مقاله پژوهشی

مجله انفورماتیک سلامت و زیست پزشکی
مرکز تحقیقات انفورماتیک پزشکی
دوره هشتم شماره دوم. ص ۲۰۷-۱۹۳، ۱۴۰۰

مدل‌سازی و پیش‌بینی احتمال ابتلاء به بیماری قلبی عروق کرونری با استفاده از الگوریتم‌های داده‌کاوی

پریا سعدی[۱]، معصومه زینال نژاد[۲]*، فرزاد موحدی سبحانی[۳]

• دریافت مقاله: ۱۴۰۰/۲/۱۹ • پذیرش مقاله: ۱۴۰۰/۶/۱

مقدمه: بیماری قلبی عروق کرونری یکی از شایع‌ترین علت‌های مرگ‌ومیر در بزرگ‌سالان است، درحالی‌که، با تشخیص سریع و دقیق، درمان به‌موقع و نجات بیمار تا حد زیادی امکان‌پذیر است. از این‌رو، هدف این پژوهش شناسایی فاکتورهای مؤثر در ابتلاء به این بیماری و ارائه مدلی داده‌محور جهت کمک به پزشکان در پیش‌بینی و تشخیص آن است.

روش: پژوهش حاضر از نوع تحقیق کاربردی-توسعه‌ای است که در آن ۲۰۳۸ رکورد گردآوری شده در مدت ۵ سال در بیمارستان قلب شهید رجایی تهران، طی عملیات پیش‌پردازش و آماده‌سازی، با استفاده از نمونه‌برداری تصادفی متوازن، به ۱۰۰۰ رکورد، ۵۰۰ بیمار و ۵۰۰ فرد سالم، کاهش یافت. مرور ادبیات تحقیق، مشاوره با پزشکان متخصص، و وزن‌دهی با استفاده از روش کای‌دو، منجر به تعیین ویژگی‌ها شد. مدل‌ها با استفاده از الگوریتم‌های ماشین بردار پشتیبان، شبکه عصبی و جنگل تصادفی در محیط نرم‌افزارهای رپیدماینر و پایتون ایجاد شدند.

نتایج: در میان ۳۵ متغیر شناسایی شده، مهم‌ترین ویژگی‌ها عبارت‌اند از بیماری دریچه‌های قلبی، درد قفسه سینه، کلسترول بد، اختلال حرکت دیواره‌ای قلب، تری‌گلیسیرید، سدیم، پتاسیم، فشارخون و وزن. معیار F، دقت، صحت، و بازخوانی، به ترتیب، برای الگوریتم جنگل تصادفی برابر با ۸۲/۱۱٪، ۸۱/۴۰٪، ۷۹/۰۷٪، ۸۵/۴۰٪ و نرخ خطای مدل ۱۸/۶٪ محاسبه شد.

نتیجه‌گیری: جنگل تصادفی با دقت قابل قبولی احتمال ابتلاء به بیماری قلبی عروق کرونری را پیش‌بینی نمود. در مقایسه مدل‌ها، به علت زیاد بودن تعداد گره‌های ورودی، خطای مدل شبکه عصبی، ۲۳/۶٪، نسبتاً بیشتر بود.

کلیدواژه‌ها: بیماری قلبی -عروق کرونری، پیش‌بینی، ماشین بردار پشتیبان، شبکه عصبی، جنگل تصادفی

•ارجاع: سعدی پریا، زینال نژاد معصومه، موحدی سبحانی فرزاد. مدل‌سازی و پیش‌بینی احتمال ابتلاء به بیماری قلبی عروق کرونری با استفاده از الگوریتم‌های داده‌کاوی. مجله انفورماتیک سلامت و زیست پزشکی ۱۴۰۰؛ ۸(۲): ۲۰۷-۱۹۳.

۱. کارشناسی ارشد مهندسی صنایع ، گروه مهندسی صنایع، دانشکده فنی و مهندسی، واحد علوم و تحقیقات، دانشگاه آزاد اسلامی، تهران، ایران
۲. دکترای مهندسی صنایع، استادیار، گروه مهندسی صنایع، دانشکده فنی و مهندسی، واحد تهران غرب، دانشگاه آزاد اسلامی، تهران، ایران
۳. دکترای مهندسی صنایع، استادیار، گروه مهندسی صنایع، دانشکده فنی و مهندسی، واحد علوم و تحقیقات، دانشگاه آزاد اسلامی، تهران، ایران
* **نویسنده مسئول:** معصومه زینال نژاد
آدرس: تهران، بلوار اشرفی اصفهانی، نرسیده به پل همت، انتهای خیابان شهید حسن آذری، دانشکده فنی و مهندسی دانشگاه آزاد اسلامی واحد تهران غرب، گروه مهندسی صنایع
• **شماره تماس:** ۴۴۲۲۰۶۷۷ – ۰۲۱ • **Email:** zeinalnezhad.m@wtiau.ac.ir

مقدمه

طی چند دهه اخیر، شیوع بیماری‌های قلبی و عروقی در جهان رو به افزایش بوده است، به‌طوری‌که اکنون آن را شایع‌ترین علت مرگ و ناتوانی افراد عنوان می‌کنند [۱]. بیماری قلبی- عروقی بر اساس اختلالات قلب در اثر اجتماع پلاکت‌ها در عروق ایجاد می‌شود و در نتیجه نارسایی قلبی، ایست قلبی، آریتمی بطنی و مرگ ناگهانی قلبی اتفاق می‌افتد [۲]. رایج‌ترین بیماری‌های قلبی، بیماری عروق کرونری قلب است که به آن تنگی عروق کرونر قلب یا بیماری قلبی نیز گفته می‌شود [۳]. در این بیماران، عروق خونی که در ابتدا نرم بوده و حالت کشسان داشته‌اند، باریک و سفت می‌شوند و ورود جریان خون به قلب مسدود می‌شود. به دنبال آن اکسیژن و مواد مغذی به قلب نمی‌رسد و نمی‌تواند به خوبی خون را پمپاژ کند. به این مشکل تصلب شرایین می‌گویند که موجب بسته شدن جریان خون می‌شود و اگر به‌موقع تشخیص و درمان نشود، می‌تواند باعث بروز حملات قلبی و مرگ شود [۴]. طبق گزارش سازمان بهداشت جهانی در ابتدای قرن بیستم، ۱۰ درصد از کل مرگ‌ها به علت بیماری‌های قلبی بوده‌اند. در پایان همین قرن، موارد مرگ ناشی از این بیماری‌ها به ۲۵ درصد افزایش یافته و پیش‌بینی می‌شود که با توجه به روند رو به رشد کنونی تا سال ۲۰۲۵ میلادی بیش از ۳۵ تا ۶۰ درصد از مرگ‌های جهان به دلیل بیماری‌های قلبی و عروقی روی دهند [۵]. بر اساس گزارش وزارت بهداشت کشور ایران، ۳۹/۹ درصد از علل مرگ ایرانیان به علت بیماری‌های قلبی است و عوامل خطرساز آن مانند افزایش کلسترول، پرفشاری خون، دیابت و مصرف سیگار در حال گسترش هستند [۴].

هزینه درمان بیماری‌های قلبی- عروقی نیز بالا است. به‌طور مثال، برای آنژیوگرافی که جهت تعیین میزان و مکان تنگی رگ‌های قلب مورد استفاده قرار می‌گیرد، بین ۳۰ تا ۴۰ میلیون ریال هزینه می‌شود. اگر پیش از آن تست ورزش و اسکن قلب نیز نیاز باشد، هزینه‌ای حدود ۵ تا ۶ میلیون ریال اضافه خواهد داشت. همچنین، اگر پس از آنژیوگرافی تشخیص گرفتگی برای بیمار داده شود و نیاز به جراحی و قرار دادن استنت باشد، بسته به بیمارستان و وسعت گرفتگی عروق بین ۱۰۰ تا ۳۰۰ میلیون ریال هزینه جراحی پرداخت می‌گردد. سالانه ۵/۷ درصد از افراد جامعه با پرداخت هزینه‌های درمان بیماری‌های سخت به زیر خط فقر می‌روند [۲]. به همین دلیل لازم است مدلی طراحی شود که بیماری قلبی را تشخیص دهد، زیرا تشخیص دقیق در مراحل اولیه و به دنبال آن درمان مناسب می‌تواند منجر به کاهش هزینه درمان و نجات تعداد قابل توجهی از بیماران شود [۶]. به دلیل اهمیت این بیماری، روش‌ها و ابزارهای مختلفی برای بررسی عملکرد قلب در علم پزشکی ابداع شده که به پزشکان توانایی تشخیص نوع بیماری و پیش‌بینی احتمال بروز آن در آینده را می‌دهد [۷]. از آن‌جمله، تحلیل داده‌های بیماران و استفاده از الگوریتم‌های مختلف داده‌کاوی در پیش‌بینی و تشخیص امراض قلبی است که امروزه در علم انفورماتیک سلامت امری مرسوم و متداول گردیده است [۸]. به عنوان نمونه، Aggrawal و Pal [۹]، با استفاده از تکنیک‌های یادگیری ماشین همچون رگرسیون لجستیک، جنگل تصادفی، ماشین بردار پشتیبان، شبکه عصبی و شبکه بیزین مدلی ارائه دادند که قادر بود حوادث مرگ و میر ناشی از بیماری‌های قلبی را پیش‌بینی کند. ارزیابی این الگوریتم‌ها نشان داد جنگل تصادفی با دقت ۸۳/۱۷٪ بالاترین دقت را دارا بوده است. در تحقیق Fitriyani و همکاران [۱۰]، علاوه‌بر الگوریتم‌های فوق‌الذکر، از درخت تصمیم نیز استفاده شد که در آن میان الگوریتم شبکه عصبی دقیق‌ترین مدل معرفی شد. به طور مشابه، Shahid و Singh [۱۱] نیز نشان دادند در صورتی‌که شبکه عصبی با روش بهینه‌سازی ازدحام ذرات PSO (Particle Swarm Optimization) ترکیب شود دقت پیش‌بینی بیماری قلبی عروق کرونری افزایش می‌یابد. شبکه عصبی در مقایسه با ماشین بردار پشتیبان هم توسط Ayatollahi و همکاران [۱۲] مورد استفاده واقع شد، ولی با توجه به مقادیر معیارهای ارزیابی از دقت نسبتاً کمتری، دقت برابر با ۷۹/۸٪ در مقابل ۸۷/۱٪، برخوردار بود. برخی از محققین نیز اذعان داشتند ترکیب الگوریتم‌های بهینه‌سازی فراابتکاری با الگوریتم‌های داده‌کاوی می‌تواند منجر به افزایش دقت پیش‌بینی‌ها شود [۱۵-۱۳].

Jan و همکاران [۱۶] و Nashif و همکاران [۱۷]، رگرسیون لجستیک، جنگل تصادفی، ماشین بردار پشتیبان، شبکه عصبی و شبکه بیزین را برای تشخیص بیماری‌های قلبی به‌کار بردند و به نتایج قابل قبولی دست یافتند. موسوی و همکاران [۱۸]، با استفاده از شبکه عصبی، درخت تصمیم‌گیری و شبکه بیزین به افتراق پنج نوع بیماری قلبی پرداختند. در این تحقیق داده‌های گردآوری شده از بیمارستان فوق‌تخصصی قائم کرج در نرم‌افزار رپیدماینر (Rapid Miner) تحلیل شدند و نتایج نشان داد مدل شبکه عصبی دارای بیشترین دقت بوده است. به‌طور مشابه، محمودی [۱۹]، با تحلیل داده‌های ۱۷۲ نفر و در نظرگرفتن ۲۱ ویژگی به طراحی مدل پیش‌بینی بیماری قلبی-

عروقی با استفاده از ماشین بردار پشتیبان پرداخت. در این تحقیق، ترکیب سیستم فازی و طبقه‌بندی کننده ماشین بردار پشتیبان با استفاده از نرم‌افزار متلب پیاده‌سازی گردید. همچنین، تحقیقات حسن‌زاده و همکاران [۲۰] و صباغ‌گل و همکاران [۲۱] نشان داد الگوریتم‌های شبکه عصبی و درخت تصمیم از کارآیی قابل قبولی در تشخیص بیماری قلبی- عروق کرونری برخوردارند. در این زمینه، طهماسبی و همکاران [۲۲] نیز نشان دادند ترکیب شبکه عصبی با شبکه بیزین و k-نزدیک‌ترین همسایه (k Nearest Neighbors) KNN منجر به افزایش دقت پیش‌بینی تا حدود ۹۰ درصد می‌شود.

تحقیقات پیشین و مصاحبه با متخصصین قلب و عروق نشان داد فاکتورها و عوامل متنوع و زیادی در ابتلاء به بیماری قلبی عروق کرونری تأثیرگذار هستند، درحالی‌که در هر یک از پژوهش‌های پیشین به طور میانگین فقط حدود ۱۷ ویژگی بررسی شده است. نمونه‌ها بعضاً از پایگاه داده UCI استخراج شده‌اند که به روز نبوده و در بسیاری از موارد تعداد نمونه‌ها کمتر از ۳۰۰ رکورد بوده است؛ لذا این ضرورت احساس شد که در تحقیق حاضر حتی‌الامکان تمامی عوامل مؤثر در ابتلاء به بیماری قلبی عروق کرونری شناسایی شوند تا بتوان مدل جامعی برای پیش‌بینی آن ارائه داد. برای این منظور، با توجه به تعداد زیاد متغیرهای (ویژگی‌های) مسئله، لازم است تعداد قابل قبولی نمونه‌های تصادفی انتخاب شوند؛ نمونه‌هایی که بومی و واقعی بوده باشند تا مدل‌سازی نتایج معتبرتری دربر داشته باشد. در اکثر تحقیقات گذشته، ماشین بردار پشتیبان SVM(Support Vector Machine) به‌کار گرفته شده که این می‌تواند بیانگر مناسب بودن این الگوریتم با فاکتورهای (ویژگی‌های) بیماری‌های قلبی باشد. همچنین ملاحظه گردید الگوریتم جنگل تصادفی RF(Random Forest) در بسیاری از مواقع از بالاترین دقت در پیش‌بینی این بیماری برخوردار بوده است. ویتن و همکاران [۲۳] نیز اذعان داشتند شبکه‌های عصبی NN (Neural Network) در برخی از عملیات مانند پیش‌بینی در مقایسه با سایر روش‌ها دارای مزایای نسبی بوده و معمولاً در کارهای اجرایی ترجیح داده می‌شوند. از سوی دیگر، پژوهشگران پیشنهاد می‌کنند به منظور اطمینان از دقت پیش‌بینی‌ها ترکیبی از تکنیک‌های داده‌کاوی مختلف استفاده شود. بنابراین، با توجه به اهمیت تشخیص زودهنگام بیماری‌های قلبی و کارآیی قابل قبول تکنیک‌های داده‌کاوی در حوزه سلامت، در این پژوهش، ابتدا و عوامل مؤثر در احتمال ابتلاء به بیماری قلبی عروق کرونری بررسی و تعیین شد، سپس با توجه به تعداد و ماهیت متغیرهای مسئله، الگوریتم‌های شبکه عصبی، ماشین بردار پشتیبان و جنگل تصادفی به‌کار گرفته شد تا امکان ابتلاء به این بیماری با دقت قابل قبولی پیش‌بینی گردد.

روش

مراحل اجرای پژوهش حاضر در نمودار ۱ نمایش داده شده است.

نمودار ۱ مطابق با استاندارد (Cross-industry standard process) CRISP ترسیم شده است. بر این اساس، در ادامه مراحل اصلی روش اجرای تحقیق در سه زیربخش تشریح می‌شود.

مرحله ۱- شناخت سیستم و داده‌ها

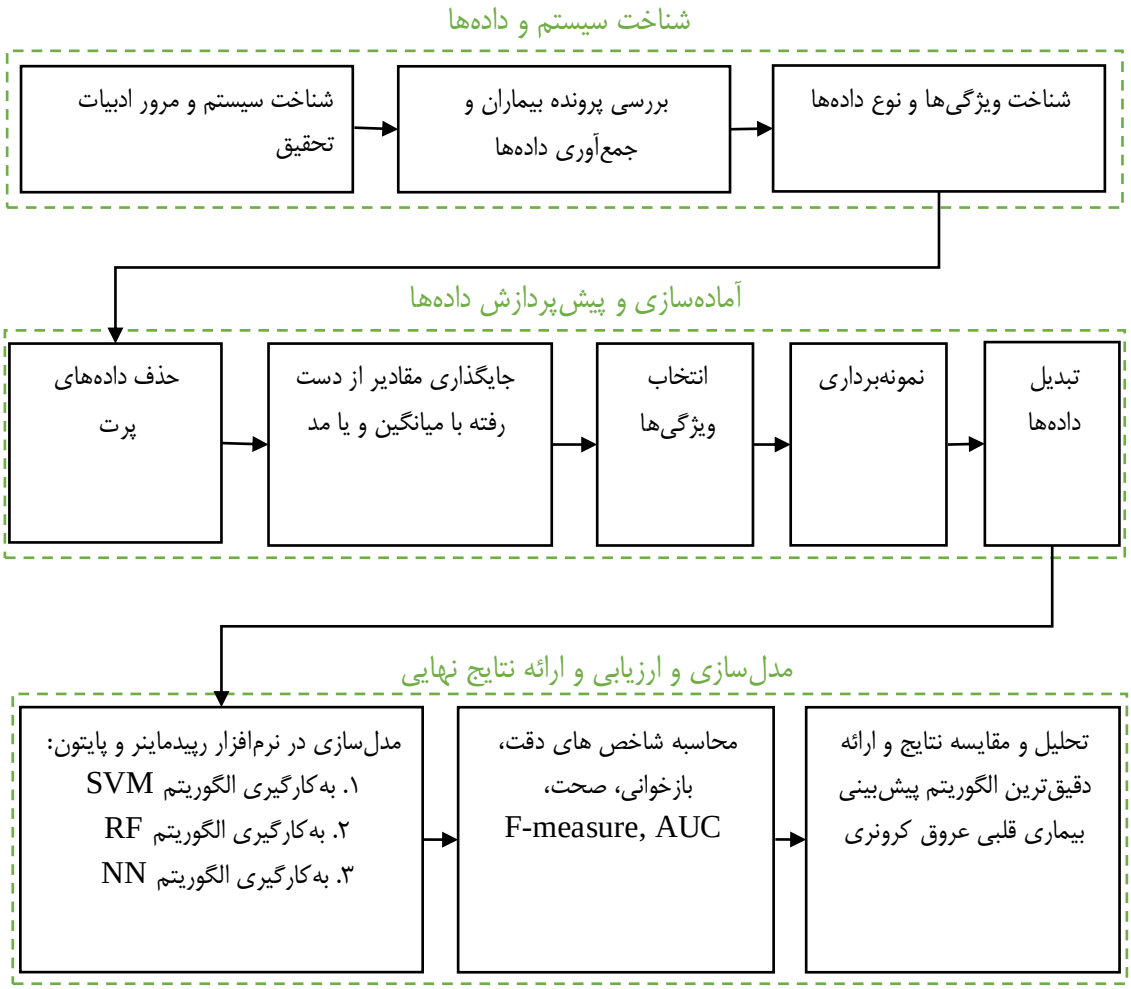
داده‌های موردنیاز، در قالب فایل اکسل، شامل اطلاعات ۲۰۳۸ نفر (۶۴ درصد بیمار مبتلا به بیماری قلبی عروق کرونری و ۳۶ درصد سالم)، از بیمارستان قلب شهید رجایی تهران طی ۵ سال، از ۱۳۹۵ تا ۱۳۹۹، درخواست و گردآوری شد. این مجموعه داده شامل ۵۹ ویژگی مرتبط با اطلاعات دموگرافیک بیماران، اطلاعات پرونده پزشکی ایشان، اطلاعات (Electro Cardio Graphic) ECG، اطلاعات معاینه و علائم بیمار، اطلاعات آزمایشگاه و اکو و داروهای مورد استفاده بود.

مرحله ۲- آماده‌سازی و پیش‌پردازش داده‌ها

در این تحقیق، با توجه به ماهیت داده‌ها و هدف تحقیق، در مرحله پیش‌پردازش و آماده‌سازی داده‌ها، عملیات پاک‌سازی داده‌ها، انتخاب ویژگی، نمونه‌برداری و تبدیل داده‌ها ضروری به‌نظر رسید. بدین‌صورت که ابتدا داده‌های پرت، کمتر از یک درصد داده‌ها، شناسایی و حذف شدند و داده‌های شامل مقادیر ازدست‌رفته (حدود ۸ درصد از داده‌ها) نیز با استفاده از میانگین یا مد هر ویژگی جای‌گذاری گردیدند. این کار در نرم‌افزار رپیدماینر به ترتیب با استفاده از عملگرهای Filter Examples و Replace Missing Values صورت گرفت.

بر اساس مرور ادبیات تحقیق و نظر پزشکان مشاور متخصص، ملاحظه گردید بسیاری از ۵۹ ویژگی موجود در مجموعه داده از نظر اطلاعاتی در نتیجه مدل‌سازی تأثیری ندارند، لذا توسط عملگر Select Attributes حذف شدند تا فضای الگوریتم بیهوده بزرگ نشود. بدین ترتیب، ۳۶ ویژگی، شامل ویژگی

کلاس، دارای ارزش اطلاعاتی برای تحلیل و مدل‌سازی باقی ماند که در جدول ۱ معرفی و تشریح شده‌اند.



نمودار ۱: مراحل اجرای تحقیق

جدول ۱: مشخصات متغیرها (فاکتورها یا ویژگی‌ها)

ردیف	نام ویژگی	سمبل	نوع داده	مقدار و دامنه	توضیح
۱	سن	Age	Numeric	۳۰ – ۸۶ سال	
۲	جنس	Sex	Binary	۰=مرد، ۱=زن	
۳	وزن	Weight	Numeric	۴۸ – ۱۲۰ کیلوگرم	
۴	شاخص توده بدنی	BMI	Numeric	۱۸ – ۴۱ کیلوگرم بر مترمربع	چربی بدن نسبت وزن به قد فرد است.
۵	دیابت قندی	DM	Binary	۰=ندارد، ۱=دارد	
۶	سیگاری فعلی	Current Smoker	Binary	۰=نیست، ۱=هست	
۷	سابقه خانوادگی	FH	Binary	۰=ندارد، ۱=دارد	
۸	چاقی	Obesity	Numeric	چاق :BMI>25	نسبت دور کمر به باسن زیاد است.
۹	نارسایی احتقانی قلب	CHF	Binary	۰=ندارد، ۱=دارد	سابقه نارسایی قلبی ریسک خطر است.
۱۰	اختلالات چربی خون	DLP	Binary	۰=ندارد، ۱=دارد	چربی رگ‌ها باعث تنگی عروق می‌شود.
۱۱	فشارخون	BP	Numeric	۹۰ – ۱۹۰ میلی‌متر جیوه	ضخیم شدن شریان‌ها و انسداد است.
۱۲	ضربان قلب	PR	Numeric	۵۰ – ۱۱۰ پالس	قلب در هر تپش صد سانتیمتر مکعب خون را در بدن پخش می‌کند.
۱۳	سوفل دیاستولیک	Diastolic Murmur	Binary	۰=ندارد، ۱=دارد.	صداهای غیرمعمول قلب در اثر گردش متلاطم خون است.
۱۴	تنگی نفس	Dyspnea	Binary	۰=ندارد، ۱=دارد.	هرگونه مشکلی روی تنفس است.
۱۵	درد قفسه سینه	Chest Pain	Numeric	۰ و ۱ و ۲ و ۳	انواع درد قفسه سینه: ۰=معمولی (مشکلات گوارشی)، ۱=غیرمعمولی (تنگی نفس و حالت تهوع)، ۲=درد بدون آنژین (بدون تنگی نفس با گلودرد)، ۳=درد هنگام (ورزش).
۱۶	کلاس عملکرد	Function Class	Numeric	۰ و ۱ و ۲ و ۳	وضعیت بیمار بعد از معاینه توسط پزشک (دامنه: ۰=معمولی، ۱=خفیف، ۲=متوسط، ۳=شدید).
۱۷	قند خون ناشتا	FBS	Numeric	۶۲ – ۴۰۰ میلی‌گرم در دسی‌لیتر	مقدار گلوکز موجود در خون انسان است.
۱۸	کراتینین	Cr	Numeric	۰/۵ – ۲/۲ میلی‌گرم در دسی‌لیتر	کراتین در عضلات، در مغز و ... است.
۱۹	تری‌گلیسیرید	TG	Numeric	۳۷ – ۱۰۵۰ میلی‌گرم در دسی‌لیتر	تری‌گلیسیرید رایج‌ترین چربی است.
۲۰	کلسترول بد	LDL	Numeric	۱۸ – ۲۳۲ میلی‌گرم در دسی‌لیتر	لیپوپروتئین با چگالی کم کلسترول بد باعث افزایش تشکیل پلاک می‌شود.
۲۱	کلسترول خوب	HDL	Numeric	۱۵ – ۱۱۱ میلی‌گرم در دسی‌لیتر	لیپوپروتئین با چگالی بالا کلسترول اضافی خون در کبد تجزیه می‌شود.
۲۲	نیتروژن اوره خون	BUN	Numeric	۶ – ۵۲ میلی‌گرم در دسی‌لیتر	اختلال کلیه‌ها، مواد زائد به خون است.
۲۳	سرعت رسوب گلبول قرمز در لوله آزمایش	ESR	Numeric	۱ – ۹۰ میلی‌متر در ساعت	سرعت رسوب گلبول‌های قرمز در لوله آزمایش است.
۲۴	هموگلوبین	HB	Numeric	۸,۹ – ۱۷,۶ میلی‌گرم در دسی‌لیتر	پروتئین حمل اکسیژن در خون است.
۲۵	لخته خون	INR	Numeric	۱ – ۱۰ میلی‌گرم در دسی‌لیتر	تعیین میزان لخته‌شدن خون است.
۲۶	پتاسیم	K	Numeric	۳ – ۶/۶ میلی مول در هر لیتر	عملکرد صحیح ماهیچه قلب است.

۲۷	سدیم	Na	Numeric	۱۲۸ – ۱۵۶ میلی مول در هر لیتر	اثرات بیشتر بر فشارخون است.
۲۸	تعداد گلبول سفید	WBC	Numeric	۳۷۰۰ – ۱۸۰۰۰ در هر میکرولیتر	میزان پایین ایمنی بدن کاهش می‌دهد.
۲۹	تعداد گلبول‌های قرمز	RBC	Numeric	۲۰۰۰ – ۱۵۰۰۰ در هر میکرولیتر	تعداد سلول‌ها در حجم خون است.
۳۰	لنفوسیت‌ها	Lymph	Numeric	۷ – ۶۰ در هر میکرولیتر	دفاع بدن در برابر باکتری‌ها است.
۳۱	نوتروفیل	NEUT	Numeric	۳۲ – ۸۹ در هر میکرولیتر	باکتری‌ها را از بین ببرند.
۳۲	پلاکت	PLT	Numeric	۲۵ – ۷۴۲ در هر میکرولیتر	سلول‌های دیسکی لخته خونی است.
۳۳	کسر تخلیه	EF	Numeric	۱۵ – ۶۰ سی سی	خون خارج شده از بطن‌ها است.
۳۴	اختلال حرکت دیواره‌ای قلب	RWMA	Numeric	۰ و ۱ و ۲ و ۳	(۰=معمولی، ۱=خفیف، ۲=متوسط، ۳=شدید).
۳۵	بیماری دریچه‌های قلبی	VHD	Numeric	۰ و ۱ و ۲ و ۳	(۰=معمولی، ۱=خفیف، ۲=متوسط، ۳=شدید).
۳۶	ویژگی کلاس	Cath	Binary	۱=مبتلا به بیماری عروق کرونری، ۰=سالم	در این‌جا دو کلاس با مقدار Cad و Normal، به ترتیب به افراد بیمار و افراد سالم تخصیص می‌یابند.

در ادامه، جهت اطمینان از تأثیرگذار بودن متغیرها و تعیین میزان اهمیت آن‌ها، ۳۵ ویژگی انتخاب شده با استفاده از روش آماره کای‌دو، عملگر Weight by Chi Squared Statistic در نرم‌افزار رپیدماینر، رتبه‌بندی شدند. پیش از این، توسط عملگر Discretize داده‌های اسمی به داده‌های عددی تبدیل شدند و نقش ویژگی سی و ششم (Cath) توسط عملگر Set Role به ویژگی برچسب (ویژگی دسته، کلاس یا label) تغییر نمود. در واقع، در پژوهش حاضر، متغیر هدف ویژگی کلاس Cad است که ابتلاء یا عدم ابتلاء به بیماری قلبی عروق کرونری را در فرد نشان می‌دهد.

در این‌جا، با توجه به عدم توازن دسته‌های موجود در مجموعه داده، از نمونه‌برداری تصادفی متوازن استفاده شد. با استفاده از عملگر Sample در نرم‌افزار، تعداد داده‌ها به طور تصادفی به ۱۰۰۰ داده کاهش یافت به‌طوری‌که تعداد داده‌های موجود در هر کلاس متوازن باشد. بدین‌ترتیب، جامعه آماری تحقیق را اطلاعات ۵۰۰ بیمار مبتلا به بیماری عروق کرونری و ۵۰۰ فرد سالم تشکیل می‌دهد.

در انتهای مرحله آماده‌سازی و پیش‌پردازش داده‌ها، به روش نرمال‌سازی z-transformation، با استفاده از عملگر Normalize در نرم‌افزار رپیدماینر، ویژگی‌های غیرنرمال به نرمال تبدیل شدند.

مرحله ۳- مدل‌سازی و ارزیابی و ارائه نتایج نهایی

در این مرحله، یادگیری مدل با استفاده از الگوریتم‌های جنگل تصادفی، ماشین بردار پشتیبان و شبکه عصبی صورت گرفت. سپس، با محاسبه معیارهای ارزیابی، الگوریتم‌ها مقایسه شده و در نهایت، دانش کشف شده به شیوه قابل درک تبدیل و ارائه شد. برای این منظور، ماتریس درهم‌ریختگی شامل مثبت حقیقی (True Positive) TP، منفی حقیقی (True Negative) TN، مثبت کاذب (False Positive) FP و منفی کاذب (False Negative) FN برای هر سه الگوریتم محاسبه گردید. TP تعداد نمونه‌های بیمار است که مدل آن‌ها را به درستی بیمار تشخیص داده است. TN تعداد نمونه‌های سالمی است که مدل آن‌ها را به درستی سالم تشخیص داده است. FP تعداد نمونه‌های بیمار است که مدل آن‌ها را به غلط سالم تشخیص داده است. FN تعداد نمونه‌های سالمی است که مدل آن‌ها را به غلط بیمار تشخیص داده است.

بر اساس مقادیر ماتریس درهم‌ریختگی، چهار معیار ارزیابی کارآیی الگوریتم‌های دسته‌بندی، یعنی دقت (Accuracy)،

بازخوانی (Recall)، صحت (Precision) و معیار اف (F-measure) محاسبه شدند. به‌منظور مقایسه کارایی دسته‌بندها، نمودار ROC (Receiver Operating Characteristic) ترسیم گردید که در آن محور عمودی (True Positive Rate) TPR و محور افقی FPR (False Positive Rate) است و از فرمول‌های ۱ و ۲ به‌دست می‌آیند:

$$TPR = \frac{TP}{TP + FN} \quad (1)$$

$$FPR = \frac{FP}{TN + FP} \quad (2)$$

همچنین، سطح زیر نمودار ROC محاسبه گردید تا AUC (Area Under Curve) یعنی میزان کارایی دسته‌بند مشخص گردد.

نرخ خطای ER (Error Rate) مدل نیز با استفاده از فرمول ۳ به دست می‌آید:

$$\text{Error Rate} = \frac{FN + FP}{TP + FP + TN + FN} \quad (3)$$

نتایج

نتایج اجرای روش کای‌دو:

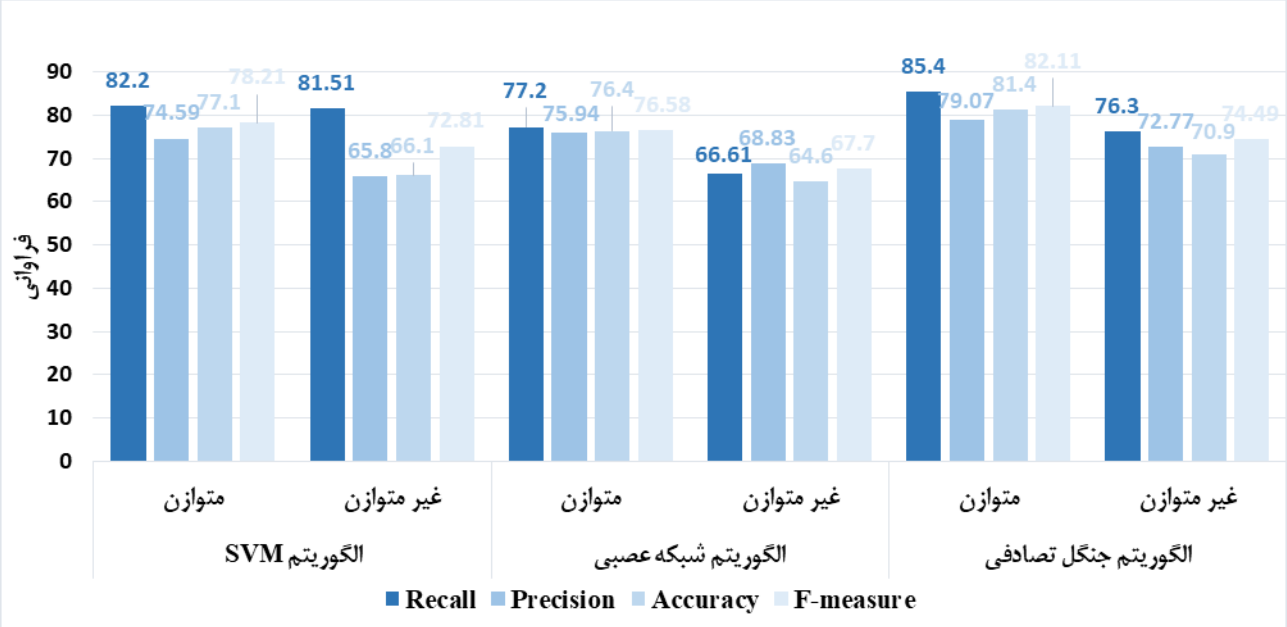
وزن و اهمیت محاسبه شده برای هر یک از ویژگی‌ها در جدول ۲ نشان داده شده است. مهم‌ترین فاکتورهای مؤثر بر بیماری قلبی عروق کرونری عبارت‌اند از بیماری دریچه‌های قلبی، درد قفسه سینه، کلسترول بد، اختلال حرکت دیواره‌ای قلب، تری‌گلیسیرید، سدیم، پتاسیم، فشارخون و وزن.

جدول ۲: وزن‌های محاسبه شده برای هر یک از متغیرها

ردیف	ویژگی‌ها	سمبل	وزن	ردیف	ویژگی‌ها	سمبل	وزن
۱	بیماری دریچه‌های قلبی	VHD	۴۰۰/۵۱۱	۱۹	سابقه خانوادگی	FH	۱۲/۷۰۱
۲	درد قفسه سینه	Chest pain	۵۵/۱۸۶	۲۰	تعداد گلبول‌های قرمز	RBC	۱۰/۲۹۷
۳	کلسترول بد	LDL	۴۵/۹۲۳	۲۱	لخته خون	INR	۹/۱۴۷
۴	اختلال حرکت دیواره‌ای قلب	RWMA	۴۲/۰۰۷	۲۲	کلاس عملکرد	Function Class	۸/۴۸۶
۵	تری‌گلیسیرید	TG	۳۲/۰۸۱	۲۳	نیتروژن اوره خون	BUN	۷/۰۷۱
۶	سدیم	Na	۳۱/۷۶۹	۲۴	قند خون ناشتا	FBS	۵/۶۳۳
۷	پتاسیم	K	۳۱/۴۷۴	۲۵	هموگلوبین	HB	۵/۵۳۹
۸	فشارخون	BP	۲۶/۶۴۱	۲۶	پلاکت	PLT	۵/۲۷۸
۹	وزن	Weight	۲۵/۸۲۵	۲۷	تعداد گلبول سفید	WBC	۴/۱۵۳
۱۰	سرعت رسوب گلبول قرمز در لوله آزمایش	ESR	۲۴/۹۱۴	۲۸	کراتینین	Cr	۳/۸۸۹
۱۱	ضربان قلب	PR	۲۳/۸۱۵	۲۹	دیابت قندی	DM	۳/۷۶۲
۱۲	شاخص توده بدنی	BMI	۲۱/۰۶۱	۳۰	جنس	Sex	۳/۷۴۴
۱۳	کلسترول خوب	HDL	۱۸/۱۶۷	۳۱	سیگاری فعلی	Current Smoker	۱/۲۸۲
۱۴	لنفوسیت‌ها نوعی گلبول سفید	Lymph	۱۷/۹۲۰	۳۲	تنگی نفس	Dyspnea	۰/۶۴۰
۱۵	سن	Age	۱۷/۱۶	۳۳	سوفل دیاستولیک	Diastolic Murmur	۰/۲۷۰
۱۶	کسر تخلیه	EF	۱۶/۹۹۱	۳۴	نارسایی احتقانی قلب	CHF	۰/۲۰۴
۱۷	اختلالات چربی خون	DLP	۱۳/۸۲۱	۳۵	چاقی	Obesity	۰/۰۴۲
۱۸	نوتروفیل	NEUT	۱۳/۱۷۴				

نتایج متوازن کردن داده‌ها

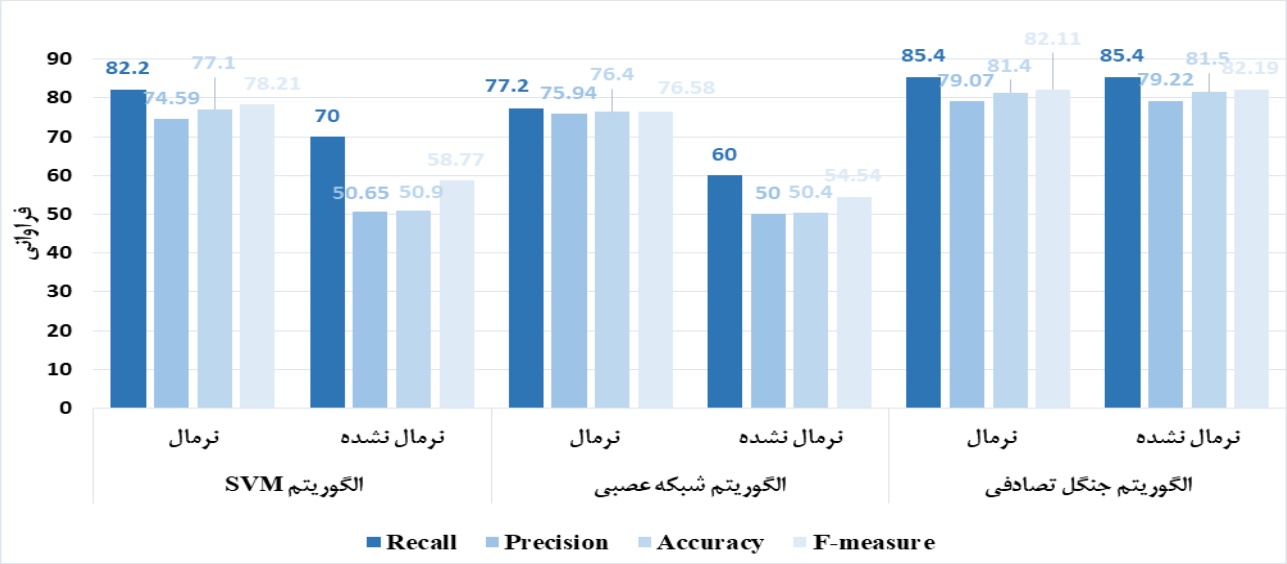
با توجه به نمودار ۲، ملاحظه گردید، با وجود بالاتر بودن دقت داده‌های غیرمتوازن، مقادیر معیارهای F-measure، Recall و Precision برای هر سه الگوریتم، به میزان قابل توجهی بالاتر است.



نمودار ۲: مقایسه داده‌های نرمال متوازن/غیرمتوازن

نتایج نرمال‌سازی داده‌ها

نمودار ۳ نشان می‌دهد نرمال کردن داده‌ها نتایج مثبتی در افزایش دقت الگوریتم‌ها داشته است.



نمودار ۳: مقایسه داده‌های متوازن نرمال شده/نرمال نشده

نتایج اجرای الگوریتم ماشین بردار پشتیبان

مقادیر درایه‌های ماتریس درهم‌ریختگی برای الگوریتم SVM در جدول ۳ نشان داده شده است. دسته مثبت معرف کلاس مربوط به افراد بیمار و دسته منفی معرف کلاس مربوط به افراد سالم است.

جدول ۳: ماتریس درهم‌ریختگی برای الگوریتم ماشین بردار پشتیبان

		مقادیر واقعی	
		دسته مثبت	دسته منفی
مقادیر پیش‌بینی شده	دسته مثبت	TP = ۴۱۱	FP = ۱۴۰
	دسته منفی	FN = ۸۹	TN = ۳۶۰

با در نظر گرفتن مقدار ۱ برای پارامتر جریمه (پارامتر C) و مقدار (۰،۱) برای پارامتر گاما، نتایج ارزیابی اجراهای الگوریتم SVM با توابع هسته مختلف در جدول ۴ نشان داده شده است.

جدول ٤: ارزیابی اجرای الگوریتم SVM با تابع‌های مختلف

نام تابع	Recall	Precision	Accuracy	F-measure
Linear	٪۷۴/۲۰	٪۷۱/۶۲	٪۷۲/۴۰	٪۷۲/۸۸
Poly	٪۷۷/۰۰	٪۷۴/۳۲	٪۷۵/۲۰	٪۷۵/۶۳
RBF	٪۸۲/۲۰	٪۷۴/۵۹	٪۷۷/۱۰	٪۷۸/۲۱
Sigmoid	٪۶۶/۴۰	٪۶۳/۹۷	٪۶۴/۵۰	٪۶۵/۱۶

نتایج اجرای الگوریتم شبکه عصبی

شبکه عصبی به‌کاررفته در تحقیق حاضر، شامل یک لایه پنهان با ۲۱ گره، لایه ورودی با ۳۶ گره و ۲ گره خروجی است. گره‌های ورودی بیانگر مهم‌ترین ویژگی‌ها و گره‌های خروجی بیانگر کلاس‌ها، یعنی بیماران قلبی عروق کرونری و افراد سالم، هستند. ماتریس درهم‌ریختگی مربوط به این الگوریتم در جدول ۵ ارائه شده است.

جدول ۵: ماتریس درهم‌ریختگی برای الگوریتم شبکه عصبی

		مقادیر واقعی	
		دسته مثبت	دسته منفی
مقادیر پیش‌بینی شده	دسته مثبت	TP = ۳۸۶	FP = ۱۱۲
	دسته منفی	FN = ۱۱۴	TN = ۳۷۸

به‌منظور ارزیابی الگوریتم شبکه عصبی نیز معیارهای فوق‌الذکر محاسبه شدند، به این ترتیب که مقادیر Recall، Precision، Accuracy و F-measure به ترتیب برابر با ۷۷/۲۰٪، ۷۵/۹۸٪، ۷۶/۴۰٪ و ۷۶/۵۸٪ شدند.

نتایج اجرای الگوریتم جنگل تصادفی

با درنظرگرفتن مقدار ۱۵۰ برای تعداد درخت‌ها و مقدار ۱۰ برای حداکثر عمق جنگل، مدل جنگل تصادفی ایجاد شد. ماتریس درهم‌ریختگی مربوط به این الگوریتم در جدول ۶ ارائه شده است.

جدول ٦: ماتریس درهم‌یختگی برای الگوریتم جنگل تصادفی

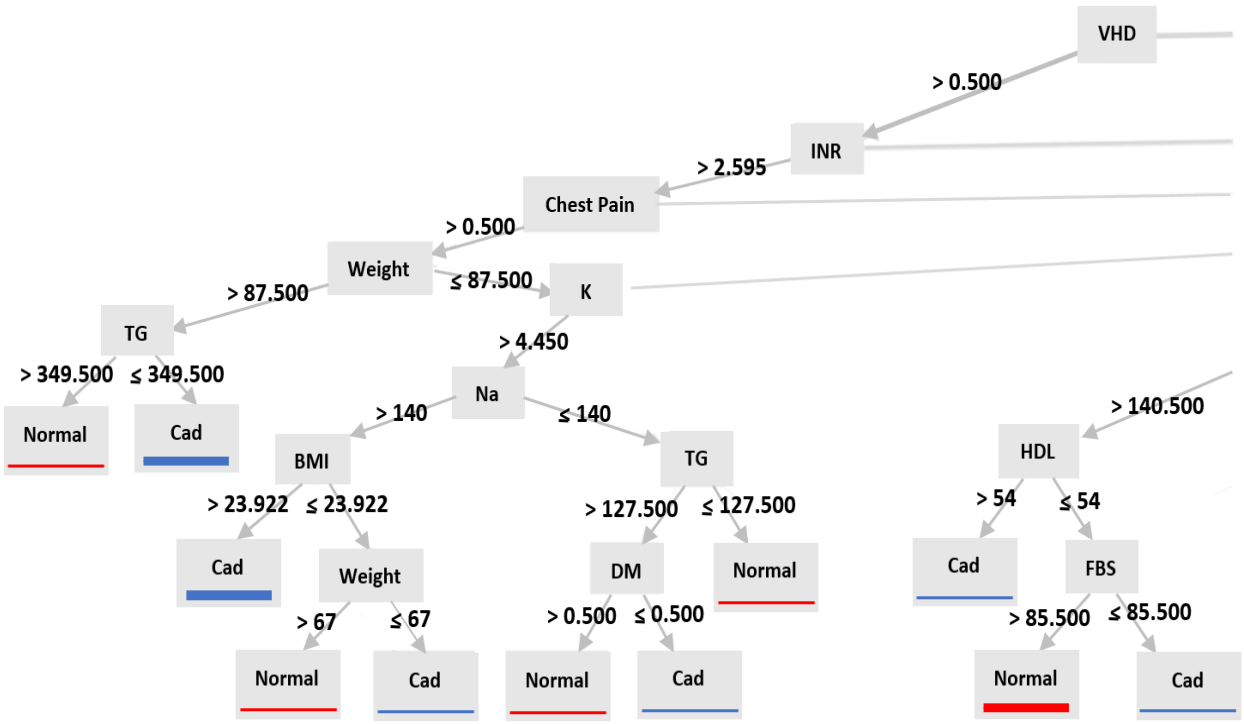
		مقادیر واقعی	
		دسته مثبت	دسته منفی
مقادیر پیش‌بینی شده	دسته مثبت	TP = ۴۲۷	FP = ۱۱۳
	دسته منفی	FN = ۷۳	TN = ۳۸۷

نتایج ارزیابی این الگوریتم با معیارهای مختلف در جدول ۷ ارائه شده است؛ که نشان داد معیار Gini Index نتایج بهتری به دنبال داشته است.

جدول ۷: ارزیابی اجرای الگوریتم جنگل تصادفی با معیارهای مختلف

نام معیار	Recall	Precision	Accuracy	F-measure
Gain Ratio	٪۸۲/۲۰	٪۷۳/۷۹	٪۷۶/۵۰	٪۷۷/۷۶
Information Gain	٪۸۲/۰۰	٪۷۸/۵۴	٪۷۹/۸۰	٪۸۰/۲۳
Gini Index	٪۸۵/۴۰	٪۷۹/۰۷	٪۸۱/۴۰	٪۸۲/۱۱

همچنین، قسمتی از جنگل تصادفی ایجاد شده در نمودار ۴ نمایش داده شده است.



نمودار ۴: قسمتی از جنگل به دست آمده از الگوریتم RF

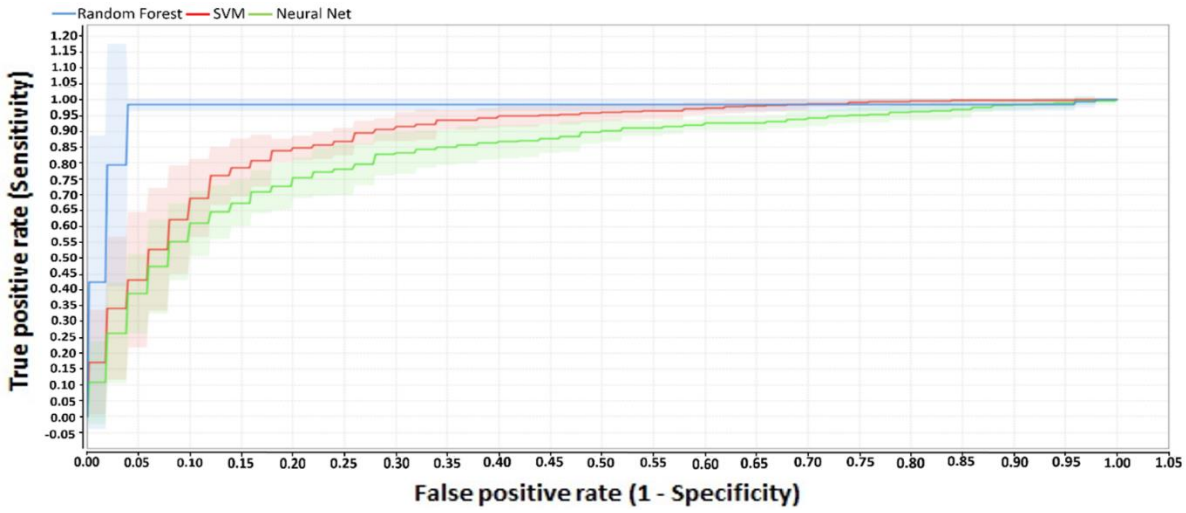
بحث و نتیجه‌گیری

در این پژوهش به منظور لحاظ نمودن تمامی متغیرهای تأثیرگذار در ابتلا به بیماری قلبی عروق کرونری و دستیابی به دقت قابل قبولی در پیش‌بینی این بیماری، الگوریتم‌های ماشین بردار پشتیبان با چهار نوع تابع هسته مختلف، جنگل تصادفی با سه معیار مختلف و شبکه عصبی استفاده شدند. کارآیی و درستی این الگوریتم‌ها با محاسبه ۱۰ معیار و شاخص در جدول ۸ با یکدیگر مقایسه شد. نتایج نشان شد الگوریتم‌های جنگل تصادفی و ماشین بردار پشتیبان کارایی بیشتری نسبت به شبکه عصبی دارند. دلیل نسبتاً پایین بودن مقادیر معیارها برای شبکه عصبی لحاظ نمودن ۳۵ ویژگی به عنوان گره‌های ورودی شبکه است که این تعداد به مراتب بیشتر از حد معمول می‌باشد. در واقع، اگر صرفاً هفت اولین ویژگی معرفی شده در جدول ۲، یعنی بیماری دریچه‌های قلبی، درد قفسه سینه، کلسترول بد، اختلال حرکت دیواره‌ای قلب، تری‌گلیسیرید، سدیم و پتاسیم که وزن بالای ۳۰ را کسب کرده‌اند، به عنوان متغیرهای ورودی مدل در نظر گرفته شود، دقت شبکه عصبی به مراتب بالاتر خواهد بود، اما هدف این پژوهش مدل‌سازی با حضور تمامی متغیرهای مؤثر در بیماری بوده است.

جدول ۸: نتایج ارزیابی الگوریتم‌ها

الگوریتم	Recall	Precision	Accuracy	F-measure	AUC	TP	TN	FP	FN	ER
SVM	٪۸۲/۲۰	٪۷۴/۵۹	٪۷۷/۱۰	٪۷۸/۲۱	٪۸۲/۶۰	۴۱۱	۳۶۰	۱۴۰	۸۹	٪۲۲/۹
NN	٪۷۷/۲۰	٪۷۵/۹۸	٪۷۶/۴۰	٪۷۶/۵۸	٪۸۱/۷۰	۳۸۶	۳۷۸	۱۲۲	۱۱۴	٪۲۳/۶
RF	٪۸۵/۴۰	٪۷۹/۰۷	٪۸۱/۴۰	٪۸۲/۱۱	٪۸۷/۰۰	۴۲۷	۳۸۷	۱۱۳	۷۳	٪۱۸/۶

منحنی‌های ROC مربوط به الگوریتم‌ها در نمودار ۵ ترسیم شده‌اند. منحنی ROC برای الگوریتم جنگل تصادفی بالاتر از سایرین ترسیم شده و نزدیک‌ترین منحنی به نقطه (۰،۱) است. این بدان معناست که در مجموع، الگوریتم جنگل تصادفی کارآترین مدل برای پیش‌بینی بیماری قلبی عروق کرونری است.



نمودار ۵: نمودارهای ROC برای مقایسه کارایی الگوریتم‌ها

به‌طور خلاصه، با توجه به آن‌که مقادیر شاخص‌های مرتبط با دقت و صحت و کارآیی مدل جنگل تصادفی حدوداً هشتاد درصد و بالاتر به‌دست آمده‌اند و نرخ خطای مدل کمتر از ۱۷ درصد است، می‌توان نتیجه گرفت با وجود لحاظ کردن تمامی متغیرهای مؤثر شناسایی شده در مدل، این الگوریتم قادر است با دقت قابل قبولی احتمال ابتلای افراد به بیماری قلبی عروق کرونری را پیش‌بینی نماید. در این‌جا، با توجه به مدل جنگل تصادفی نمایش داده شده در نمودار ۴، به چند نمونه از قوانین کشف شده توسط این الگوریتم که مورد تأیید پزشکان متخصص نیز قرار گرفته‌اند، اشاره می‌شود:

۱) اگر فرد دارای بیماری دریچه قلبی، لخته خون بیشتر از ۲/۵۹۵ میلی‌گرم در دسی‌لیتر، درد قفسه سینه، وزن بیشتر از ۸۷/۵۰۰ کیلوگرم و تری‌گلیسیرید کمتر یا مساوی ۳۴۹/۵۰۰ میلی‌گرم در دسی‌لیتر باشد، آنگاه به بیماری قلبی عروق کرونری مبتلا می‌شود.

۲) اگر فرد دارای بیماری دریچه قلبی، لخته خون بیشتر از ۲/۵۹۵ میلی‌گرم در دسی‌لیتر، درد قفسه سینه، وزن کمتر یا مساوی ۸۷/۵۰۰ کیلوگرم، پتاسیم بیشتر از ۴/۴۵۰ میلی‌مول در لیتر، سدیم بیشتر از ۱۴۰ میلی‌مول در لیتر و میزان توده بدنی بیشتر از ۲۳/۹۲۲ کیلوگرم در متر مربع باشد، آنگاه به بیماری قلبی عروق کرونری مبتلا می‌شود.

۳) اگر فرد دارای بیماری دریچه قلبی، لخته خون بیشتر از ۲/۵۹۵ میلی‌گرم در دسی‌لیتر، درد قفسه سینه، وزن کمتر یا مساوی ۸۷/۵۰۰ کیلوگرم، پتاسیم بیشتر از ۴/۴۵۰ میلی‌مول در لیتر، سدیم کمتر یا مساوی با ۱۴۰ میلی‌مول در لیتر، تری‌گلیسیرید بیشتر از ۱۲۷/۵۰۰ میلی‌گرم در دسی‌لیتر باشد و دیابت قندی نداشته باشد، آنگاه به بیماری قلبی عروق کرونری مبتلا می‌شود.

۴) اگر فرد دارای بیماری دریچه قلبی، لخته خون بیشتر از ۲/۵۹۵ میلی‌گرم در دسی‌لیتر، درد قفسه سینه، وزن کمتر یا مساوی ۸۷/۵۰۰ کیلوگرم، پتاسیم کمتر یا مساوی ۴/۴۵۰ میلی‌مول در لیتر، اختلالات چربی خون، سدیم بیشتر از ۱۴۰/۵۰۰ میلی‌مول در لیتر باشد و کلسترول خوب او بیشتر از ۵۴ میلی‌گرم در دسی‌لیتر باشد، آنگاه به بیماری قلبی عروق کرونری مبتلا می‌شود.

با مقایسه دقت به‌دست آمده برای الگوریتم جنگل تصادفی در پژوهش حاضر، ۸۱/۴۰٪، با مقدار مشابه در تحقیق Aggrawal و Pal [۹]، ۸۳/۱۷٪، نتایج تا حدود زیادی

مشابه‌اند و هر دو پژوهش کارآیی الگوریتم جنگل تصادفی را در پیش‌بینی نسبتاً دقیق احتمال ابتلاء به بیماری قلبی عروق کرونری تأیید می‌کنند. دقت محاسبه شده برای الگوریتم شبکه عصبی در پژوهش حاضر، ۷۶/۴۰٪، بالاتر از مقدار مشابه در تحقیق حسن‌زاده و همکاران [۲۰]، ۷۱/۷٪، ولی پایین‌تر از آن در تحقیقات Fitriyani و همکاران [۱۰]، ۸۵/۵۶٪ و موسوی و سپهری [۱۸]، ۸۰٪، است. دلیل این امر، با توجه به هدف مطالعه یعنی لحاظ نمودن تمامی ۳۵ ویژگی مؤثر در بیماری قلبی عروقی، تعداد بسیار زیاد متغیرهای مسئله است. شبکه عصبی معمولاً با تعداد گره ورودی تک رقمی نتایج بهتری از خود نشان می‌دهد و دو تحقیق فوق، فقط به ترتیب، ۱۳ و ۱۹ متغیر را لحاظ نموده‌اند. از سوی دیگر، در پژوهش حاضر ۱۰۰۰ نمونه بومی واقعی تحلیل شده‌اند، درحالی‌که دو تحقیق مذکور، به ترتیب، کمتر از ۳۰۰ و ۲۰۰ نمونه را مدل‌سازی نموده‌اند. از این‌رو، منطقی است فرض کنیم نتایج پژوهش حاضر واقعی‌تر و قابل اعتمادتر است. با استناد به این دلایل، کمتر بودن مقدار دقت الگوریتم SVM در پژوهش حاضر، ۷۸/۲۲٪، در مقایسه با تحقیق محمودی [۱۹]، ۸۵/۵٪، نیز قابل توجیه است. توجه به متوازن نمودن دسته‌ها و نرمال‌سازی متغیرها مسئله دیگری است که در تحقیقات مشابه کمتر به آن پرداخته شده است. معمولاً بسیاری از محققین برای متوازن ساختن مجموعه داده از روش Smote استفاده می‌کنند [۲۶-۲۴] که در آن با ایجاد داده‌های کاذب کلاس‌ها متوازن می‌شوند و این‌گونه بر دقت مدل به صورت غیرواقعی افزوده می‌شود، درحالی‌که پژوهش حاضر با استفاده از روش نمونه‌برداری متوازن به تحلیل داده‌های واقعی پرداخته است.

البته در پژوهش حاضر، همچون سایر تحقیقات داده‌محور، محدویت‌هایی از جمله احتمال اشتباه در ثبت داده‌ها توسط اپراتور، خطای اندازه‌گیری، وجود داده‌های پرت، مقادیر از دست رفته و نرمال نبودن داده‌ها وجود دارد که تلاش گردید با کمک تکنیک‌های آماده‌سازی و پاک‌سازی تا حد ممکن کاهش یابند. همچنین، با وجود آن‌که در بسیاری از تحقیقات مرور شده، از مجموعه داده‌های موجود در پایگاه داده UCI استفاده شده، در پژوهش حاضر تلاش گردید به داده‌های واقعی و بومی معتبر دست یافت که این امر با دشواری‌هایی از قبیل نگرانی از محرمانگی داده‌ها همراه بود و مستلزم دریافت معرفی‌نامه و کد اخلاق گردید. با توجه به دقت نسبتاً بالای کسب شده در تحقیقات آقائی‌نژاد و همکاران [۱۳]، Abdar و همکاران [۱۴]، نوشیار و همکاران [۱۵] و طهماسبی و همکاران [۲۲]، پیشنهاد می‌گردد در پژوهش‌های آتی، روش‌های ترکیبی با استفاده از الگوریتم ژنتیک، شبکه بیزین و KNN مدنظر قرار گیرند. مدل پیشنهادی حاضر قابلیت اجرا برای تمامی ۴۱ نوع بیماری قلبی، به شرط شناسایی دقیق متغیرهای مربوطه، را دارا است.

تشکر و قدردانی

این مقاله مستخرج از پایان‌نامه کارشناسی ارشد گروه مهندسی صنایع دانشگاه آزاد اسلامی واحد علوم تحقیقات تهران با عنوان «استفاده از روش‌های داده‌کاوی برای پیش‌بینی و تشخیص بیماری قلبی عروق کرونری» است.

تعارض منافع

نویسندگان مقاله اعلام می‌کنند که این پژوهش هیچ‌گونه تعارض منافعی ندارد.

References

1. Zegard A, Okafor O, Bono JD, Kalla M, Lencioni M, Arshall H, et al. Myocardial fibrosis as a predictor of sudden death in patients with coronary artery disease. J Am Coll Cardiol 2021;77(1):29-41. doi: 10.1016/j.jacc.2020.10.046.

2. Tabreer TH, Manal HJ, Ivan AH. Heart disease diagnosis system. Int J Curr Eng Technol 2017; 77(55): 2277-4106.

3. Piri Z, Dehghani Sufi M, Salimzadeh Z, Ashragh B, Alizadeh G. Heart Failure Management via Mobile Phones: A Systematic Review. Journal of Health and Biomedical Informatics 2017; 4(3):232-41. [In Persian]

4. Norouzkhani N, Sepehri MM. Designing and evaluating an education-based follow-up system for cardiac patients. Journal of Health and Biomedical Informatics 2020; 7(2): 113-23. [In Persian]

5. Kamali Yousef Abad M, Tara SM, Mouhebati M, Azizi A, Kiani B, Hasibian MR. Designing and Evaluation of the Local Cardiovascular Terms Coding System Based on Concept Mapping. Journal of Health and Biomedical Informatics 2015; 1(2):83-94. [In Persian]

6. Ghazisaeedi M, Shahmoradi L, Ranjbar A, Sahraei Z, Tahmasebi F. Designing a Mobile-Based Self-Care Application for Patients with Heart Failure. Journal of Health and Biomedical Informatics 2016; 3(3):195-204. [In Persian]

7. Srinivas K, Rani BK, Govrdhan A. Applications of data mining techniques in healthcare and prediction of heart attacks. International Journal on Computer Science and Engineering 2010;2(02):250-5.

8. Alizadehsani R, Khosravi A, Roshanzamir M, Abdar M, Sarrafzadegan N, Shafie D, et al. Coronary artery disease detection using artificial intelligence techniques: A survey of trends, geographical differences and diagnostic features 1991–2020. Comput Biol Med 2021;128:104095. doi: 10.1016/j.compbiomed.2020.10409

9. Aggrawal RI, Pal SA. Multi-Machine Learning Binary Classification, Feature Selection and Comparison Technique for Predicting Death Events Related to Heart Disease. International Journal of Pharmaceutical Research 2021;13(1): 428-39. doi:10.31838/ijpr/2021.13.01.080

10. Fitriyani NL, Syafrudin M, Alfian G, Rhee J. HDPM: an effective heart disease prediction model for a clinical decision support system. IEEE Access 2020;8:133034-50.
doi: 10.1109/ACCESS.2020.3010511

11. Shahid AF, Singh MP. A Novel Approach for Coronary Artery Disease Diagnosis using Hybrid Particle Swarm Optimization based Emotional Neural Network. Biocybernetics and Biomedical Engineering 2020; 40(4): 1568-85. https://doi.org/10.1016/j.bbe.2020.09.005

12. Ayatollahi H, Gholamhosseini L, Salehi M. Predicting coronary artery disease: a comparison between two data mining algorithms. BMC Public Health 2019;19(1):448. doi:10.1186/s12889-019-6721-5

13. Aghaienejad E, Teymourei-Yansary R, Riahi A. A hybrid model of heart anomalies detection by processing heart sounds Journal of Health and Biomedical Informatics 2019; 6(2): 101-10. [In Persian].

14. Abdar M, Ksiazek W, Acharya UA, Tan RS, Makarenkov V, Pławiak P. A new machine learning technique for an accurate diagnosis of coronary artery disease. Comput Methods Programs Biomed 2019;179:104992. doi: 10.1016/j.cmpb.2019.104992.

15. Nooshyar M, Momeni M, Gharravi S, Hourali F. Using gbc algorithm to optimize support vector machine parameters for predicting the relationship between cancer and cardiac infarction: a case study. Journal of Health and Biomedical Informatics 2018; 5(3): 361-72. [In Persian]

16. Jan M, Awan AA, Khalid MS, Nisar S. Ensemble approach for developing a smart heart disease prediction system using classification algorithms. Research Reports in Clinical Cardiology 2018; 9(1): 33-45. doi:10.2147/RRCC.S172035

17. Nashif S, Raihan MR, Islam MR, Imam MH. Heart disease detection by using machine learning algorithms and a real-time cardiovascular health monitoring system. World Journal of Engineering and Technology 2018; 6(4): 854-73. doi: 10.4236/wjet.2018.64057

18. Mousavi SR, Sepehri MM. Comparison of three decision-making models in differentiating informatics2018five types of heart disease: a case study in ghaem sub-specialty hospital. Journal of Health and Biomedical 2018;5(4): 457-68. [In Persian]

19. Mahmoodi MS. Designing a Heart Disease prediction System using Support Vector Machine. Journal of Health and Biomedical Informatics 2017; 4 (1):1-10. [In Persian]

20. Hassanzadeh M, Zabbah I, Layeghi K. Diagnosis of Coronary Heart Disease using Mixture of Experts Method. Journal of Health and Biomedical Informatics 2018; 5(2):274-85. [In Persian]

21. Sabbagh Gol H. Detection of coronary artery disease using C4.5 decision tree. Journal of Health and Biomedical Informatics 2017; 3(4): 287-99. [In Persian]

22. Tahmasbi H, Jalali M, Shakeri H. An Expert System for Heart Disease Diagnosis Based on Evidence Combination in Data Mining. Journal of Health and Biomedical Informatics 2017; 3(4):251-8. [In Persian]

23. Witten IH, Frank E, Hall MA, Pall C. Data Mining: Practical Machine Learning Tools and Techniques. 4th ed. St. Canada: Montreal; 2016.

24. Veloso R, Portela F, Santos MF, Silva A, Rua F, Abelha A, et al. A clustering approach for predicting readmissions in intensive medicine. Procedia Technology 2014;16(2):1307-16. doi:10.1016/j.protcy.2014.10.147

25. Cui S, Wang D, Wang Y, Yu PW, Jin Y. An improved support vector machine-based diabetic readmission prediction. Comput Methods Programs Biomed 2018;166:123-35. doi: 10.1016/j.cmpb.2018.10.012.

26. Srivastava S, Sharma L, Sharma V, Kumar A, Darbari H. Prediction of diabetes using aritificial neural network approach. Lecture Notes in Electrical Engineering 2019; 478(75): 679-87. https://doi.org/10.1007/978-981-13-1642-5_59

Original Article

Journal of Health and Biomedical Informatics
Medical Informatics Research Center
2021; 8(2):193-207

Modeling and Predicting the Risk of Coronary Artery Disease Using Data Mining Algorithms

Paria Saadi[1], Masoomeh Zeinalnezhad[2]*, Farzad Movahedi Sobhani[3]

• **Received: 9 May 2021** • **Accepted: 23 Aug 2021**

Introduction: Coronary artery disease (CAD) is one of the most common causes of death in adults while accurate and early diagnosis can lead to treatment and survival of patients to a great extent. Therefore, the objective of this study was to identify the effective factors leading to this disease and develop a data-driven model to assist physicians in predicting and diagnosing it.

Method: This is an applied research, considering 2038 medical records, collected from Shahid Rajaei Heart Hospital in Tehran, during 5 years. A data preprocessing was carried out and random balanced sampling reduced the dataset into 1000 records, with 500 CAD and 500 Normal. Literature review, consultation with specialist physicians, and weighting using the Chi-square method led to the determination of important features. Support Vector Machine, Neural Network and Random Forest algorithms were applied in RapidMiner and Python.

Results: Among the 35 identified variables, the most important features included VHD, Chest pain, LDL, RWMA, TG, Na, K, BP, and weight. The F-measure, precision, accuracy, and recall for random forest algorithm were calculated as 82.11%, 81.40%, 79.07%, and 85.40%, respectively, and the error rate was 18.6%.

Conclusion: Random Forest predicted the risk of CAD with a reasonable precision. In comparison, due to the large number of input nodes, the error rate of the Neural Network model was relatively higher (23.6%).

Keywords: Coronary Artery Disease, Prediction, Support Vector Machine, Neural Network, Random Forest

•**Citation:** Saadi P, Zeinalnezhad M, Movahedi Sobhani F. Modeling and Predicting the Risk of Coronary Artery Disease Using Data Mining Algorithms. Journal of Health and Biomedical Informatics 2021; 8(2): 193-207. [In Persian]

1. M.Sc. in Industrial Engineering, Industrial Engineering, Faculty of Engineering Dept., Science and Research Branch, Islamic Azad University, Tehran, Iran
2. Ph.D. in Industrial Engineering, Assistant Professor, Industrial Engineering Dept., Faculty of Engineering, West Tehran Branch, Islamic Azad University, Tehran, Iran
3. Ph.D. in Industrial Engineering, Assistant Professor, Industrial Engineering Dept., Faculty of Engineering, Science and Research Branch, Islamic Azad University, Tehran, Iran
***Corresponding Author:** Masoomeh Zeinalnezhad
Address: Industrial Engineering, Faculty of Engineering Dept., West Tehran Branch, Islamic Azad University, Shahid Hasan Azari St., Hemmat Bridge, Ashrafi Esfahani Hwy, Tehran, Iran
• **Tel:** 021-44220677 • **Email:** zeinalnezhad.m@wtiau.ac.ir