‌های فاز اول مطالعه سلامت مردم یزد (یاس) که شامل ۱۰۰۰۰ شرکت‌کننده و با استفاده از ۲۱ ویژگی آنان مانند سن، نوع درد قفسه سینه، میزان قند خون، وضعیت شغلی، مصرف الکل، شاخص توده بدنی و غیره که از سال ۱۳۹۳ تا کنون جمع‌آوری شده بود استفاده شد.

نتایج: تجزیه و تحلیل داده‌ها جمع‌آوری شده با استفاده از الگوریتم‌های Random Forest و Naive Bayes، دقت ۷۴/۵۱ درصد را در پیش‌بینی بیماری کرونر قلبی نشان داد.

نتیجه‌گیری: می‌توان نتیجه گرفت که با الگوریتم‌های ساده فوق می‌توان پیش‌بینی بیماری ایسکمیک قلب را با دقت بالا انجام داده و با غربالگری زودهنگام و درمان به‌موقع در مراحل اولیه باعث کاهش مرگ و میر مرتبط شد.

واژه‌های کلیدی: داده‌کاوی، غربالگری، بیماری ایسکمیک قلبی - عروقی، Naïve Bayes، Random Forest، مطالعه سلامت مردم یزد

ارجاع: برزگری اعظم، نورانی سیده فاطمه، میرزائی مسعود. **مقایسه عملکرد الگوریتم‌های داده کاوی در پیش‌بینی بیماری‌های عروق کرونر قلبی با استفاده از داده‌های مطالعه سلامت مردم یزد (یاس)**. مجله علمی پژوهشی دانشگاه علوم پزشکی شهید صدوقی یزد ۱۴۰۲؛ ۳۱ (۷): ۳۵-۶۸۲۴.

۱- معاونت تحقیقات و فناوری، دانشگاه علوم پزشکی شهید صدوقی، یزد، ایران.
۲- دانشکده فنی و مهندسی، گروه کامپیوتر و فناوری اطلاعات، دانشگاه پیام نور، تهران، ایران.
۳- مرکز تحقیقات قلب و عروق، پژوهشکده بیماریهای غیرواگیر، دانشگاه علوم پزشکی شهید صدوقی، یزد، ایران.
*(نویسنده مسئول)؛ تلفن: ۰۹۱۹۷۷۲۴۶۵۷، پست الکترونیکی: sf.noorani@pnu.ac.ir، صندوق پستی: ۴۶۹۷- ۱۹۳۹۴

[DOI: 10.18502/ssu.v31i7.13693]

مقدمه

طبق گزارش سازمان بهداشت جهانی بیماری‌های قلبی علت اصلی مرگ و میر در جهان و ۸۲ درصد مرگ و میرها در کشورهای در حال توسعه است. همچنین بر اساس گزارش وزارت بهداشت و سازمان بهداشت جهانی، ۳۵ درصد علل مرگ و میر در ایران بر اثر بیماری‌های قلبی است (۱). به دنبال بیماری‌های قلبی، بیماران مشکلات متعددی مانند درد، تغییر در جریان خون بافتی، تحمل نکردن فعالیت، ناسازگاری با بیماری، استرس مزمن، اضطراب و تظاهرات روانی شدید را تجربه می‌کنند. لذا با وجود این مشکلات روند بهبودی بیماری به تأخیر می‌افتد و احتمال مرگ در ماه‌های اول افزایش می‌یابد (۲). به‌دلیل وجود تعداد کم علائم در بیماری‌های قلب و عروق، پیش‌بینی ابتلا به این بیماری‌ها به روش‌های سنتی، زمان بر بوده یا دشوار است. با پیش‌بینی زودهنگام و کسب بینش نسبت به آینده وضعیت بیماری در افراد، می‌توان با اتخاذ روش‌های پیشگیری از ابتلا به این نوع بیماری‌ها، هزینه‌های اقتصادی درمان‌های مداخله‌ای و تهاجمی در سطوح خانواده و جامعه و کشور را تا حد قابل‌ملاحظه‌ای کاهش داد (۳). از آنجا که زمان در پیش‌بینی بیماری‌های قلب و عروق از اهمیت زیادی برخوردار است، بهره‌برداری از تکنیک‌های داده‌کاوی به دلیل استنتاج از حجم زیادی از داده‌های موجود در مدت زمان کوتاه کارآمدتر است. در انتخاب تکنیک های یادگیری ماشین، تمام عوامل ذکر شده در هنگام تجزیه و تحلیل و درک بیماران توسط پزشک از طریق معاینات دستی در فواصل زمانی معین در نظر گرفته می‌شود. تکنیک‌های مختلف داده‌کاوی متناسب با حوزه‌های علمی و تخصصی متنوع ارائه شده و این تکنیک‌ها به‌راحتی برای توسعه چارچوب‌ها یا یافتن استنتاج‌ها و نتیجه‌گیری‌های مهم از مجموعه داده‌های به‌دست آمده استفاده می‌شوند (۴). پژوهش‌هایی که تاکنون انجام‌شده حاکی از این است که پیش‌بینی ابتلا به این بیماری در مراحل اولیه دشوار بوده بنابراین، توسعه نرم‌افزاری که از فاکتورهای شناخته شده این بیماری، در انتخاب الگوریتم‌های تخصصی بهره‌برداری کند، می‌تواند آسیب‌پذیری بیماری‌های قلبی را با توجه به علائم اولیه پیش‌بینی کرده و هزینه‌های درمان را نیز کاهش می‌دهد (۵). در مطالعه سلامت مردم یزد (یاس) اطلاعات سلامت و بیماری بیش از ده هزار نفر از ساکنان شهرستان یزد از طریق پرسش‌نامه الکترونیکی جمع‌آوری و ثبت شده است که شامل فاکتورهای خطر بیماری‌های قلبی نیز بوده است. تجزیه و تحلیل این داده‌ها جهت ارتقا سلامت مردم یزد و تشخیص زودهنگام بیماری قلبی مفید می‌باشد. با توجه به افزایش تورم و هزینه‌های واردات و یا ساخت تجهیزات پزشکی تشخیصی و تجهیزات مورد نیاز برای درمان بیماری‌های قلب و عروق که عموماً نیاز به استفاده از انواع روش‌های مداخله‌ای مانند آنژیوپلاستی و جراحی قلب دارد، استفاده از روش‌های غربالگری کم‌هزینه، بیش از پیش حائز اهمیت است. روش داده‌کاوی با پیش‌بینی زودهنگام و آینده‌نگری، کاهش بار هزینه‌های تشخیص و درمان در نظام سلامت کشور را در پی داشته و با ارائه اطلاعات بهینه، به موقع و مرتبط با وضعیت خطر ابتلا به این نوع بیماری در افراد جامعه، این مشکل را مرتفع می‌کند. این مطالعه با هدف پیش‌بینی خطر ابتلا به بیماری عروق کرونر با استفاده از تکنیک‌های داده کاوی در پی آن است تا در مراحل اولیه و با پیشنهاد تغییر سبک زندگی از پیشرفت آن جلوگیری کرد. باقری و همکاران در پژوهشی به مطالعه تشخیص بقا در بیماران نارسایی قلبی با استفاده از داده‌کاوی و دو روش درخت تصمیم و رگرسیون اقدام و نتایج این دو روش را باهم مقایسه نمودند. این تحقیق از دو تکنیک درخت تصمیم و رگرسیون جهت انجام کار پیاده‌سازی کمک گرفته و درنهایت کارایی هر یک مورد بررسی و مقایسه قرار گرفت. نتایج حاصل از این تحقیق نشان داد که میزان دقت تشخیص در این تحقیق با روش درخت تصمیم برابر با ۹۵/۶۵٪ و با تکنیک رگرسیون دارای میزان دقت ۹۱/۲۸٪ است (۶). مطالعه Mahmoodi و همکاران (۷) با هدف طراحی یک سیستم هوشمند برای تشخیص بیماری قلبی با استفاده از کامپیوتر انجام شد. مجموعه داده مورد استفاده ۲۷۰ بیمار مراجعه‌کننده که دارای ۱۳ ویژگی (سن، جنس، ضربان قلب، فشارخون در حالت استراحت، نوع درد قفسه‌سینه، کلسترول

[DOI: 10.18502/ssu.v31i7.13693]

خون، قند خون ناشتا، نتایج الکتروکاردیوگرافی در حالت استراحت و....) بودند توسط تکنیک فازی و الگوریتم ماشین بردار پشتیبان در نرم‌افزار متلب جهت تشخیص درست و سریع که درصد نجات بیمار را افزایش می‌دهد استفاده شد. هم‌چنین معیارهای ارزیابی در این سیستم نرخ دسته‌بندی و حساسیت بود که عملکرد این سیستم بر اساس این شاخص‌ها به ترتیب ۸۵٪ و ۸۵/۸٪ به‌دست آمده بود که سیستم پیشنهادی با دقت نسبتاً بالایی افراد مبتلا به بیماری قلبی را تشخیص داده بود. داده‌کاوی نتایج قابل‌توجهی در پیش‌بینی و کشف بیماری نشان داده است، تکنیک حذف داده‌ها به‌طور گسترده برای پیش‌بینی، شناسایی و برای انواع مختلف بیماری‌های قلبی کاربرد دارد. داده کاوی می‌تواند روش مناسبی برای حمایت از متخصصان پزشکی در تشخیص بیماری با به دست آوردن اطلاعات و دانش در مورد بیماری و علائم از مجموعه داده‌های بیمار باشد. تکنیک‌های حذف اطلاعات شامل روش‌های پنهان برای ایجاد آگاهی در محیط سازمان است. این می‌تواند به‌طور گسترده‌ای برای بهبود از عملکرد همراه با عالی‌بودن تصمیم پزشکی پشتیبانی کند. بر اساس مطالعه تناسبی، تکنیک‌های مختلف داده کاوی برای پیش بینی بیماری‌های قلبی و مشکلات پزشکی مشابه مورد استفاده قرار گرفت. از این‌رو الگوریتم‌های داده کاوی مختلفی وجود دارد که باید مورد استفاده قرار گرفته و از نظر کارایی بالاتر با هم مقایسه شوند (۸). این مطالعه با هدف استفاده از تکنیک داده کاوی و الگوریتم‌های ترکیبی جهت غربالگری و شناسایی زودهنگام افراد مستعد بیماری قلبی، در کوتاه‌ترین زمان، ارائه شده است. یافته‌ها می‌تواند با فراهم‌کردن آموزش و تغییر سبک زندگی باعث کاهش فاکتورهای خطر و افزایش طول عمر و امید به زندگی افراد مستعد به بیماری شود.

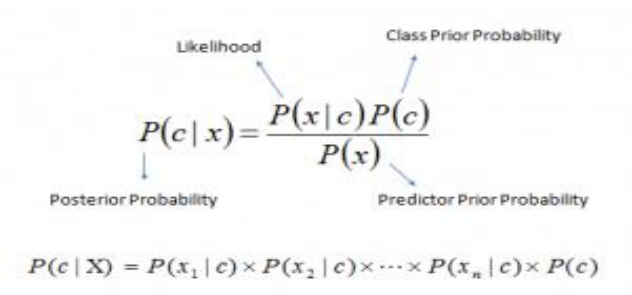
روش بررسی

تحقیق حاضر یک مطالعه کاربردی است که به پیش‌بینی بیماری عروق کرونر قلبی پرداخته است. جامعه آماری استفاده شده در این تحقیق داده‌های فاز اول مطالعه سلامت مردم یزد که روی ۱۰۰۰۰ نفر از ساکنان ۶۹-۲۰ سال شهرستان یزد که در طی سال‌های ۱۳۹۴-۱۳۹۳ جمع‌آوری شده، می‌باشد. در ادامه به نحوه جمع‌آوری داده‌ها پرداخته می‌شود. نمونه‌گیری مطالعه یاس به‌صورت خوشه‌ای تصادفی در دو مرحله روی ۱۰۰۰۰ نفر انجام شده است. روش نمونه‌گیری این مطالعه دومرحله‌ای و بدین شرح بوده است: مرحله اول، در هر بلوک، بر اساس لیست فهرست برداری خانوار سال ۱۳۹۲، ۵۰ سرخوشه انتخاب و با حرکت از سمت راست نسبت به تکمیل پرسش‌نامه اقدام شده و خانوارهای بعدی به‌ترتیب انتخاب شدند. در صورتی‌که در یک پلاک چند خانوار وجود داشت (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. پرسشگران در زمینه‌های پرسشگری، اخذ رضایت آگاهانه و رعایت اصول اخلاق پژوهش آموزش دیده و پس از شرکت در آزمون تئوری (پروتکل مطالعه) و عملی (پرسشگری و اندازه‌گیری فشارخون و شاخص‌های آنتروپومتریک) برای انجام مصاحبه تأیید شدند. جمع‌آوری اطلاعات مورد نیاز از طریق پرسش‌نامه و به‌صورت 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 شد. 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 ۸۱٪ بوده بنابراین پرسش‌نامه معتبر در نظر گرفته شده است. جزییات روش مطالعه قبلاً منتشر شده است (۹).

الگوریتم نایوبیز (Naïve Bayes): نایوبیز یک الگوریتم یادگیری ساده است که از قانون بیز همراه با یک فرض قوی مبنی بر اینکه ویژگی‌ها با توجه به کلاس مستقل هستند، استفاده می‌کند. در حالیکه این فرض استقلال اغلب در عمل نقض می‌شود، با این وجود نایوبیز اغلب دقت طبقه‌بندی رقابتی را ارائه

[DOI: 10.18502/ssu.v31i7.13693]

می‌دهد. همراه با کارایی محاسباتی و بسیاری از ویژگی‌های مطلوب دیگر، این امر منجر به استفاده گسترده نایوبیز در عمل می‌شود (۱۰). ساخت مدل نایوبیز آسان است و به ویژه برای مجموعه داده‌های بسیار بزرگ مفید است. در عین سادگی، از روش‌های طبقه‌بندی بسیار پیچیده نیز بهتر عمل می‌کند. قضیه بیز راهی برای محاسبه احتمال خلفی P(c|x) از P(c)، P(x) و P(x|c) ارائه می‌دهد. به معادله زیر نگاه کنید (۱۱).



(x، ویژگی‌ها) است.

P(c) احتمال قبلی کلاس است.

P(x|c) احتمالی است که احتمال کلاس داده شده پیش‌بینی کننده است.

P(x) احتمال قبلی پیش‌بینی‌کننده است.

الگوریتم جنگل تصادفی (Random Forest): جنگل تصادفی یک الگوریتم جدید یادگیری ماشین و یک الگوریتم ترکیبی جدید است که ترکیبی از یک سری طبقه‌بندی‌کننده ساختار درختی است. جنگل تصادفی به‌طور گسترده در طبقه‌بندی و پیش‌بینی استفاده شده است و در رگرسیون نیز استفاده می‌شود. در مقایسه با الگوریتم‌های سنتی، جنگل تصادفی دارای ویژگی‌های خوبی است. بنابراین دامنه کاربرد جنگل تصادفی بسیار گسترده است (۱۲).

الگوریتم درخت تصمیم (Decision Tree): الگوریتم درخت تصمیم یکی از تکنیک‌های پرکاربرد در داده‌کاوی، سیستم‌هایی است که طبقه‌بندی کننده‌ها را ایجاد می‌کنند در داده‌کاوی، الگوریتم‌های طبقه‌بندی قادر به مدیریت حجم وسیعی از اطلاعات هستند. می‌توان از آن برای ایجاد مفروضات در مورد نام‌های طبقه‌بندی‌شده، طبقه‌بندی دانش بر اساس مجموعه‌های آموزشی و برچسب‌های کلاس و طبقه‌بندی داده‌های تازه به‌دست‌آمده استفاده کرد (۱۳). در این تحقیق از ۲۱ سؤال مرتبط پرسش‌نامه فاز اول طرح یاس استفاده شده است. قبل از تحلیل داده باید آن‌ها را پاکسازی کرد. داده‌ها را از فایل اکسل به محیط نرم‌افزار رپیدماینر که یکی از نرم‌افزارهای داده‌کاوی است وارد کرده است. تعداد داده‌های جمع‌آوری‌شده در این پژوهش ۱۰۰۰۰ رکورد بوده که بعد از عملیات پاک‌سازی به روش حذف داده‌های گم‌شده (Missing Value) به تعداد ۹۹۶۶ 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 گوناگون استفاده شد.

جهت ارزیابی معیارها روش‌های مختلفی وجود دارد که در این پژوهش برای ارزیابی معیارها از ماتریس درهم ریختگی (Confusion Matrix) استفاده شده است.

- معیارهای ارزیابی شده
  - Accuracy (دقت)
  - Precision (صحت)
  - Recall (بازخوانی)
  - F Score (حاصل از میانگین‌هارمونیک دقت و بازخوانی)
- ماتریس درهم ریختگی

TP: نشان دهند تعداد رکوردهایی است که دسته واقعی آن‌ها مثبت بوده و الگوریتم دسته‌بندی نیز دسته آن‌ها را به‌درستی مثبت تشخیص داده است. (پیش‌بینی بله است و آنها این بیماری را دارند).

TN: نشان دهنده تعداد رکوردهایی است که دسته واقعی آن‌ها منفی بوده و الگوریتم دسته‌بندی نیز دسته آن‌ها را به‌درستی منفی تشخیص داده است. (پیش‌بینی منفی است و آن‌ها این بیماری را ندارند).

FP: نشان دهنده تعداد رکوردهایی است که دسته واقعی آن‌ها منفی بوده و الگوریتم دسته‌بندی دسته آن‌ها را به اشتباه مثبت تشخیص داده است. (ما پیش‌بینی کردیم بله، اما آنها در واقع این بیماری را ندارند. (هم‌چنین به عنوان "خطای نوع اول" شناخته می‌شود).

[DOI: 10.18502/ssu.v31i7.13693]

FN: نشان‌دهنده تعداد رکوردهایی است که دسته واقعی آن‌ها مثبت بوده و الگوریتم دسته‌بندی، آن‌ها را به اشتباه منفی تشخیص داده است. (یعنی ما پیش‌بینی کردیم آن‌ها بیماری ندارند، اما آن‌ها در واقع این بیماری را داشتند). جهت ارزیابی دسته‌ها از مقادیر ماتریس درهم ریختگی استفاده می‌شود. جدول ۱ و ۲ نحوه محاسبه معیارهای ارزیابی را براساس ماتریس درهم ریختگی نشان می‌دهد. یکی از مهم‌ترین معیارها از بین معیارهای استفاده شده برای کارایی الگوریتم، معیار دقت با نرخ تشخیص است که میزان پیش‌بینی صحیح به کل نمونه‌ها را نشان می‌دهد. داده‌های مورد استفاده در این پژوهش مجموعه داده یاس (مطالعه سلامت مردم یزد) می‌باشد که شامل ۱۰۰۰۰ رکورد و ۳۰۰ پارامتر (متغیر) در فاز اول بود که از ۲۱ پارامتر از ۳۰۰ پارامتر در این پژوهش استفاده شد. داده‌ها شامل ۲۱ ستون مانند سن، جنس، قند خون در حال استراحت، درد قفسه سینه، کلسترول سرم، قندخون ناشتا، نتایج الکتروگرافی در حالت استراحت و غیره بود که با الگوریتم‌های منتخب نایوبیز و جنگل تصادفی پیاده‌سازی شد.

جدول ۱: ماتریس درهم ریختگی

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

جدول ۲: نحوه محاسبه معیارهای ارزیابی

Evaluation Methods	Equations
Accuracy	$\frac{(TP+TN)}{(TP+FN+FP+TN)}$
Precision	$\frac{TP}{(TP+FP)}$
Recall	$\frac{TP}{(TP+FN)}$
F Score	$\frac{2\times(Precision\times Recall)}{(Precision+Recall)}$

مراحل انجام پژوهش

مراحل انجام تحقیق به‌صورت استاندارد (CRISP: Cross-Industry Standard Process) به روش زیر می‌باشد:

مرحله اول: جمع‌آوری و پیش پردازش داده‌ها

جمع‌آوری داده از طریق پرسش‌نامه انجام شده پرسش‌نامه به روش الکترونیکی بود و پاسخ‌های ثبت شده در پرسش‌نامه‌های اسکن شده توسط رایانه خوانده شده در این گام به جمع‌آوری داده‌های اولیه، توصیف داده‌ها، بازرسی و بررسی داده‌ها پرداخته شد.

آماده‌سازی داده: در ابتدا برای جمع و آماده‌سازی داده‌ها از کوئری‌های Select، Where، Top و Distinct کوئری‌های Join کردن در جداولی مانند Inner Join و ساخت View در نرم‌افزار SQL، استفاده گردید. نرم‌افزار رپیدماینر مجهز به ابزارهای بسیار قوی است تا بتواند مجموعه داده را در پایگاه داده

[DOI: 10.18502/ssu.v31i7.13693]

داخلی یا محلی نرم‌افزار بارگذاری نموده و این مجموعه داده را برای ارائه به عملگرهای یادگیری مدل آماده کند.

مرحله دوم: مدل‌سازی

در مدل‌سازی روش‌های داده کاوی زیادی وجود دارد. در این مرحله تکنیک‌های مختلف داده‌کاوی به رسم مدل و الگوی بهبود یافته می‌پردازیم.

مرحله سوم: نتایج

در این مرحله پیش‌بینی می‌گردد که دقت هر مدل چند درصد می‌باشد.

مرحله چهارم: ارزیابی

برای رسیدن به نتیجه و هدف در این مرحله مدل ارزیابی می‌شود تا ببینیم آیا به هدف رسیده‌ایم یا نه؟ قسمت‌هایی که نتیجه بخش نبوده و به هدف نرسیده را تکرار می‌کنیم یا بعضی مواقع ممکن است به تغییر هدف تبدیل شود و یا مجبور به تغییر اعداد اولیه شود.

مرحله پنجم: توسعه

پایان یک پروژه ساخت مدل نیست و هدف از کشف دانش و استفاده از این دانش کشف‌شده در آینده است.

تجزیه و تحلیل آماری

داده‌ها با استفاده از الگوریتم‌های ترکیبی و نرم‌افزار Rapid Miner نسخه ۷ (محصول شرکت رپیدماینر شهر بوستون آمریکا) تجزیه و تحلیل و پیاده سازی شد. برای ارزیابی داده‌ها و هم‌چنین میزان کیفیت پیش‌بینی مدل‌ها دسته‌بند از عملگر X-Validation استفاد و جهت حداقل کردن واریانس مدل از تکنیک Bagging استفاده شد و در نهایت جهت بهبود دقت تشخیص از عملگر Vote استفاده کردیم.

ملاحظات اخلاقی

پروپوزال این تحقیق توسط دانشگاه علوم پزشکی شهید صدوقی یزد تایید شده است (کد اخلاق: IR.SSU.REC.1401.016)

نتایج

داده‌هایی که از طریق بکارگیری ابزارهای جمع‌آوری در نمونه (جامعه) آماری فراهم آمده‌اند، خلاصه، کدبندی و دسته‌بندی و در نهایت پردازش می‌شوند تا زمینه برقراری انواع تحلیل‌ها و ارتباط‌ها بین این داده‌ها به منظور آزمون فرضیه‌ها و پاسخ به سؤالات تحقیق فراهم آید. بدین منظور، در ادامه به پرسش‌های پژوهش پاسخ داده می‌شود. مدل‌سازی با استفاده از عملگر جنگل تصادفی، الگوریتم درخت تصمیم و عملگر نایوبیز مدل‌های مورد استفاده قرار گرفته در این پژوهش، ترکیبی از عملگر جنگل تصادفی با استفاده از الگوریتم درخت تصمیم با پارامترهای مختلف و عملگر نایوبیز بود، در این مدل‌سازی پارامترهای مختلف با حالات و مقادیر مختلف مورد بررسی قرار گرفت آزمایش و همچنین در وضعیت عدم هرس و هرس کردن، که بهترین و بالاترین دقت به‌دست آمده از مدل‌سازی با عملگر جنگل تصادفی با استفاده از الگوریتم درخت تصمیم با پارامترهای ذکر شده در جدول ۳ نشان داده شده است.

ارزیابی داده‌ها: جدول ۴ داری دو ستون عمودی به نام سالم و بیمار (دسته واقعیت که همان دیتاست می‌باشد) و دو ستون افقی سالم و بیمار (دسته پیش‌بینی) می‌باشد. در دسته، در پاسخ به این سوال که آیا بیماری قلبی بوده یا نه مقدار ۱ داشته یعنی فرد بیماری قلبی داشته در دسته واقعیت بیماری را تشخیص می‌دهد و در دسته پیش‌بینی بیماری را پیش‌بینی می‌کند. چیزی که مدل تشخیص داده برای مدل ترکیبی این است: جمع ستون سالم (۲۷۶۳=۲۵۱۸+۲۴۵) مدل تشخیص داده که ۲۴۵ تا درست تشخیص داده که تقسیم بر ۲۵۱۸ می‌شود و ۸/۸۷ درصد دیتا را درست تشخیص داده است. در دسته واقعیت دوم بیمار جمع ستون (۷۲۰۳=۲۲+۷۱۸۱) که مدل ۷۱۸۱ را با دقت ۹۹/۶۹ درصد درست تشخیص داده است. در قسمت ستون‌های افقی سالم جمع ستون (۲۶۷=۲۴۵+۲۲) که مدل ۲۴۵ تا را با دقت ۹۱/۷۶ درصد به درستی تشخیص داده است. در قسمت ستون‌های افقی بیمار (۹۶۹۹=۲۵۱۸+۷۱۸۱) که مدل ۷۱۸۱ تا را با دقت ۷۵/۰۴ درصد به درستی تشخیص داده است. بقیه مدل‌های جدول نیز مشابه این توضیحات می‌باشد. طبق نتایج به‌دست آمده از جدول ۵ مشاهده شد که مدل ترکیبی جنگل تصادفی و نایوبیز جهت پیش‌بینی و طبقه‌بندی بهترین عملکرد را نسبت به استفاده از این مدل‌ها به صورت تفکیکی داشته است و دقت ۷۴/۵۱ درصد و صحت ۹۹/۶ درصد را نشان داده است.

[DOI: 10.18502/ssu.v31i7.13693]

جدول ۳: پارامترهای استفاده شده در مدل‌سازی با عملگرها

عملگر	پارامتر	مقدار/حالت	دقت مدل
Random Forest	معیار برش	Gini_index	۷۵ درصد
	حداکثر تعداد عمق	۱۰	
	نسبت نمونه آزمایش	۰.۳	
	عدم هرس و وجود هرس	وجود هرس	
Naive Bayes	Laplace correction	وجود اصلاح لاپلاس	

جدول ۴: ماتریس درهم ریختگی ارزیابی با کل داده‌ها

ترکیب جنگل تصادفی و نایوبیز			پیش‌بینی واقعی
جمع	سالم	بیمار	
۹۶۹۹	۲۵۱۸	۷۱۸۱	بیمار
۲۶۷	۲۴۵	۲۲	سالم
۹۹۶۶	۲۷۶۳	۷۲۱۳	مجموع
نایوبیز			پیش‌بینی واقعی
جمع	سالم	بیمار	
۷۵۵۸	۱۳۷۵	۶۱۸۳	بیمار
۲۴۰۸	۱۳۸۸	۱۰۲۰	سالم
۹۹۶۶	۲۷۶۳	۷۲۰۳	مجموع
جنگل تصادفی			پیش بینی واقعی
جمع	سالم	بیمار	
۹۷۹۵	۲۶۰۱	۷۱۹۴	بیمار
۱۷۱	۱۶۲	۹	سالم
۹۹۶۶	۲۷۶۳	۷۲۰۳	مجموع

جدول ۵: نتیجه ارزیابی با کل داده‌های فاز اول مطالعه سلامت مردم یزد ۱۳۹۳-۹۴

عملگرهای مورد استفاده .	Accuracy	Precision	Recall	F-Score
Decision Tree	79.75%	99%	78.66%	87.67%
Random Forest	73.81%	99.88	73.45%	84.65%
Naive Bayes	75.97%	85.84%	81.81%	83.78%
Vote, Bagging (Random Forest, Naive Bayes)	74.51%	99.6%	72.05%	85.24%

بحث

هدف از این پژوهش مقایسه طبقه‌بندی بیماری‌های ایسکمیک قلب با توجه به علائم اولیه بیمار و تکنیک‌های داده کاوی بود. با پیش‌بینی و تشخیص زودهنگام این بیماری‌ها می‌توان درمان‌های لازم را در مراحل اولیه انجام داده و باعث کاهش مرگ و میر بیماران شد. در این راستا قبلاً پژوهش‌هایی انجام شده که نتایج آن با نتایج این پژوهش همسو می‌باشد. به‌طور مثال Rubini و همکاران (۱۴) پژوهشی با هدف غربالگری و طبقه‌بندی بیماری‌های قلبی با توجه به علائم اولیه مانند سن، جنس، ضربان قلب، فشارخون در حالت استراحت، کلسترول، قند خون ناشتا، نتایج الکتروکاردیوگرافی در حالت

[DOI: 10.18502/ssu.v31i7.13693]

استراحت، آنژین ناشی از ورزش، افسردگی ST، ST بخش شیب انجام دادند. این مقاله یک تجزیه و تحلیل مقایسه‌ای از تکنیک‌های یادگیری ماشین مانند جنگل تصادفی (RF:Random Forest)، رگرسیون لجستیک، ماشین بردار پشتیبان (SVM: Support Vector Machin) و Naive Bayes در طبقه‌بندی بیماری‌های قلبی عروقی ارائه داد. با تجزیه و تحلیل مقایسه‌ای، الگوریتم یادگیری ماشین جنگل تصادفی دقیق‌ترین و قابل‌اطمینان‌ترین الگوریتم است و بنابراین در این پژوهش مورد استفاده قرار گرفت. این سیستم هم‌چنین ارتباط بین دیابت و میزان تأثیر آن بر بیماری‌های قلبی را ارائه داد. در اینجا از ۴ الگوریتم استفاده و پیاده‌سازی و نتایج را مقایسه کرده بود. مانند الگوریتم جنگل تصادفی رگرسیون لجستیک؛ ماشین بردار پشتیبان و Naïve Bayes یک تجزیه و تحلیل مقایسه‌ای جهت طبقه‌بندی بیماری ارائه می‌دهد و با توجه به تجزیه وت حلیلی که با این ۴ روش انجام گرفت نشان داد که الگوریتم یادگیری ماشین جنگل تصادفی دقیق‌ترین و قابل‌اطمینان‌ترین الگوریتم است و مورد استفاده قرار می‌گیرد و هم‌چنین ارتباط بین دیابت و میزان تأثیر بر بیماری قلبی را ارائه داده است با استفاده از ۴ الگوریتم جنگل تصادفی، ماشین بردار پشتیبان، رگرسیون لجستیک و Naïve Bayes مجموعه داده‌ها تجزیه و تحلیل شد و الگوریتم جنگل تصادفی بالاترین دقت را ارائه نمود و از این‌رو جنگل تصادفی در سیستم پیشنهادی پیاده‌سازی شده. دقت الگوریتم جنگل تصادفی: ۸۱/۸۴٪، رگرسیون خطی: ۸۳/۸۲٪، و وکتور پشتیبانی: ۷۴/۰۵٪ بود. در پژوهش علی و همکاران (۱۵) از مجموعه داده بیماری قلبی جمع آوری‌شده از سه طبقه‌بندی Kaggle بر اساس الگوریتم‌های k-نزدیک‌ترین همسایه (KNN)، درخت تصمیم (DT) و جنگل‌های تصادفی (RF)، استفاده شد. روش RF دقت ۱۰۰ درصد همراه با حساسیت ۱۰۰ درصد نشان داد. بنابراین، مشخص شد که یک الگوریتم یادگیری ماشینی نظارت شده نسبتا ساده می‌تواند برای پیش‌بینی بیماری قلبی با دقت بسیار بالا استفاده شود. در تحقیق Ishqa و همکاران (۱۶) برای پیش‌بینی بیماران قلبی از نه مدل استفاده کرد، درخت تصمیم (DT)، طبقه‌بندی کننده سازگار (AdaBoost)، رگرسیون لجستیک (LR)، طبقه‌بندی گرادیان تصادفی (SGD) جنگل تصادفی (RF)، طبقه‌بندی کننده افزایش گرادیان (GBM)، طبقه‌بندی کننده درخت اضافی (ETC)، طبقه‌بندی کننده Gaussian Naive Bayes (G-NB) و ماشین بردار پشتیبانی (SVM) مشکل کلاس عدم تعادل توسط تکنیک ابر نمونه‌گیری اقلیت مصنوعی (SMOTE) مدیریت شد. نتایج تجربی نشان داد که ETC در پیش‌بینی بقای بیماران قلبی عملکرد بهتری نسبت به سایر مدل‌ها داشت و با SMOTE به میزان دقت ۰/۶۲۹۲ رسید. Tougui و همکاران (۱۷) در پژوهش خود شش ابزار رایج داده‌کاوی را با هم مقایسه کردند: Orange، Weka، RapidMiner، Knime، Matlab و Scikit-Learn. با استفاده از شش تکنیک یادگیری ماشین: رگرسیون لجستیک، ماشین بردار پشتیبانی، K نزدیکترین همسایگان، شبکه عصبی مصنوعی، بیز ساده و جنگل تصادفی با طبقه‌بندی بیماری قلبی. مجموعه داده مورد استفاده کلیولند که دارای ۱۳ ویژگی، یک متغیر هدف و ۳۰۳ مورد است که در آن ۱۳۹ مورد از بیماری‌های قلبی عروقی و ۱۶۴ فرد سالم هستند. سه معیار عملکرد برای مقایسه عملکرد تکنیک‌ها در هر ابزار استفاده شد: دقت، حساسیت و ویژگی. نتایج نشان داد که Matlab بهترین ابزار و مدل شبکه عصبی مصنوعی Matlab بهترین عملکرد را داشتند. در پژوهش Premsmith و همکاران (۱۸) مدلی برای پیش‌بینی بیماری از تکنیک داده کاوی استفاده کرد. الگوریتم داده‌کاوی از مدل رگرسیون لجستیک و مدل شبکه عصبی استفاده می‌کند. مجموعه داده این مقاله از داده‌های بیماری قلبی در دانشگاه کالیفرنیا ارواین (UCI) با همان ۱۴ ویژگی استفاده شد. معیارهای ارزیابی با استفاده از جدول ماتریس سردرگمی مانند دقت، صحت، فراخوان و اندازه‌گیری F. نتایج نشان داد که مدل رگرسیون لجستیک عملکرد بهتری نسبت به مدل شبکه عصبی دارد. مدل رگرسیون لجستیک دارای دقت ۹۵/۴۵ درصد و دقت ۹۱/۶۵ درصد است. در مطالعه Kavitha و همکاران (۱۹) مدل ترکیبی یک تکنیک جدید است که با استفاده از احتمالات به دست آمده از یک مدل یادگیری ماشین به عنوان ورودی به مدل یادگیری ماشین دیگر داده شد. این مدل ترکیبی بر اساس هر دو الگوریتم یادگیری ماشین که برای پیاده‌سازی‌ها در نظر گرفته شد. کار پیشنهادی با کتابخانه‌های sklearn،

[DOI: 10.18502/ssu.v31i7.13693]

پانداها، matplotlib و سایر کتابخانه‌های اجباری اجرا شده و از مجموعه داده‌های کلیولند و الگوریتم‌های یادگیری ماشینی به همراه مدل ترکیبی مانند درخت تصمیم و جنگل تصادفی استفاده شد. نتایج نشان داد که تشخیص بیماری قلبی با استفاده از الگوریتم جنگل تصادفی و یک مدل ترکیبی موثر است. Decision Tree حدود ۷۹٪ دقت و جنگل تصادفی ۸۱٪ دقت و مدل Hybrid دقت ۸۸٪ را نشان داد. در مطالعه Kazemi و همکاران (۲۰) باهدف پیش‌بینی دقیق‌تر و تصمیم‌گیری مؤثرتر در درمان بیماران انجام شد. داده‌های مورد استفاده در این مطالعه مربوطه به اطلاعات ۲۷۰ بیمار از داده‌های سایت (UCI: University of California Irvine) استخراج شده بود که شامل ۱۴ متغیر بود که با استفاده از الگوریتم شبکه عصبی جهت پیش‌بینی بیماری قلبی و عروقی استفاده شده بود که نتیجه مدل با دقتی برابر ۸۸/۳۳٪ را برای مجموعه مشاهدات نشان داده است. مطالعه‌ای Pavithra و همکاران (۴) داده‌های لازم از بیمار مانند: سن، نوع درد قفسه سینه، میزان قند خون و غیره را برای پیش‌بینی بیماری قلبی مورد استفاده قرار داده بود. نتایج نشان داد که با استفاده از تکنیک داده‌کاوی جمع‌آوری و طبقه‌بندی شده و بیماری به‌راحتی قابل تشخیص بوده است. بنابراین می‌توان درمان لازم را در مراحل اولیه و کاهش میزان مرگ‌ومیر انجام داد. روش تحقیق به‌صورت داده‌کاوی- کتابخانه‌ای (بر اساس داده‌های موجود در بانک اطلاعاتی بیماری‌های قلبی مربوط به ۱۴ پارامتر ارزشمند در تشخیص بیماری قلبی در پایگاه Kaggle) الگوریتم استفاده‌شده، الگوریتم C4.5 یک طبقه‌بندی درخت تصمیم است که خروجی را در داده‌های، داده‌شده طبقه‌بندی و پیش‌بینی می‌کند که این مقادیر می‌تواند پیوسته یا گسسته باشد. دقت این روش داده‌کاوی، نسبت به روش‌های موجود، بالاتر است. در مطالعه Bagheri و همکاران (۶) به مطالعه تشخیص بیماران نارسایی قلبی با استفاده از داده‌کاوی، در دو روش درخت تصمیم و رگرسیون انجام و نتایج باهم مقایسه گردید که این تحقیق با استفاده از داده‌های مربوط به بیماران نارسائی قلبی در انستیتوی قلب و عروق فیصل‌آباد و بیمارستان متفقی فیصل‌آباد، جهت شناسایی عوامل مؤثر در وقوع مرگ بیماران عملیات پیاده سازی انجام شد. نتایج حاصل از این تحقیق نشان داد که میزان دقت تشخیص در این تحقیق با روش درخت تصمیم برابر با ۹۵/۶۵٪ و با تکنیک رگرسیون دارای میزان دقت ۹۱/۲۸٪ است. در مطالعه Bhatt و همکاران (۲۱) از ابزار داده کاوی Weka به منظور پیش‌بینی بیماری قلبی با استفاده از دو تکنیک طبقه‌بندی استفاده کردند J48 که در مجموعه داده مجارستانی استفاده شد و Naïve Bayes که در پایگاه داده اکوکاردیوگرام به کار رفت. برای ارزیابی مدل‌های طبقه‌بندی از ماتریس سردرگمی و معیارهای عملکرد استفاده شد. مجموعه داده اول دارای ۱۴ ویژگی با متغیر هدف ۵ مقدار و مجموعه داده دوم دارای ۱۳۲ نمونه و ۱۲ ویژگی بود. دو آزمون با استفاده از الگوریتم‌های J48 و Naive Bayes با تمام ویژگی‌ها و با استفاده از گروهی از ویژگی‌های خاص برای مقایسه نتایج برای انتخاب ویژگی انجام شد. با استفاده از اولین مجموعه داده، دقت طبقه‌بندی ۸۲/۳٪ با استفاده از تمام ویژگی‌هایی که از دقت ۶۵/۶۴٪ با ویژگی‌های انتخاب شده بهتر است، به دست آمد. با استفاده از مجموعه داده دوم، نتایج نشان می‌دهد که دقت طبقه‌بندی ۹۸/۶۴ درصد با استفاده از تمام ویژگی‌ها و دقت ۹۳/۲۴ درصد با ویژگی‌های انتخاب شده به دست آمده است.

نتیجه‌گیری

استفاده از روش داده‌کاوی در غربالگری افراد مستعد بیماری‌های ایسکمیک قلب و عروق کارایی مناسب دارد و با کمک آن می‌توان این افراد را سریع‌تر و با هزینه کمتر نسبت به غربالگری سنتی شناسایی و درمان کرد. استفاده از داده‌کاوی نسبت به روش سنتی اهمیت و دقت بالاتری داشته و به دلیل اهمیت زمان در پیش‌بینی بیماری قلبی، داده کاوی به دلیل استفاده از حجم زیادی از داده‌های موجود در مدت زمان کوتاه‌تر، کارآمدتر است در نتیجه با پیش‌بینی زودهنگام امکان درمان زودهنگام بیماری را فراهم کرده و موجب کاهش مرگ و میر ناشی از این بیماری شده و هم‌چنین بار هزینه‌های تشخیص و درمان را کاهش می‌دهد.

پیشنهادات کاربردی

با استفاده از پیش‌بینی‌های مربوط به مدل‌های این پژوهش می‌توان زودتر و بهتر به عوامل موثر در بهبود درمان این بیماران توسط مراکز بهداشتی - درمانی رسید. مصرف سیگار یکی از

[DOI: 10.18502/ssu.v31i7.13693]

مهم‌ترین و تأثیرگذارترین عوامل در پیش‌بینی بیماری‌های ایسکمیک قلب در تمامی مدل‌ها بود، که با برنامه‌ریزی پویش‌های ترک سیگار می‌توان این عامل خطر را در زندگی بسیاری از مردم جامعه کاهش داده و زمینه ارتقاء سلامت را فراهم نمود.

پیشنهادات پژوهشی

در این پژوهش از درخت تصادفی و نزدیک‌ترین همسایگی و نایوبیز برای مدل‌سازی و پیش‌بینی عوامل مؤثر بر بیماری‌های قلبی، استفاده شد، پیشنهاد می‌شود در پژوهش‌های آینده بر شبکه‌های عصبی و الگوریتم‌های دیگر تمرکز شده و هم‌چنین از پارامترهای دیگر استفاده کرد.

سپاس‌گزاری

این مقاله بخشی از پایان‌نامه کارشناسی ارشد رشته مهندسی کامپیوتر گرایش نرم‌افزار دانشگاه پیام نور تهران می‌باشد که بدون حمایت مالی انجام شده است. در پایان از تمامی شرکت‌کنندگان و مجریان طرح یاس که امکان انجام این تحقیق را فراهم نموده‌اند، تشکر می‌گردد.

حامی مالی: ندارد.

تعارض در منافع: وجود ندارد.

References:

1-Bahrambagi Z, Lotfi Kashani F, Vaziri S. ***Effectiveness of Mindfulness-Based Therapy on Chronic Stress and Disease Perception in Heart Patients.*** Medical Sciences 2023; 33(1): 70-9. [Persian]

2-Malekyian Fini E, Ahmadizad S. ***Effect of Resistance Exercise and Training and Principles of prescribing it for Cardiovascular Patients.*** J Shahid Sadoughi Univ Med Sci 2021; 29(8): 3955-75. [Persian]

3-Tougui I, Jilbab A, El Mhamdi J. ***Heart Disease Classification Using Data Mining Tools and Machine Learning Techniques.*** Health Technol 2020; 10: 1137-44.

4-Pavithra M, Sindhana AM, Subajanaki T, Mahalakshmi S. ***Effective Heart Disease Prediction Systems Using Data Mining Techniques.*** Annals of R.S.C.B 2021; 25(3): 6566-71.

5-Premsmith, J, Ketmaneechairat H. ***A Predictive Model for Heart Disease Detection Using Data Mining Techniques.*** Journal of Advances in Information Technology 2021; 12(1): 14-20.

6-Bagheri A, Kilini Mina. ***Diagnosis of Survival in Heart Failure Patients Using Data Mining, in Two Methods of Decision Tree and Regression and Comparing the Results of These Two Methods.*** The 4th International Conference on Information Technology, Computer and Telecommunication Engineering of Iran, Tehran, August 1400.

7-Mahmoodi MS. ***Heart Disease Prediction System Using Support Vector Machine.*** Journal of Health and Biomedical Informatics 2017; 4(1): 1-10. [Persian]

8-Yadav SK, Chouhan Y, Choubisa M. ***Predictive Hybrid Approach Method to Detect Heart Disease.*** Mathematical Statistician and Engineering Applications 2022; 71(1): 36-47.

9-Mirzaei M, Salehi-Abargouei A, Mirzaei M, Mohsenpour MA. ***Cohort Profile: The Yazd Health***

[DOI: 10.18502/ssu.v31i7.13693]

Study (Yahs): A Population-Based Study of Adults Aged 20-70 Years (Study Design and Baseline Population Data). Int J Epidemiol 2017; 47(3): 697-8h. [Persian]

10-Webb GI, Keogh E, Miikkulainen R. ***Naïve Bayes.*** Encyclopedia of machine learning 2010; 15: 713-14.

11-Pouriyeh S, Vahid S, Sannino G, De Pietro G, Arabnia H, Gutierrez J. ***A Comprehensive Investigation and Comparison of Machine Learning Techniques in the Domain of Heart Disease.*** IEEE Symposium on Computers and Communications (ISCC) 2017; 204-7.

12-Liu Y, Wang Y, Zhang J. ***New Machine Learning Algorithm: Random Forest.*** ICICA 2012; 246-52.

13-Charbuty B, Abdulazeez A. ***Classification Based on Decision Tree Algorithm for Machine Learning.*** JASTT 2021; 2(01): 20-8.

14-Rubini PE, Subasini CA, Katharine AV, Kumaresan V, Kumar SG, Nithya TM. ***A Cardiovascular disease Prediction Using Machine Learning Algorithms.*** Annals of the Romanian Society for Cell Biology 2021; 25(2): 904-12.

15-Ali MM, Paul BK, Ahmed K, Bui FM, Quinn JMW, Moni MA. ***Heart Disease Prediction Using Supervised Machine Learning Algorithms: Performance Analysis and Comparison.*** Comput Biol Med 2021; 136: 104672.

16-Ishaq A, Sadiq S, Umer M, Ullah S, Mirjalili S, Rupapara V, et al. ***Improving the Prediction of Heart Failure Patients' Survival Using SMOTE and Effective Data Mining Techniques.*** IEEE Access 2021; 9: 39707-16.

17-Tougui I, Jilbab A, El Mhamdi J. ***Heart Disease Classification Using Data Mining Tools and Machine Learning Techniques.*** Health and Technology 2020; 10(5): 1137-44.

18-Premsmith J,Ketmaneechairat H. ***A Predictive Model for Heart disease Detection Using Data Mining Techniques.*** Journal of Advances in Information Technology 2021; 12(1): 14-20.

19-Kavitha M, Gnaneswar G, Dinesh R, Sai YR, Suraj RS. ***Heart Disease Prediction Using Hybrid Machine Learning Model.*** In 2021 6th international conference on inventive computation technologies (ICICT) 2021; 1329-33.

20-Kazemi M, Mehdizadeh M, Shiri A. ***Heart Disease Forecast Neural Network Data Mining Techniques.*** Journal of Ilam University of Medical Sciences 2017; 25(1): 20-32. [Persian]

21-Bhatt A, Dubey SK, Bhatt AK, Joshi M. ***Data mining approach to predict and analyze the cardiovascular disease***. InProceedings of the 5th International Conference on Frontiers in Intelligent Computing: Theory and Applications 2017, 1:117-126.

[DOI: 10.18502/ssu.v31i7.13693]

Journal of Shahid Sadoughi University of Medical Sciences
Vol 31| Nov 7 | Oct 2023

Received: 22 Oct 2022
Accepted: 15 Mar 2023

Comparison of Data Mining Algorithms in Prediction of Coronary Artery Diseases Using Yazd Health Study (YaHS) Data

Azam Barzegari[1], Seyede Fatemah Noorani[*2], Masoud Mirzaei[3]

Original Article

Introduction: Cardiovascular diseases, including ischemic heart disease (IHD), are one of the main cause of mortality and morbidity worldwide and are currently one of the top ten causes of death. Ischemic heart disease is a type of heart disease that is caused by narrowing of arteries feeding the heart itself. The present study aimed to use data mining algorithms in screening and early prediction of IHD according to the patient's characteristics and risk factors.

Methods: In this research, data of the first phase of Yazd Health Study (YaHS), focusing on 21 characteristics of 10,000 participants aged 20-70 years such as age, type of chest pain, blood sugar level, body mass index, employment status, etc. which have been collected since 2013 were analyzed.

Results: Data analysis was conducted using Random Forest and Naive Bayes algorithms which showed 74.51% accuracy in predicting IHD.

Conclusion: The study findings revealed that via applying Random Forest and Naive Bayes algorithms, ischemic heart disease can be predicted with high accuracy. Moreover, early screening and timely treatment in the early stages of disease may reduce mortality and morbidity.

Keywords: Data mining, Screening, Coronary heart Disease, Naive Bayes, Random Forest, YaHS.

Citation: Barzegari A, Noorani N, Mirzaei M**. Comparison of Data Mining Algorithms in Prediction of Coronary Artery Diseases Using Yazd Health Study Data (YaHs)**. J Shahid Sadoughi Uni Med Sci 2023; 31(7): 6824-35.

[DOI: 10.18502/ssu.v31i7.13693]

[1]Shahid Sadoughi University of Medical Sciences, Yazd, Iran.

[2]Faculty of Engineering, Department of Computer Engineering and Information Technology, Payame Noor University, Tehran, Iran.

[3]Cardiovascular Research Center, Shahid Sadoughi University of Medical Sciences, Yazd, Iran.

***Corresponding author: Tel:** 09197724657**, email:** sf.noorani@pnu.ac.ir