مجله دانشگاه علوم پزشکی قم
دوره دهم، شماره چهارم، تیر ۹۵
صفحه ۲۲ الی ۳۵

مقاله پژوهشی
(Original Article)

پیش‌بینی دقیق بیماری عروق کرونری با استفاده از الگوریتم‌های بیوانفورماتیک

هاجر شفیعی[۱]، منصور ابراهیمی[۲*]

چکیده

زمینه و هدف: بیماری قلبی - عروقی یکی از مهم‌ترین علل مرگ و میر در کشورهای پیشرفته و جهان سوم است. طبق اعلام سازمان بهداشت جهانی پیش‌بینی می‌شود مرگ و میر ناشی از بیماری‌های قلبی تا سال ۲۰۳۰ به ۲۳ میلیون نفر افزایش می‌یابد. در جدیدترین آمار وزیر بهداشت ایران، ۳۹/۳٪ کل مرگ و میرها ناشی از بیماری‌های قلبی - عروقی و ۱۹/۵٪ مربوط به سکته‌های قلبی گزارش شده است. این پژوهش با هدف پیش‌بینی بیماری عروق کرونر قلبی با استفاده از الگوریتم‌های داده‌کاوی انجام شد.

روش بررسی: در این مطالعه از الگوریتم‌های مختلف بیوانفورماتیک از جمله درخت تصمیم، شبکه‌های عصبی، ماشین‌بردار پشتیبان، خوشه‌بندی و ...، برای پیش‌بینی بیماری عروق کرونر قلب استفاده شد. در این مطالعه داده‌ها از چندین پایگاه معتبر (شامل ۱۴ داده) گرفته شدند.

یافته‌ها: در این تحقیق از تکنیک‌های داده‌کاوی، جهت تشخیص بیماری‌های مختلف از جمله بیماری عروق کرونری استفاده شد که مؤثر بود. همچنین برای اولین بار یک سیستم پیش‌بینی مبتنی بر ماشین‌بردار پشتیبان با بهترین دقت ممکن معرفی گردید.

نتیجه‌گیری: نتایج نشان داد بین ویژگی‌ها؛ متغیر اسکن‌تالیوم به‌عنوان مهم‌ترین ویژگی در تشخیص بیماری‌های قلبی می‌باشد، و طراحی مدل‌های پیش‌بینی ماشینی از جمله الگوریتم یادگیری بردار پشتیبان ماشین با دقت ۱۰۰٪ می‌تواند بین افراد بیمار و سالم تمایز قائل شود.

کلید واژه‌ها: بیماری‌های قلبی؛ بیماری‌های عروق کرونر؛ زیست‌شناسی محاسباتی؛ ماشین‌بردار پشتیبان.

[۱]دانشکده مهندسی، دانشگاه قم، قم، ایران.

[۲]گروه بیولوژی، دانشکده علوم پایه، دانشگاه قم، قم، ایران.

[*]نویسنده مسئول مکاتبات:
منصور ابراهیمی، گروه بیولوژی، دانشکده علوم پایه، دانشگاه قم، قم، ایران؛

آدرس پست الکترونیکی:
mansour@future.edu

تاریخ دریافت: ۹۴/۳/۲۶

تاریخ پذیرش: ۹۴/۶/۲۹

لطفاً به این مقاله به‌صورت زیر استناد نمایید:

Shafiee H, Ebrahimi M. Accurate prediction of coronary artery disease using bioinformatics algorithms. Qom Univ Med Sci J 2016;10(4):22-35. [Full Text in Persian]

مقدمه

بیماری قلبی - عروقی یکی از مهم‌ترین علل مرگ و میر در کشورهای پیشرفته و جهان سوم است. از نظر پاتولوژی، شایع‌ترین علت بیماری‌های قلبی - عروقی "آترواسکلروز یا بیماری تصلب شرائین" می‌باشد. این بیماری یک بیماری پیشرونده است که از دوران کودکی آغاز و تظاهرات بالینی آن به‌طور عمده در بزرگسالی و میانسالی بوده و تا هنگامی‌که بیش از ۶۰٪ قطر داخلی رگ را درگیر نکند، علائم بروز نمی‌کنند. در اثر فعال‌شدن پلاکت‌ها، موادی نظیر ترومبوکسان و سروتونین ترشح شده که باعث تحریک انقباض عروق و انتشار لخته می‌شوند. هنگامی‌که وسعت تجمع پلاکت‌ها و لخته تشکیل‌شده به‌حدی برسد که بتواند باعث انسداد عروق و جریان خون (به‌صورت نسبی یا کامل) گردد، یک واقعه حاد کرونری یا انفارکتوس میوکارد رخ می‌دهد. از نظر بالینی تغییرات آترواسکلروزی در قلب به‌عنوان بیماری شریان کرونری (CAD) شناخته می‌شود (۱).

این بیماری می‌تواند بر جنبه‌های جسمی، روانی و اجتماعی سلامتی و درک فرد از خوب بودن تأثیرگذار باشد. طبق شواهد موجود، این بیماری موجب بروز ناتوانی در بسیاری از ابعاد زندگی بیمار شده و با ایجاد اختلال در تحرک جسمی و خواب، بر عملکرد روزانه بیماران تأثیرگذار است و موجب کاهش انرژی و بروز واکنش‌های هیجانی اجتناب‌ناپذیر می‌شود (۲). بیماری‌های قلبی بعد از تصادفات، بالاترین نرخ مرگ و میر را به همراه دارند، از این‌رو تلاش می‌شود با استفاده از راهکارهای پیش‌بینی بیماری‌های قلبی، میزان تلفات ناشی از بیماری کاهش یابد. در صورت وجود متغیرهای اشاره‌شده در یک فرد می‌توان اقدامات لازم را جهت پیشگیری از بروز سکته قلبی و فجایع دیگر انجام داد. در تشخیص بیماری‌های قلبی، داده‌کاوی یک فرآیند کشف اطلاعات است که برای استخراج الگوها و قوانین نهفته از داده‌ها مورد استفاده قرار می‌گیرد. از ابزارها و تکنیک‌های داده‌کاوی برای تجزیه و تحلیل هوشمندانه اطلاعات نیز استفاده می‌شود (۳). تاکنون در تحقیقات متعدد انجام‌شده در این زمینه، از روش‌های متفاوتی برای تشخیص بیماری‌های قلبی استفاده شده است. مطالعات انجام‌شده (سال ۱۳۸۲) در بیمارستان امام خمینی جهت انجام آنژیوگرافی عروق کرونر (ABI) از ۱۰۰ بیمار، نشان داد ABI≤۰/۹ دارای ارزش تشخیص مثبت بالا جهت بیماری کرونری می‌باشد، اما تست حساسی برای تشخیص بیماری کرونری نیست؛ زیرا این سطح از ABI (ABI≤۰/۹) در تعداد زیادی از بیماران نمی‌تواند وجود بیماری کرونری را رد کند (۱). Rani در مطالعه خود به تجزیه و تحلیل یک مجموعه داده‌های قلبی و تکنیک شبکه‌های عصبی پرداخت و برای افزایش راندمان کار در مرحله آموزش از روش موازی استفاده کرد که به نتایج رضایت‌بخشی دست یافت (۴).

از الگوریتم بهینه‌سازی ذرات (PSO) با رویکرد افزایشی، برای استخراج قوانین و به رسمیت شناختن وجود یا عدم وجود بیماری عروق کرونر در یک بیمار استفاده می‌شود. در این روش به‌نظر می‌رسد برخلاف دیگر کارها، پیچیدگی افزایش پیدا کرده و پیچیدگی محاسباتی زیاد شده است، اما چون زمان انجام تست بسیار سریع است، شرایط عملکرد الگوریتم مناسب و قابل‌قبول می‌باشد (۵). Chau و همکاران پیشنهاد کردند از الگوریتم درخت تصمیم و الگوریتم کیسه‌ای یا متراکم شدن خودکار (Bagging) که برای ترکیب رده‌بندی‌های پیش‌بینی‌شده می‌باشد می‌توان برای پیش‌بینی بیماری‌های قلبی نیز استفاده کرد. این الگوریتم‌ها با ترکیب روش بیز ساده برای شناسایی بیماری‌های قلبی از افراد سالم، نتایج بسیار امیدوارکننده‌ای را نشان داده‌اند (۶).

در مطالعه حاضر، داده‌های موجود مربوط به ۲۷۰ فرد بیمار و سالم که ۷۶ متغیر مختلف در آنها اندازه‌گیری شده بود، جهت پیش‌بینی بیماری عروق کرونری در افراد سالم و بیمار با استفاده از الگوهای بیوانفورماتیک مورد استفاده قرار گرفت (۷)، تا بتوان روشی دقیق و سریع برای شناسایی افراد بیمار و سالم ارائه داد.

روش بررسی

داده‌های این مطالعه از پایگاه داده‌های بیماری‌های قلبی استخراج شد که در این پایگاه، داده‌هایی از چهار مجموعه داده متفاوت برای تشخیص بیماری‌های قلبی قرار دارند. داده‌های فوق از چهار منبع (بنیاد کلنیک کلیولند، انستیتوکاردیولوژی مجارستان، مرکز پژشکی لانگ‌بیچ کالیفرنیا و بیمارستان دانشگاه زوریخ سویس) جمع‌آوری شده بود (۸). در مجموع، در این پایگاه‌ها، داده‌ها با ۷۶ ویژگی یا متغیر مختلف اندازه‌گیری شدند که به‌طور مستقیم یا

غیرمستقیم با بیماری‌های قلبی مرتبط بودند. در پیش‌پردازش داده‌ها، جهت کم‌کردن تعداد متغیرها، کاهش زمان پردازش و اجرای مدل‌های داده‌کاوی و بیوانفورماتیک، ۱۴ ویژگی مهم انتخاب شد. متغیر هدف در این مطالعه، وجود یا عدم وجود بیماری قلبی بود که در مورد هرکدام از افراد مورد بررسی یکی از این دو حالت ثبت گردید (مقدار متغیر هدف یا num برابر با ۱، نشان‌دهنده وجود بیماری قلبی و صفر، نشان‌دهنده عدم وجود بیماری می‌باشد).

ویژگی‌های استفاده‌شده در این مقاله شامل: سن اشخاص، جنسیت، نوع درد قفسه سینه (که به ۴ دسته تقسیم شده بود)، فشار خون در حال استراحت، سطح کلسترول خون، میزان قند خون ناشتا، نتایج نوار قلب در حالت استراحت، حداکثر ضربان قلب، درد ناشی از ورزش، پایین‌افتادگی فاصله ST در نوار قلب ناشی از تمرین نسبت به مرحله استراحت، انحراف در شیب ST در زمان ورزش، تعداد عروق اصلی درگیر در فلوروسکوپی و نتایج اسکن‌تالیوم بود (جدول شماره ۱).

جدول شماره ۱: متغیرهای اصلی انتخاب‌شده، توصیف و نوع آنها جهت بررسی در این مطالعه

ویژگی‌ها	مقادیر به‌کار رفته برای هر ویژگی	ویژگی‌ها	مقادیر به‌کار رفته برای هر ویژگی
سن	عدد	حداکثر ضربان قلب	عدد
جنس	۱: زن، صفر: مرد	درد ناشی از ورزش	۱: بله ۰: خیر
نوع درد قفسه سینه	۱: آنژین معمول ۲: آنژین غیرمعمول ۳: درد غیرآنژینی ۴: بدون علامت	پایین‌افتادگی فاصله ST در نوار قلب ناشی از تمرین نسبت به مرحله استراحت	عدد
فشار خون در حال استراحت (میلی‌متر جیوه)	عدد	انحراف در شیب ST در زمان ورزش	۱: شیب به سمت بالا ۲: صاف ۳: شیب به سمت پایین
سطح کلسترول خون (میلی‌گرم بر دسی‌لیتر)	عدد	تعداد عروق اصلی درگیر در فلوروسکوپی	عدد
میزان قند (میلی‌گرم بر دسی‌لیتر)خون ناشتا در ۱۲۰	درست غلط	اسکن‌تالیوم	۳: نرمال ۶: نقص ثابت ۷: نقص برگشت پذیر
نتایج نوار قلب در حالت استراحت	موج اس-تی غیر نرمال	متغیرهدف برای تشخیص فرد بیمار از سالم	۰: سالم ۱: مریض

به کمک انتخاب ویژگی‌ها (Attribute Selection)، ویژگی‌های مورد نیاز از بین سایر ویژگی‌ها انتخاب، و اهمیت و سهم هر ویژگی با توجه به متغیر هدف به‌وسیله الگوریتم ویژگی وزنی (Attribute Weighting) مورد بررسی قرار گرفت که در نهایت، وزنی بین ۰ و ۱ به هر ویژگی اختصاص داده شد (۱۱-۹). این عمل، به‌ویژه برای مجموعه داده‌های بزرگ و پیچیده مناسب بوده و اجازه می‌دهد ویژگی‌های مهم مورد نیاز، به‌راحتی انتخاب شود. خوشه‌بندی به عمل تقسیم جمعیت ناهمگن در تعدادی از زیرمجموعه‌ها یا خوشه‌های همگن گفته می‌شود. تفاوت اساسی خوشه‌بندی و دسته‌بندی در این است که در دسته‌بندی براساس یک مدل، هرکدام از دسته‌ها از پیش تعیین شده یا از قبل وجود داشته است یا از طریق یافته‌های پیشین این دسته‌بندی‌ها وجود دارد، اما در خوشه‌بندی، هیچ دسته‌ای از قبل وجود نداشته و براساس تشابه این دسته‌بندی‌ها انجام می‌شود و دسته‌بندی‌ها بدین‌صورت است که درون هر دسته، بیشترین تشابه بین داده‌ها وجود دارد و بین هر دو دسته تشابه، حداقل می‌شود (۱۳،۱۲). داده‌ها ممکن است دارای ساختارهای پیچیده‌ای باشند که هیچ تکنیکی قادر به استخراج الگوی معنی‌داری از آنها نباشد، اما خوشه‌بندی به بهترین وجه‌ممکن این عمل را انجام می‌دهد (۱۴).

اساس روش شبکه عصبی برپایه ساختار مغز انسان است. پس از آنکه ایده شبکه عصبی به ذهن خطور کرد مبانی ریاضی آن پایه‌ریزی می‌شود. برمبنای کارکرد آماری، شبکه‌های عصبی در رگرسیون و سری‌های زمانی برای مدل‌سازی و پیش‌بینی به‌کار می‌رود، و از آنجا که شبکه‌ای از اجزای به‌هم مرتبط است آن را شبکه عصبی نامیده‌اند. این اجزاء از مطالعات سیستم‌های عصبی زیستی الهام می‌گیرد. به عبارت دیگر، شبکه عصبی کوششی است تا با استفاده از اجزایی که شبیه سلول‌های عصبی زیستی رفتار می‌کنند ماشین‌هایی ایجاد گردد که مانند مغز انسان کار کند. شبکه عصبی دارای قابلیت‌های طبقه‌بندی الگوها، قابلیت یادگیری و تعمیم بوده و بنابراین، از این ساختارها می‌توان برای پیش‌بینی شرایط آینده براساس تجارب گذشته استفاده کرد. شبکه عصبی یک پردازنده توزیع‌شده موازی بزرگی است که از واحدهای پردازشی ساخته شده و دارای یک تمایل طبیعی برای ذخیره دانش تجربی و ارائه پیشنهادات مناسب است. شبکه‌ها باید طوری طراحی شوند که بتوانند براساس مجموعه‌های ورودی؛ خروجی‌های مناسب و مدنظر را ایجاد کنند. برای آموزش شبکه‌های عصبی می‌توان از الگوریتم‌هایی چون یادگیری رو به جلو یا مدل پرسپترون چند لایه بهره جست (۱۵،۱۴،۱۱). یک روش بسیار مهم، روش بیز ساده است که بیز سطحی و بیز مستقل نیز نامیده می‌شود. ساخت این روش بسیار ساده است و نیازی به برنامه‌های تخمین پارامتر تکرارشونده پیچیده ندارد؛ یعنی می‌توان از آن برای مجموعه داده‌های بسیار وسیع استفاده کرد که درنهایت، این روش معمولاً فوق‌العاده عمل می‌کند. بیز ساده یکی از قدیمی‌ترین الگوریتم‌های دسته‌بندی رسمی است و هنوز حتی در ساده‌ترین شکل بسیار مؤثر است. از این مدل در دسته‌بندی متون و جداسازی اسپم‌ها به‌طور گسترده استفاده می‌شود. هدف در این بخش ایجاد قانون است که بتوان به‌راحتی افراد بیمار و سالم را در دو دسته مجزا قرار داد (۱۶). در کاربردهای امروزی یادگیری ماشین، ماشین‌بردار پشتیبان به‌عنوان یکی از قدیمی‌ترین و دقیق‌ترین متدها در میان الگوریتم‌های معروف شناخته می‌شود. همچنین ماشین‌بردار پشتیبان یکی از روش‌های یادگیری با ناظر است که از آن برای طبقه‌بندی و رگرسیون استفاده می‌کنند. مبنای کاری دسته‌بندی‌کننده این مدل، دسته‌بندی خطی داده بوده و در تقسیم خطی داده‌ها سعی بر آن است خطی انتخاب شود که حاشیه اطمینان بیشتری را داشته باشد. البته ماشین‌بردار پشتیبان در دسته‌بندی غیرخطی هم کاربرد دارد. به‌طورکلی این الگوریتم از یک نگاشت غیرخطی برای تبدیل داده‌های اصلی به ابعاد بالاتر استفاده می‌کند. هدف، یافتن بهترین تابع برای طبقه‌بندی بوده که به‌نحوی بتوان اعضای دسته‌های سالم و بیمار را در مجموعه داده‌ها از هم تشخیص داد (۱۶،۱۵).

یکی دیگر از الگوریتم‌های دسته‌بندی، درخت تصمیم است که نتایج مدل را براساس یک درخت پیاده‌سازی می‌کند. درخت تصمیم در رده‌بندی، فضای جستجو را به نواحی مستطیلی تقسیم می‌کند. یک نمونه براساس ناحیه‌ای که در آن قرارگرفته، رده‌بندی می‌شود. درخت تصمیم نیز از طریق جداسازی متوالی داده‌ها به گروه‌های مجزا ساخته می‌شود که هدف در این فرآیند، افزایش فاصله بین گره‌ها در هر جداسازی است. یکی از تفاوت‌های بین روش‌های ساخت درخت تصمیم این است که این فاصله چگونه اندازه‌گیری می‌شود. درخت‌های تصمیمی‌که برای پیش‌بینی متغیرهای دسته‌ای استفاده می‌شوند درخت‌های دسته‌بندی نام دارند؛ زیرا نمونه‌ها را در دسته‌ها یا رده‌ها قرار می‌دهند. درخت‌هایی که برای پیش‌بینی متغیرهای پیوسته استفاده می‌شوند درخت‌های رگرسیون نامیده می‌شود (۱۳،۱۲،۹). ویژگی‌های درخت تصمیم برای تقریب توابع گسسته به کار رفته و نسبت به اختلال داده‌های ورودی، مقاوم و برای داده‌های با حجم بالا بسیار مناسب است، از این‌رو در داده‌کاوی استفاده می‌شود. می‌توان درخت را به‌صورت قوانین اگر - آنگاه نمایش داد که برای کاربران قابل‌فهم بوده و امکان ترکیب عطفی و فصلی فرضیه‌ها را می‌دهد (۱۴). اغلب الگوریتم‌های یادگیری درخت تصمیم برپایه یک عمل جستجو حریصانه بالا به پایین در فضای درخت‌های موجود عمل می‌کند (۱۷).

یافته‌ها

در این مطالعه، داده‌های مورد استفاده از ۲۷۰ رکورد و ۷۶ ویژگی تشکیل می‌شد. از بین این ویژگی‌ها، ۱۳ ویژگی اصلی انتخاب و ویژگی آخر (وجود و یا عدم بیماری قلبی)، به‌عنوان متغیر هدف تعیین گردید. همه متغیرها با وزن بالاتر از ۰/۵۰ انتخاب شدند و

۱۰ مجموعه داده‌های جدید شامل تحلیل مؤلفه‌های اصلی (PCA)، ماشین‌بردار پشتیبان (SVM)، امداد (Relief)، عدم قطعیت (Uncertainty)، شاخص جینی (Gini Index)، کای‌مربع (Chi Squared)، انحراف (Deviation)، قانون (Rule)، اطلاعات به‌دست‌آمده (Information Gain) و نرخ اطلاعات حاصله (Information Gain Ratio) ایجاد شد. براین اساس، ویژگی‌های نتایج اسکن تالیوم، نوع درد سینه، شیب خط ST و تعداد رگ‌های اصلی درگیر؛ بیشترین وزن را به ترتیب دریافت کردند که به‌همین ترتیب مهم‌ترین متغیرها در تشخیص افراد سالم و بیمار بودند. متغیرهای با وزن بالاتر از ۰/۵۰ که در هر روش وزن‌دهی مشخص شده بودند، براساس نوع روش وزن‌دهی، انتخاب و یک پایگاه جدید ساخته شد. بنابراین، علاوه بر پایگاه اولیه، ۱۰ پایگاه جدید از نتایج الگوهای وزن‌دهی ساخته شد (جدول شماره ۲).

جدول شماره ۲: وزن متغیرها و مجموعه‌های جدید ایجادشده

ویژگی‌ها	تحلیل مؤلفه‌های اصلی	ماشین‌بردار پشتیبان	انحراف	امداد	عدم قطعیت	شاخص جینی	کای‌اسکوئر	قانون	اطلاعات به‌دست‌آمده	نرخ اطلاعات به‌دست‌آمده	تأثیر وزن ویژگی‌ها
نتایج اسکن تالیوم	۰/۰	۰/۸	۰/۰	۱	۱	۱	۱	۱	۱	۱	۸
نوع درد قفسه سینه	۰/۰	۰/۷	۰/۰	۰/۳	۰/۷	۰/۹	۰/۹	۱	۰/۹	۰/۹	۷
تعداد عروق بزرگ رنگی‌شده فلوروسکوپی	۰/۰	۱	۰/۰	۰/۴	۰/۷	۰/۸	۰/۸	۰/۶	۰/۸	۰/۸	۷
تورفتگی ST ناشی از تمرین نسبت به استراحت	۰/۰	۰/۵	۰/۰	۰/۱	۰/۵	۰/۶	۰/۷	۰/۶	۰/۸	۰/۶	۶
حداکثر ضربان قلب	۰/۰	۰/۶	۰/۴	۰/۱	۰/۴	۰/۶	۰/۷	۰/۱	۰/۷	۰/۶	۵
شیب خط ST در اوج ورزش	۰/۰	۰/۱	۰/۰	۰/۲	۰/۵	۰/۵	۰/۵	۰/۰	۰/۵	۰/۵	۵
کلسترول	۱	۰/۲	۱	۰/۰	۰/۱	۰/۱	۰/۲	۰/۰	۰/۵	۰/۱	۲
سن	۰/۰	۰/۰	۰/۲	۰/۰	۰/۲	۰/۳	۰/۴	۰/۰	۰/۶	۰/۳	۱
فشار خون در هنگام استراحت	۰/۱	۰/۳	۰/۳	۰/۰	۰/۱	۰/۱	۰/۱	۰/۳	۰/۷	۰/۱	۱
نتایج نوار قلب در حالت استراحت	۰/۰	۰/۳	۰/۰	۰/۱	۰/۱	۰/۱	۰/۱	۰/۲	۰/۱	۰/۱	۰

دو الگوریتم خوشه‌بندی بدون ناظر (k-means و k-medoids) در ۱۱ مجموعه پایگاه‌ها (پایگاه اصلی و ۱۰ مجموعه ایجادشده با الگوریتم وزن‌دهی) اجرا شدند. بهترین عملکرد را امداد با ۷۶٪ در هر دو الگوریتم k-means و k-medoids و بدترین عملکرد را در k-means،Chi Squared با ۵۰٪ و در k-medoids، بدتـرین عملکرد را SVM و rule با مقدار ۵۲٪ به خود اختصاص دادند. میانگین در الگوریتم k-means برابر ۶۳٪ برآورد شد، که به مجموعه SVM نزدیک بود و در الگوریتم k-medoids برابر ۶۴٪ بود که مقدار آن به مجموعه‌های کای‌اسکوئر، شاخص جینی و اطلاعات به‌دست‌آمده نزدیک بود (جدول شماره ۳).

جدول شماره۳: خوشه‌بندی ویژگی‌ها با استفاده از k-means و k-medoids

پایگاه	شماره ردیف	k-means	k-medoids
کای‌اسکوئر	۱	۰/۵۰۴	۰/۶۴۸
انحراف	۲	۰/۵۴۴	۰/۵۶۷
شاخص جینی	۳	۰/۵۰۷	۰/۶۴۸
اطلاعات به‌دست‌آمده	۴	۰/۵۲۲	۰/۶۴۸
نرخ اطلاعات به‌دست‌آمده	۵	۰/۵۱۵	۰/۶۲۶
تحلیل مؤلفه‌های اصلی	۶	۰/۵۹۳	۰/۵۶۷
امداد	۷	۰/۷۶۳	۰/۷۶۳
قانون	۸	۰/۷۰۴	۰/۵۲۶
ماشین‌بردار پشتیبان	۹	۰/۶۲۶	۰/۵۲۶
عدم قطعیت	۱۰	۰/۶۵۲	۰/۷۵۹
داده‌های خودمان	۱۱	۰/۵۱۵	۰/۶۲۲

بهترین عملکرد مربوط به شبکه عصبی در الگوریتم پرسپترون برای مجموعه داده‌های ماشین‌بردار پشتیبان با ۸۰٪ و بدترین عملکرد برای مجموعه داده‌های نرخ اطلاعات به‌دست‌آمده، شاخص جینی، کای‌اسکوئر و داده‌های مطالعه با ۴۴٪ بود که میانگین عملکرد این الگوریتم برابر ۶۲٪ برآورد شد. بهترین عملکرد در الگوریتم شبکه عصبی برای مجموعه داده ماشین‌بردار پشتیبان با ۸۳٪ و بدترین عملکرد برای مجموعه داده انحراف با ۵۵٪ بود که میانگین عملکرد این الگوریتم برابر ۶۹٪ برآورد شد. بهترین عملکرد در الگوریتم Auto Mlp برای مجموعه داده عدم قطعیت با ۸۴٪ و بدترین عملکرد برای مجموعه داده تحلیل مؤلفه‌های اصلی با ۵۴٪ بود که میانگین عملکرد این الگوریتم برابر ۶۹٪ برآورد شد (جدول شماره ۴).

جدول شماره ٤: پیاده‌سازی شبکه عصبی با استفاده از ۳ الگوریتم پرسپترون، شبکه عصبی و Auto Mlp

پایگاه	شماره ردیف	Auto Mlp	شبکه عصبی	پرسترون
عدم قطعیت	۱	۰/۸۴۱	۰/۸۱۹	۰/۵۵۲
ماشین‌بردار پشتیبان	۲	۰/۸۳۰	۰/۸۳۳	۰/۸۰۴
قانون	۳	۰/۸۰۷	۰/۷۸۱	۰/۵۵۲
امداد	۴	۰/۷۵۹	۰/۷۵۹	۰/۶۳۰
تحلیل مؤلفه‌های اصلی	۵	۰/۵۴۸	۰/۵۶۷	۰/۴۵۶
نرخ اطلاعات به‌دست‌آمده	۶	۰/۷۸۹	۰/۸۰۴	۰/۴۴۴
اطلاعات به‌دست‌آمده	۷	۰/۸۰۷	۰/۷۹۶	۰/۴۴۴
شاخص جینی	۸	۰/۷۹۳	۰/۸۰۴	۰/۴۴۴
انحراف	۹	۰/۵۵۶	۰/۵۵۶	۰/۴۵۶
کای‌اسکوئر	۱۰	۰/۸۲۶	۰/۷۹۳	۰/۴۴۴
داده‌های خودمان	۱۱	۰/۸۰۴	۰/۸۰۰	۰/۴۴۴

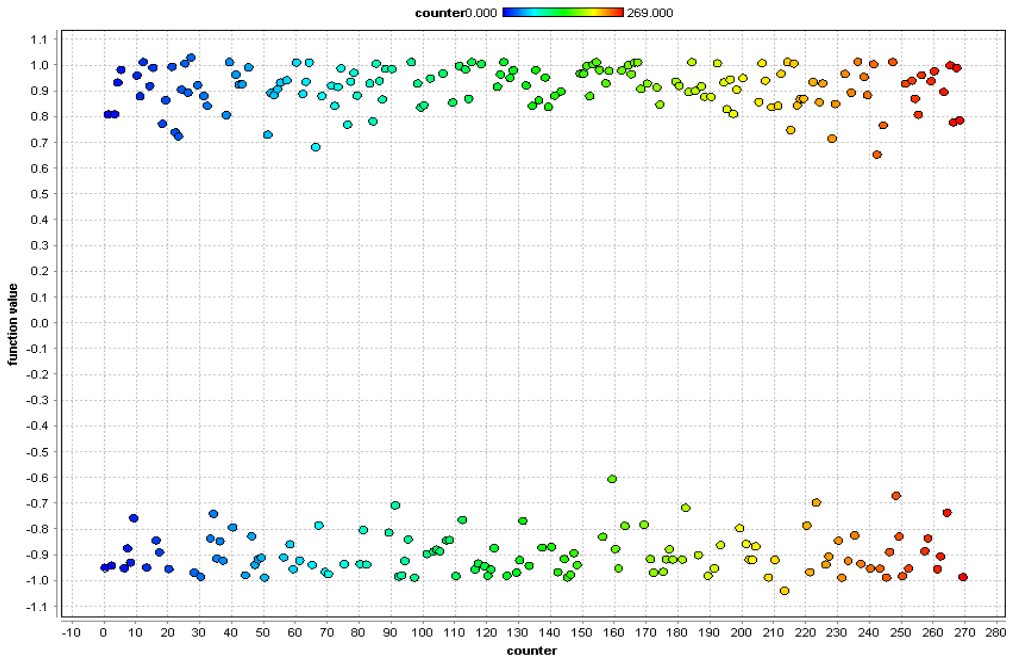
بهترین عملکرد در الگوریتم بیز هسته (Bayse Kernel) برای مجموعه داده عدم قطعیت با مقدار ۸۵٪ و بدترین عملکرد برای مجموعه داده‌های انحراف و تحلیل مؤلفه‌های اصلی با مقدار ۵۶٪ بود که میانگین عملکرد با ۷۱٪ برآورد شد.

بهترین عملکرد در الگوریتم بیز ساده برای مجموعه داده‌های اطلاعات به‌دست‌آمده، شاخص جینی و کای‌اسکوئر با مقدار ۸۵٪ و بدترین عملکرد برای مجموعه داده‌های انحراف و تحلیل مؤلفه‌های اصلی با مقدار ۵۲٪ بود که میانگین عملکرد با ۶۸٪ برآورد شد (جدول شماره ۵).

جدول شماره ۵: پیاده‌سازی الگوریتم بیز ساده و بیز هسته

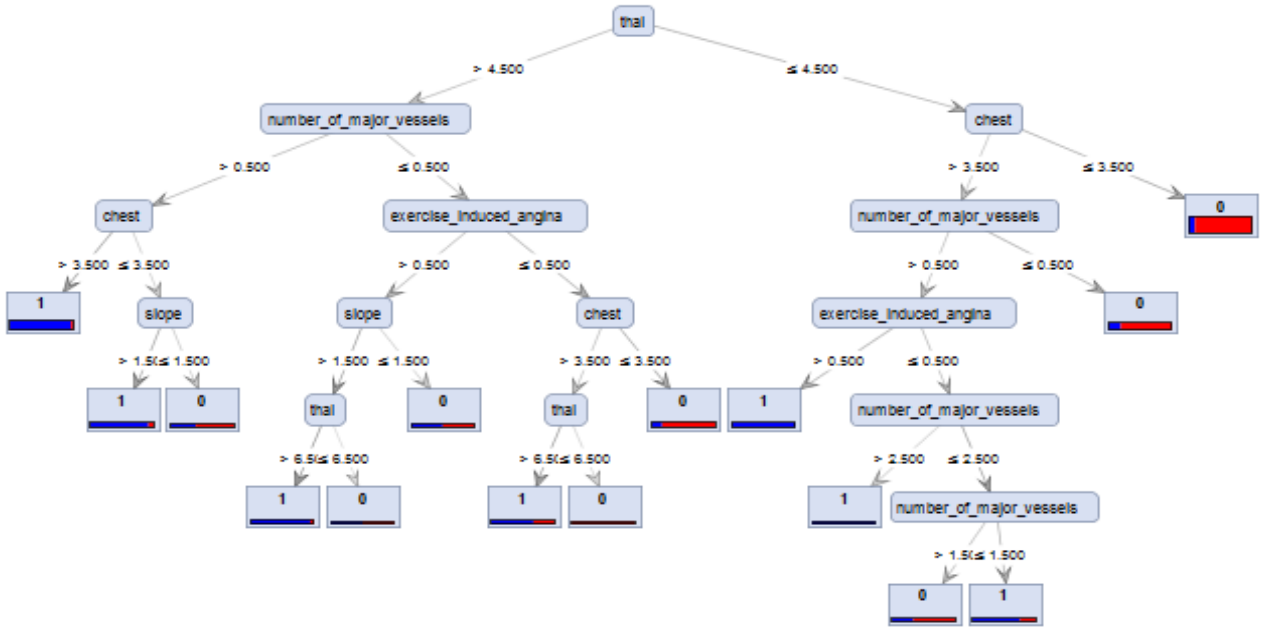
شناسه	شماره ردیف	بیز هسته	بیز ساده
قانون	۱	۰/۸۲۶	۰/۸۲۲
عدم قطعیت	۲	۰/۸۲۲	۰/۸۵۹
امداد	۳	۰/۷۶۳	۰/۷۶۳
نرخ اطلاعات به‌دست‌آمده	۴	۰/۸۴۱	۰/۷۹۶
اطلاعات به‌دست‌آمده	۵	۰/۸۵۲	۰/۸۱۵
شاخص جینی	۶	۰/۸۵۲	۰/۸۱۵
کای‌اسکوئر	۷	۰/۸۵۲	۰/۸۱۵
داده‌های خودمان	۸	۰/۸۳۷	۰/۷۸۱
انحراف	۹	۰/۵۲۶	۰/۵۶۳
تحلیل مولفه‌های اصلی	۱۰	۰/۵۲۶	۰/۵۶۳
ماشین‌بردار پشتیبان	۱۱	۰/۸۳۷	۰/۸۰۴

براساس عملکرد برای الگوریتم فراخواندن ماشین‌بردار پشتیبان تکاملی (SVM Evolutionary Recall)، نرخ اطلاعات به‌دست‌آمده برابر ۱۰۰٪ و بدترین عملکرد برای الگوریتم فراخواندن ماشین‌بردار پشتیبان فوق‌العاده (SVM Hyper Hyper Recall) بود که در مجموع، داده عدم قطعیت برابر ۳۶٪ به دست آمد (شکل شماره ۱).



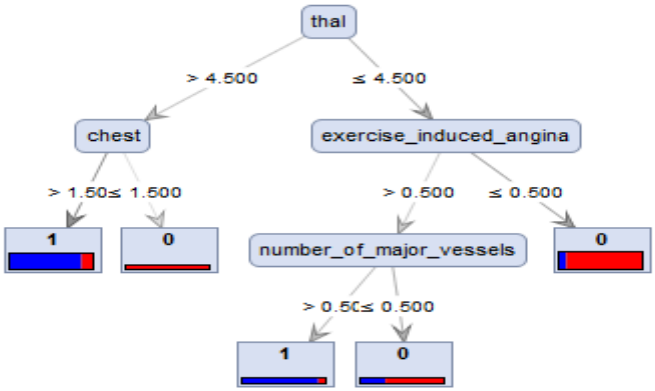
شکل شماره ۱: عملکرد ماشین‌بردار پشتیبان تکاملی.

بهترین عملکرد در میان ۱۷۶ درخت تصمیم (۱۶ مدل در ۱۱ مجموعه داده) برای مجموعه داده‌های ماشین‌بردار پشتیبان با ۸۴٪ و بدترین عملکرد برای مجموعه داده‌های انحراف و تحلیل مؤلفه‌های اصلی با ۵۴٪ بود (شکل شماره ۲).



شکل شماره ۲: درخت تصمیم مربوط به مجموعه عدم قطعیت.

thal >4.500
| number_of_major_vessels > 0.500
| | chest>3.500: 1 {1=46, 0=2}
| | chest≤3.500
| | | slope>1.500: 1 {1=11, 0=1}
| | | slope≤1.500: 0 {1=2, 0=3}
| number_of_major_vessels ≤ 0.500
| | exercise_induced_angina > 0.500
| | | slope>1.500
| | | | thal>6.500: 1 {1=15, 0=1}
| | | | thal≤6.500: 0 {1=1, 0=1}
| | | slope≤1.500: 0 {1=3, 0=3}
| | exercise_induced_angina ≤ 0.500
| | | chest>3.500
| | | | thal>6.500: 1 {1=6, 0=3}
| | | | thal≤6.500: 0 {1=0, 0=2}
| | | chest≤3.500: 0 {1=3, 0=15}
thal ≤ 4.500
| chest > 3.500
| | number_of_major_vessels>0.500
| | | exercise_induced_angina>0.500: 1 {1=11, 0=0}
| | | exercise_induced_angina≤0.500
| | | | number_of_major_vessels>2.500: 1 {1=2, 0=0}
| | | | number_of_major_vessels≤2.500
| | | | | number_of_major_vessels>1.500: 0 {1=1, 0=2}
| | | | | number_of_major_vessels≤1.500: 1 {1=3, 0=1}
| | number_of_major_vessels≤0.500: 0 {1=6, 0=25}
| chest ≤ 3.500: 0 {1=10, 0=91



شکل شماره ۳: درخت تصمیم مربوط به مجموعه عدم قطعیت

thal>4.500
| chest>1.500:1 {1=84, 0=15}
| chest≤1.500:0 {1=0, 0=6}
thal≤4.500
| exercise_induced_angina > 0.500
| | number_of_major_vessels>0.500:1 {1=18, 0=2}
| | number_of_major_vessels≤ 0.500:0 {1=7, 0=17}
| exercise_induced_angina≤0.500:0 {1=10, 0=111}

بحث

عروق کرونر، رگ‌های اصلی خون‌رسانی به خود عضله قلبی بوده و بیماری‌های عروق کرونر قلب ناشی از تنگ شدن یا اسپاسم این عروق و در نتیجه کاهش خون‌رسانی به قلب است که عوامل مختلفی باعث رخداد این بیماری در افراد می‌گردد. برای شناسایی و تشخیص این بیماری از روش‌های متفاوتی استفاده می‌شود که هر روش یک‌سری مزایا و معایب داشته و دقت آنها متفاوت است (۱۸). داده‌های حاصل از کاربرد تکنیک‌های مختلف در طول زمان جمع‌آوری شده و داده‌کاوی به‌عنوان یک روش مطلوب می‌تواند ضمن آنالیز داده‌های مختلف؛ بهترین نتایج و الگوهای پنهان را برای توصیه بهترین روش تشخیصی فراهم کند. از الگوریتم‌های داده‌کاوی در زمینه‌های مختلف تشخیص بیماری‌های قلبی استفاده می‌شود (۳).

تحقیقات متعددی برای پیش‌بینی بیماری عروق کرونر با استفاده از تکنیک‌های داده‌کاوی و مجموعه داده‌های مختلف انجام شده است به‌طور مثال در تحقیقی از سه تکنیک رگرسیون لجستیک، درخت تصمیم‌گیری، طبقه‌بندی و شبکه‌های عصبی به همراه داده‌های ۸ ریسک فاکتور مربوط به ۱۲۴۵ نفر استفاده شد که در نهایت، مدل شبکه عصبی پرسپترون چند لایه با دقت ۷۸/۷٪، بهترین مدل معرفی گردید (۱۹).

در مطالعات دیگر نیز با استفاده از شبکه عصبی بر روی مجموعه داده‌هایی با ریسک فاکتورها و اندازه‌های متفاوت به مدل‌سازی بیماری عروق کرونر پرداخته شد و مدل‌هایی با دقت ۸۹٪ و ۷۲٪ به دست آمد (۲۰).

برخی از الگوریتم‌های داده‌کاوی همچون روش‌های یادگیری برپایه ماشین می‌توانند علاوه بر آنالیز داده‌ها و استخراج الگوهای پنهان در آنها، با یادگیری تدریجی این الگوها و شرایط موجود، نسبت به پیش‌بینی وضعیت‌های مختلف در آینده اقدام کنند. به‌طور مثال از شبکه‌های عصبی کامپیوتری برای پیش‌بینی بیماری عروق کرونری در بیماران استفاده شده است (۳).

در یک مطالعه توصیفی - تحلیلی، مجموعه داده‌ها شامل ۹ ریسک فاکتور از اطلاعات ۱۳۲۲۸ نفر که در مرکز قلب تهران آنژیوگرافی شده بودند (۴۰۵۹ نفر فاقد بیماری عروق کرونر و ۹۱۶۹ نفر مبتلا به این بیماری) مورد استفاده قرار گرفت. تولید مدل پیش‌بینی بیماری عروق کرونر براساس شبکه عصبی پرسپترون چندلایه و روش گزینش متغیر، مبتنی بر درخت رگرسیون و طبقه‌بندی می‌باشد که هر دو با استفاده از نرم‌افزار Statistica انجام شده است. برای مقایسه و انتخاب بهترین مدل، از آنالیز منحنی راک استفاده گردید.

پس از ۷ مرتبه مدل‌سازی و مقایسه مدل‌های تولیدشده، مدل نهایی تشکیل‌شده از کل ریسک فاکتورهای موجود با سطح زیرمنحنی راک ۰/۷۵۴، دقت ۷۴/۱۹٪، حساسیت ۹۲/۴۱٪ و ویژگی ۳۳/۲۵٪ به دست آمد. در نتیجه انجام گزینش متغیر نیز مدلی متشکل از ۴ ریسک فاکتور با سطح زیرمنحنی راک ۰/۷۳۷، دقت ۷۴/۱۹٪، حساسیت ۹۳/۳۴٪ و ویژگی ۳۱/۱۷٪ تولید شد (۲۱). در مطالعه‌ای که توسط Ghadiri Hedeshi و Saniee Abadeh انجام گرفت، روش PSO فازی تقویت‌شده، برای تولید قوانین مناسب برای تشخیص بیماری شریان قلبی پیشنهاد گردید. این الگوریتم پیشنهادشده به تولید قوانین بهینه برای پوشش موارد، بیشتر کمک می‌کند. در این الگوریتم هر فرد شامل مجموعه‌ای از قوانین است (۳).

Sajja از الگوریتم‌های بیز ساده، درخت تصمیم و پرسپترون چندلایه بر روی مجموعه داده‌ها استفاده کرده است. دقت به‌دست‌آمده برای هریک از مدل‌های استفاده‌شده بدین صورت می‌باشد: در الگوریتم درخت تصمیم، دقت برابر ۸۱/۱۹٪؛ در الگوریتم بیز ساده، دقت برابر ۶۳/۹۷٪ و در الگوریتم پرسپترون چندلایه، دقت به‌دست‌آمده برابر ۹۱/۷۵٪ بوده که بهترین عملکرد آن در استفاده از الگوریتم پرسپترون چند لایه است (۲۲).

در مطالعه حاضر برخی از متغیرهای اندازه‌گیری‌شده مربوط به بیمارانی‌که مشکل عروق کرونر داشتند در مقایسه با افراد سالم، از چهار منبع علمی متفاوت مورد بررسی قرار گرفت. ویژگی‌های استفاده‌شده در این مطالعه عبارت از سن اشخاص، جنسیت، نوع درد قفسه سینه (که به ۴ دسته تقسیم شده بود)، فشار خون در حال استراحت، سطح کلسترول خون، میزان قند خون ناشتا، نتایج نوار قلب در حالت استراحت، حداکثر ضربان قلب، درد ناشی از ورزش، پایین‌افتادگی فاصله ST در نوار قلب ناشی از تمرین نسبت به مرحله استراحت، انحراف در شیب ST در زمان ورزش، تعداد عروق اصلی درگیر در فلوروسکوپی و نتایج اسکن‌تالیوم بود. اولین قدم در آنالیز داده‌های مورد مطالعه، تعیین مهم‌ترین متغیرها یا ویژگی‌هایی بود که باوجود و یا عدم وجود بیماری قلبی ارتباط داشتند.

روش‌های وزن‌دهی در داده‌کاوی، الگوریتم‌های مناسبی برای شناسایی متغیرهای مهم نسبت به یک متغیر هدف می‌باشند. متغیری که حداکثر وزن را دریافت کند با متغیر هدف، ۱۰۰٪ ارتباط داشته و براساس آن می‌توان افراد بیمار و سالم را به‌طور کامل از هم جدا کرد. نتایج این مطالعه برای اولین بار نشان داد متغیر نتایج اسکن‌تالیوم، بالاترین وزن‌دهی را در بین ۱۰ روش وزن‌دهی مورد استفاده در این مطالعه به خود اختصاص داده است. به عبارت دیگر، براساس یافته‌های این بخش می‌توان توصیه کرد فقط اسکن‌تالیوم می‌تواند به‌طور کامل بین افراد بیمار و سالم تفاوت قائل شده و نیازی به انجام سایر آزمایشهای تکمیلی نیست. البته در مطالعات دیگری نیز به این مطلب اشاره شده که اسکن‌تالیوم قلب یک روش آزمایشگاهی مطمئن می‌باشد (۲۵-۲۳)، ولی یافته‌های این مطالعه برای اولین بار با استفاده از الگوریتم‌های وزن‌دهی آماری و داده‌کاوی، این مطلب را به‌طور کامل اثبات کرده است. همان‌گونه که در جدول شماره ۶ مشاهده می‌شود متغیرهای نوع درد سینه و تعداد رگ‌های بسته‌شده نیز توسط حدود ۹۰٪ از روش‌های وزن‌دهی به‌عنوان متغیرهای مهم معرفی شده‌اند. از آنجا که شناسایی نوع درد ممکن براساس تجربیات و اطلاعات پزشک انجام شده و ممکن است در همه متخصصین به یک اندازه صحت و دقت نداشته باشد (۲۶). بنابراین، انتظار می‌رود این متغیر نتواند در همه موارد برای طبقه‌بندی افراد سالم و بیمار قلبی مورد استفاده قرار گیرد. اگر بتوان نوع درد قفسه صدری را به متغیرهای عددی تبدیل کرد که تمایز بین ۴ نوع درد به‌وسیله کامپیوتر و براساس روش‌های داده‌کاوی انجام شود انتظار می‌رود که این دقت نزدیک به ۱۰۰٪ شود. از طرف دیگر، اگرچه متغیر تعداد رگ‌های قلبی بسته‌شده نیاز به عملیات کلینیکی داشته و به دلیل شرایط بیماران ممکن است در همه افراد، قابل انجام نبوده و نتایج متغیر باشند. بنابراین، وزن اختصاص داده‌شده به این متغیر به‌وسیله روش‌های وزن‌دهی نیز کامل نبوده است.

جدول شماره ٦: وزن‌دهی ویژگی‌های مؤثر در پیش‌بینی بیماری عروق کرونر قلبی

متغیر	ویژگی‌ها
٨	نتایج اسکن تالیوم
٧	نوع درد قفسه سینه
٧	تعداد عروق بزرگ رنگی‌شده فلوروسکوپی
٦	تورفتگی ST ناشی از تمرین نسبت به استراحت
٥	حداکثر ضربان قلب
٥	شیب بخش ST در اوج ورزش
٢	کلسترول
١	سن
١	فشار خون در هنگام استراحت
٠	نتایج نوار قلب در حالت استراحت

طبقه‌بندی بیماران و افراد سالم براساس پایگاه‌های موجود نشان داد بالاترین درصد دقت به‌دست‌آمده در این الگوریتم‌ها، ۷۶٪ برای پایگاه داده‌های امداد بوده است. اگرچه این دقت از نظر آماری در حد قابل‌قبولی است، ولی با توجه به اینکه نتایج این مدل‌ها باید بتواند انسانهای بیمار و سالم را از هم تشخیص دهد و بهترین کارآیی این مدل از هر ۴ نفر نیز ممکن است یک‌نفر را به‌طور اشتباهی تشخیص دهد، بنابراین این الگوریتم، مدل بهینه در این مطالعه معرفی نشد.

همان‌گونه که قبلاً بیان گردید یکی از مهم‌ترین استفاده‌هایی که از مدل‌های داده‌کاوی در بخش بیماری‌های انسانی می‌شود، طراحی ابزارهای پیش‌بینی یادگیری ماشینی است. در این ابزارها، ماشین براساس الگوریتم‌های خاصی شروع به کنکاش در رابطه بین متغیر هدف و سایر ویژگی‌ها کرده و سعی می‌کند با استخراج الگوهای پنهان، مدلی را استخراج کند که بتوان بر اساس آن بین افراد بیمار و سالم تمایز قائل شد و سپس از این قابلیت برای شناسایی افراد مشکوک به بیماری در آینده استفاده کرد. در این مطالعه از چندین روش یادگیری ماشینی استفاده گردید که روش شبکه عصبی در الگوریتم پرسپترون بر روی پایگاه داده‌ای ماشین‌بردار پشتیبان با دقت ۸۰٪ به‌عنوان بهترین مدل معرفی شد. این دقت در الگوریتم‌های بیز ساده بر روی پایگاه داده عدم قطعیت به ۸۵٪ رسید که در سطح استاندارد دقت‌های قابل‌قبول در روش داده‌کاوی منظور می‌گردد. کارآیی روش‌های یادگیری ماشینی برای درخت‌های تصمیم نیز در همین حدود بود و بهترین دقت الگوریتم‌های درخت تصمیم زمانی حاصل شد که این مدل بر روی پایگاه داده ماشین‌بردار پشتیبان اجرا گردید (۸۴٪ دقت در تمایز بین گروه‌های سالم و بیمار). برای اولین بار در این مطالعه یک سیستم بهینه یادگیری ماشینی مبتنی بر مدل ماشین‌بردار پشتیبان (SVM Evolutionary Recall) ارائه شد که می‌تواند براساس پایگاه داده‌های نسبت اطلاعات به‌دست‌آمده به‌طور کامل (دقت ۱۰۰٪)، افراد بیمار قلبی و سالم را از هم تشخیص دهد. همچنین این سیستم می‌تواند به‌عنوان ابزاری در کلینیک‌های بیماری‌های قلبی به متخصصان کمک کرده تا با حداکثر دقت ممکن، افراد سالم و بیمار را از هم تشخیص دهند. طراحی و ارائه این سیستم برای اولین بار در این مطالعه گزارش شده است.

نتیجه‌گیری

نتایج این تحقیق نشان داد مدل‌های داده‌کاوی می‌توانند به‌عنوان ابزارهایی مبتنی بر یادگیری ماشینی و با دقت‌های فوق‌العاده نسبت به تشخیص گروه افراد بیمار قلبی از سالم مورد استفاده قرار گیرند. در الگوریتم خوشه‌بندی بدون ناظر، بهترین عملکرد را امداد در هر دو الگوریتم k-means و k-medoids با ۷۶٪ دارا می‌باشد. در الگوریتم‌های شبکه عصبی، بهترین عملکرد با ۸۰٪ مربوط به الگوریتم پرسپترون برای مجموعه داده‌های ماشین‌بردار پشتیبان است. بهترین عملکرد در الگوریتم بیز هسته برای مجموعه داده عدم قطعیت با مقدار ۸۵٪ است. بهترین عملکرد برای الگوریتم، فراخواندن ماشین‌بردار پشتیبان تکاملی است که در مجموعه داده‌های مطالعه و نرخ اطلاعات به‌دست‌آمده برابر با ۱۰۰٪ می‌باشد.

بهترین عملکرد در درخت تصمیم مربوط به مجموعه داده‌های ماشین‌بردار پشتیبان با ۸۴٪ است. همچنین در این تحقیق برای اولین بار نشان داده شد نتایج اسکن‌تالیوم، بهترین ویژگی برای تشخیص این دو گروه بوده و برای اولین بار یک سیستم یادگیری بر پایه ماشین ارائه شد که می‌تواند با دقت ۱۰۰٪ افراد بیمار را از سالم تشخیص دهد، از این سیستم می‌توان به‌عنوان ابزارهای کمکی در کلینیک‌های تشخیص طبی قلبی استفاده کرد.

References:

1. Neamatipour E, Sadat Heidari B. Evaluation of ankle brachial index as a predictive factor for diagnosis of coronary artery disease. Tehran Univ Med J 2006;64(1):45-48. [Full Text in Persian]

2. Ghasemi E, Mohammad Aliha J, Bastani F, Samiei N, Haghani H. General health status in women with coronary artery disease. Tehran Univ Med Sci J (Koomesh) 2013;14(4):474-82. [Full Text in Persian]

3. Herron P. Machine learning for medical decision support: Evaluating diagnostic performance of machine learning classification algorithms. INLS 110 Data Mining 2004 Spring;1-15.

4. Rani KU. Analysis of heart disease dataset using neural network approach. Int J Data Min Knowl Manage Process 2011;1(5):1-8.

5. Ghadiri Hedeshi N, Saniee Abadeh M. Coronary artery disease detection using a fuzzy-boosting PSO approach. Comput Intell Neurosci 2014;2014:783734.

6. Chau M, Shin D, Shin DK. A comparative study of medical data classification methods based on decision tree and bagging algorithms. 8th IEEE International Conference on Dependable Autonomic and Secure Computing, DASC 2009 12-14 December. China: Chengdu; 2009. p. 178-83.

7. Alizadehsania R, Habibia J, Hosseinia MJ, Mashayekhi H, Boghrati R, GHandeharioun A, et al. A data mining approach for diagnosis of coronary artery disease. Comput Methods Programs Biomed 2013;111(1):52-61.

8. Itu L, Rapaka S, Passerini T, Georgescu B, Schwemmer C, Schoebinger M, et al. A Machine learning approach for computation of fractional flow reserve from coronary computed tomography. J applied physiology. Accessed Apr 14, 2016.

9. Beiki AH, Saboor S, Ebrahimi M. A new avenue for classification and prediction of olive cultivars using supervised and unsupervised algorithms. PLoS ONE 2012;7(9):e44164.

10. Ebrahimi M, Lakizadeh A, Agha Golzadeh P, Ebrahimie E. Prediction of thermo stability from amino acid attributes by combination of clustering with attribute weighting: A new vista in engineering enzymes. PLoS ONE 2011;6(8):e23146.

11. KayvanJoo AH, Ebrahimi M, Haqshenas G. Prediction of hepatitis C virus interferon/ribavirin therapy outcome based on viral nucleotide attributes using machine learning algorithms. BMC Res Notes 2014;7:565.

12. Bagherzadeh-Khiabani F, Ramezankhani A, Azizi F, Hadaegh F, Steyerberg EW, Khalili D. A tutorial on variable selection for clinical prediction models: Feature selection methods in data mining could improve the results. J Clin Epidemiol 2016;71:76-85:26475568.

13. Alirezadei E, Forouzideh F. Behavioral analysis of false codes. [MSc Thesis]. International Kish University Paradise; 2012. [Text in Persian]

14. Mahdavi M, Taheri M, Lotfi F. Application of data mining in neural networks. In: Proceedings of 8th Symposium on progress in science and technology 2013. Mashhad. Iran; 2013. [Text in Persian]

15. Breuel TM, Shafait F. Automlp: Simple, effective, fully automated learning rate and size adjustment. University of Kaiserslautern, 67663, Germany: Kaiserslautern; 2010. Available From: http://snowbird.djvuzone.org/2010/abstracts/163.pdf. Accessed Apr 14, 2016.

16. Salari M, Adib F. Ten best data mining algorithms. 13th Student conference on electrical engineering. Tehran: Tarbiat Modares University; 2010. [Text in Persian]

17. Fazli H, Momeni H. Comparison and evaluation of data mining algorithms, decision tree and SVM application for intrusion detection. In: Proceedings of 8th Symposium progress in science and technology 2013, Mashhad. Iran; 2013. [Text in Persian]

18. Shah NP, Cainzos-Achirica M, Feldman DI, Blumenthal RS, Nasir K, Miner MM, et al. Cardiovascular disease prevention in men with vascular erectile dysfunction: The view of the preventive cardiologist. Am J Med 2016;129(3):251-9.

19. Kurt I, Ture M, Kurum AT. Comparing performances of logistic regression, classificationand regression tree, and neural networks for predicting coronary artery disease. Expert Sys Appl Sci Direct 2008;34(1):366-74.

20. Mobley BA, Schechter E, Moore WE, McKee PA, Eichner JE. Predictions of coronary artery stenosis by artificial neural network. Artif Intell Med 2000;18(3):187-203.

21. Mahmoudi I, Askari 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 Univ Med Sci 2013;15(5):47-56. [Full Text in Persian]

22. Sunitha Sajja. Data mining of medical datasets with missing attributes from different sources. [MSc Thesis]. Department of Mathematics and Statistics. Youngstown State University; 2010.

23. Ebrahimie E, Ebrahimi M, Sarvestani NR. Protein attributes contribute to halostability, bioinformatics approach. Saline Systems 2011;7(1):1.

24. Hosseinzadeh F, Ebrahimi M, Goliaei B, Shamabadi N. Classification of lung cancer tumors based on structural and physicochemical properties of proteins by bioinformatics models. PLoS One 2012;7(12).

25. Srinivas K, Raghavendra Rao G, Govardhan A. Analysisof coronary heart disease and prediction of heart attack in coalmining regions using data mining techniques. In: Proceedings of the 5th International Conference on Computer Science and Education (ICCSE ’10); August 2010; Hefei, China; 2010.

26. Sabe MA, Claggett B, Burdmann EA, Desai AS, Ivanovich P, Kewalramani R, et al. Coronary artery disease is a predictor of progression to dialysis in patients with chronic kidney disease, type 2 diabetes mellitus, and anemia: An Analysis of the Trial to Reduce Cardiovascular Events With Aranesp Therapy (TREAT). J Am Heart Assoc 2016;5(4).

Qom Univ Med Sci 2016 June

Original Article

Accurate Prediction of Coronary Artery Disease Using Bioinformatics Algorithms

Hajar Shafiee[1], Mansour Ebrahimi[2*]

[1]Faculty of Engineering, University of Qom, Qom, Iran.

[2]Department of Biology, Faculty of Basic Sciences, University of Qom, Qom, Iran.

[*]Corresponding Author:
Mansour Ebrahimi,
Department of Biology, Faculty of Basic Sciences, University of Qom, Qom, Iran.

Email:
mansour@future.edu

Received: 16 Jun, 2015

Accepted: 20 Sep, 2015

Abstract

Background and Objectives: Cardiovascular disease is one of the main causes of death in developed and Third World countries. According to the statement of the World Health Organization, it is predicted that death due to heart disease will rise to 23 million by 2030. According to the latest statistics reported by Iran's Minister of health, 3.39% of all deaths are attributed to cardiovascular diseases and 19.5% are related to myocardial infarction. The aim of this study was to predict coronary artery disease using data mining algorithms.

Methods: In this study, various bioinformatics algorithms, such as decision trees, neural networks, support vector machines, clustering, etc., were used to predict coronary heart disease. The data used in this study was taken from several valid databases (including 14 data).

Results: In this research, data mining techniques can be effectively used to diagnose different diseases, including coronary artery disease. Also, for the first time, a prediction system based on support vector machine with the best possible accuracy was introduced.

Conclusion: The results showed that among the features, thallium scan variable is the most important feature in the diagnosis of heart disease. Designation of machine prediction models, such as support vector machine learning algorithm can differentiate between sick and healthy individuals with 100% accuracy.

Keywords: Heart diseases; Coronary artery diseases; Computational biology; Support vector machine.