ISSN: 2745-3673
http://www.Rcsj.ir

مجله‌ی علمی پژوهش در علوم رایانه (سال چهارم)
شماره ۱۳ / بهار ۱۳۹۸ / ص ۶۱- ۴۹

تشخیص بیماری عروق کرونر قلبی با استفاده از روشی مبتنی بر درخت تصمیم و بیزساده

محسن غلامی [۱] ، علی برومندنیا [۲]

[۱] دانشجوی کارشناسی ارشد دانشگاه آزاد اسلامی، واحد بوشهر، ایران.

[۲] استادیار، دانشگاه آزاد اسلامی، تهران جنوب، تهران، ایران.

نام نویسنده مسئول:
محسن غلامی

چکیده

بیماری‌های قلبی یکی از شایع‌ترین علل مرگ‌ومیر در سراسر جهان است. از این‌رو محققین بدنبال یافتن راهکارهایی جهت تشخیص و پیش‌بینی سریع و ارزان این بیماری هستند. این پژوهش با هدف شناسایی متغیرهای تاثیر گذارتر، کاهش ابعاد و ارائه سیستم هوشمند به کمک داده‌کاوی جهت پیش‌بینی بیماران قلبی صورت گرفته است.

این مقاله از روش توصیفی- تحلیلی است. داده‌های بکار رفته شامل ۲۷۰ نفر از مخزن یادگیریUCI است. جهت پیش‌بینی بیماران قلبی از درخت‌تصمیم در فاز پیش‌پردازش داده‌ها برای شناسایی متغیرهای مستقل با بهره‌اطلاعاتی بیشتر برای کاهش ابعاد و بهبود کارایی طبقه‌بندی کننده بیزساده در مواجه با متغیرهای غیر مستقل استفاده گردیده است. برای مدل‌سازی از نرم‌افزارRapid Miner نسخه۹ و جهت آزمون‌های آماری و پیش‌پردازش داده¬ها از نرم‌افزارSPSS Statistics نسخه۲۵ بر روی سیستمی با مشخصات ویندوز۱۰ و مدل HP 15 Notebook PC, 4 Core(S) استفاده گردیده است.

در این مقاله از روش ترکیبی درخت‌تصمیم و بیزساده استفاده گردید که نتایج حاصل شده با دقت۸۷,۱۶٪ و صحت۸۶,۴۲٪ است. روش پیشنهادی علاوه‌بر کاهش ابعاد، باعث بهبود عملکرد طبقه‌بندی کننده بیزساده و افزایش سرعت تشخیص و بهبود بیماران می‌شود.

با مقایسه پارامترهای مختلف، روش پیشنهادی علاوه بر استفاده از متغیرهای کمتر و کاهش ابعاد، دقت و صحت بالاتری نسبت به سایر روش‌های پیشین دارد.

واژگان کلیدی: بیماری‌قلبی عروقی، کاهش ابعاد، درخت‌تصمیم، بهره‌اطلاعاتی، بیزساده.

مجلـه‌ی علمی پژوهش در علوم رایانه ، شماره ۱۳ ، بهار ۱۳۹۸، ص ۶۱- ۴۹

ISSN: 2745-3673
http://www.Rcsj.ir

مقدمه

امروزه بیماری‌های قلبی علت مرگ یک سوم مردم در سراسر جهان و علت اصلی مرگ و میر در کشورهای در حال توسعه و از مهمترین عوامل تهدید کننده سلامت انسان‌ها می‌باشد[1]. مطالعات اخیر نشان می‌دهد که علی‌رغم جوان بودن جمعیت کشور ایران، بیماری‌های قلبی و عروقی از عوامل اصلی مرگ‌ومیر است[3-2]. از این‌رو ایجاد سیستم‌های سریع، دقیق و هوشمند جهت پیش‌بینی و تشخیص زودهنگام بیماران قلبی و عروقی با استفاده از روش‌های داده‌کاوی، هوش مصنوعی و یادگیری ماشین بسیار ضروری به‌نظر می‌رسد.

در سال‌های اخیر مطالعات زیادی در زمینه بیماری‌های قلبی انجام شده و در علوم پزشکی و اطلاع‌رسانی سلامت، پیشرفت‌های چشمگیری حاصل شده است. در عین‌حال با گسترش حجم داده‌ها و فناوری‌های نوین، پتانسیل مناسبی برای تجزیه و تحلیل داده‌های پزشکی فراهم گردیده است[4]. Chenو همکاران در[1] از الگوریتم شبکه‌عصبی مصنوعی استفاده شده است. مجموعه داده بکار رفته در این پژوهش از سایتUCI و شامل ۱۳ متغیر، با۳۰۳ پرونده است. نرخ دسته‌بندی در این روش۸۰ درصد است. Safdariو همکاران در[5] الگوریتم‌های شبکه عصبی، درخت تصمیم و قوانین انجمنی را با هدف جمع‌آوری عوامل خطرزای بیماری قلبی مقایسه کردند. مطالعات مربوط به پرونده۳۵۰ بیمار بستری شده در سال۱۳۹۶ در بیمارستان قلب شهید رجایی است. نتایج این مطالعه بر اساس قوانین انجمنی، نشان داد که پنج عامل مهم سکته‌قلبی، فشارخون بالا، DLP، مصرف دخانیات، دیابت و افراد گروه خون+A بیشتر در خطر این بیماری هستند. Mahmudi در[6] از تکنیک فازی و الگوریتم ماشین بردارپشتیبان جهت تشخیص بیماری قلبی بر روی مجموعه داده‌ای با ۱۳ ویژگی و۲۷۰ پرونده از سایتUCI انجام داد. نتایج حاصل از این روش با نرخ دسته‌بندی۸۵ درصد، نرخ صحت۸۳,۵ و حساسیت ۸۵,۸ درصد است. Sabbaghgol در[7] از الگوریتمC4.5 با استفاده از مجموعه داده سایتUCI با۳۰۳ پرونده و۱۳ متغیر مدل‌سازی انجام گرفت که نرخ ویژگی۷۲,۶ و دقت۸۰,۲ درصد است. Suganya و Tamije selvy در[8] با توجه به تاثیر نامطلوب داده‌های پرت و نویز با استفاده از تکنیک‌های فازی سعی در شناسایی و حذف اینگونه داده‌ها در مجموعه داده بیماران قلبی و سپس پیش‌بینی بیماران نمودند. Srinivas و همکاران در[9] الگوریتم‌های شبکه‌عصبی و بیزساده بر روی بیماران قلبی مقایسه شدند که الگوریم بیزساده با دقت۸۴,۱۴ درصد پیشنهاد گردید. حسین‌خانی و همکاران در[10] عوامل خطر بیماری‌های قلبی‌عروقی در بالغین شهر قزوین بررسی گردید که نتایج این مقاله نشان از شیوع بالای عوامل خطرساز بیماری‌های قلبی و عروقی و سایر بیماری‌های غیرواگیر مانند دیابت و فشار خون شریانی است.

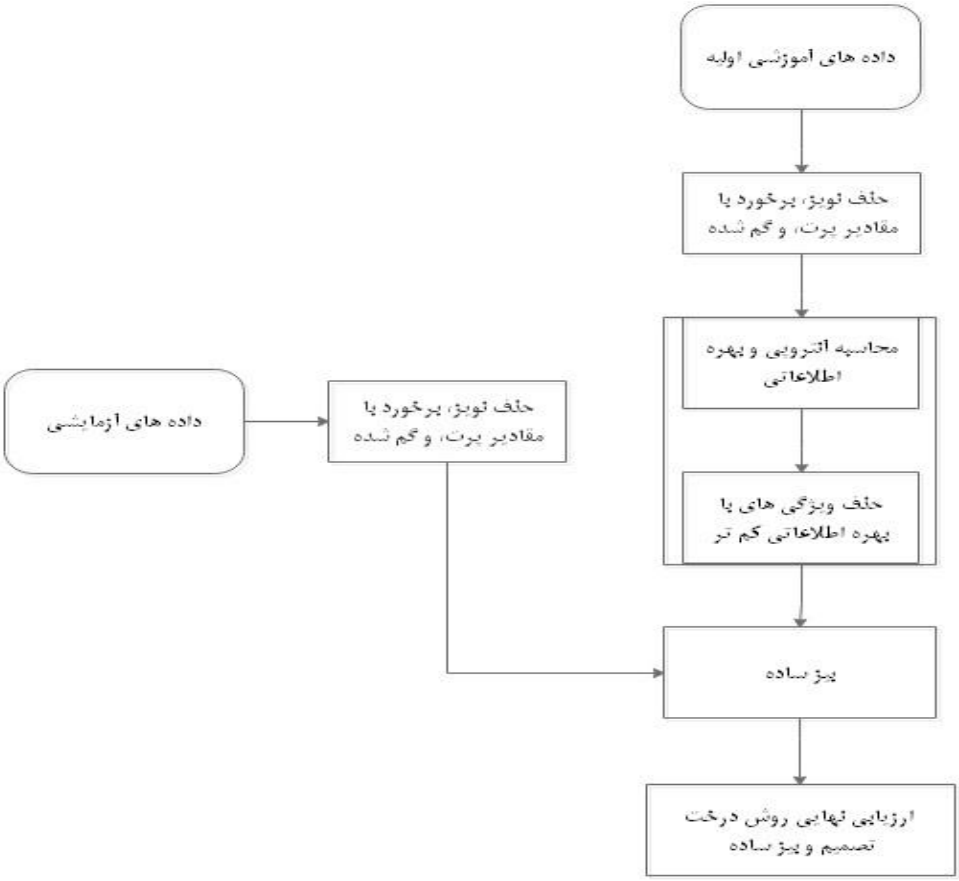
در اکثر تحقیقات صورت گرفته با مقدار بسیار زیاد و متنوعی از داده‌ها با ابعاد بالا مدل‌سازی صورت گرفته است که باعث پیچیدگی، کاهش سرعت و غیر مشهود بودن است. از این‌رو انجام تحقیقاتی مبتنی بر کاهش ابعاد و شناسایی متغیرهای تاثیر گذارتر در فاز پیش-پردازش بسیار ضروری است که در تحقیقات اخیر کمتر به آن پرداخته شده یا در بعضی تحقیقات، همچون[6] به عنوان کارهای آتی پیشنهاد داده شده است.

هدف اصلی این مقاله کاهش تعداد متغیرها با استفاده از اضافه کردن یک مرحله به فاز پیش‌پردازش داده‌ها جهت شناسایی متغیرهای با تاثیر گذاری بالا و استقلال بیشتر و حذف متغیرهای دارای ارزش کم‌تر جهت افزایش سرعت، دقت و صحت تشخیص بیماران قلبی به کمک درخت‌تصمیم و معیار بهره‌اطلاعاتی برای حل مشکل بیز ساده در مجموعه‌های با مشخصه های وابسته است و در ادامه با استفاده از الگوریتم بیز ساده جهت پیش‌بینی و تشخیص بیماران قلبی عروقی بر روی مجموعه دادهHeart از مخزن یادگیریUCI است.

روش

این مطالعه از نوع، توصیفی- تحلیلی است که شامل سه مرحله پیش‌پردازش، مدل‌سازی و تست است. داده‌های بکار رفته در این پژوهش شامل۱۳ ویژگی مربوط به۲۷۰ نفر از مخزن یادگیری ماشینUCI و فاقد مقادیر گم شده است[11]. در این مقاله از نرم‌افزار Rapid Miner نسخه۹ جهت مدل‌سازی و ارزیابی استفاده گردیده است. این نرم‌افزار شامل طیف گسترده‌ای از امکانات جهت آماده‌سازی اطلاعات اولیه، بصری‌سازی، مدل‌سازی و تحلیل و ارزیابی است[12]. همچنین برای پیش‌پردازش داده‌ها و آزمون‌های آماری از نرم‌افزار SPSS Statistics نسخه۲۵ استفاده شده است. این نرم‌افزار شامل طیف گسترده‌ای از آزمون‌های آماری است. مدل پیشنهادی بر روی لپ‌تاپHP با مشخصات(AMD A4-6210 APU with AMD Radeon R3 Graphics, 1800 Mhz, 4 Core(s مدل‌سازی شده است. سیستم-عامل بکار رفته در این پژوهش ویندوز۱۰، ۶۴بیت است. در این روش سعی شده متغیرهای تاثیرگذارتر انتخاب و کاهش ابعاد صورت گیرد و بتواند هرچه بهتر بیماران قلبی را تشخیص دهد. در ادامه مراحل روش پیشنهادی شرح و فلوچارت آن در (شکل۳) آمده است.

ISSN: 2745-3673
http://www.Rcsj.ir



شکل۳: فلوچارت روش پیشنهادی

در جدول۵ متغیرهای مجموعه‌داده معرفی و جامعه آماری مورد ارزیابی قرار گرفته است.

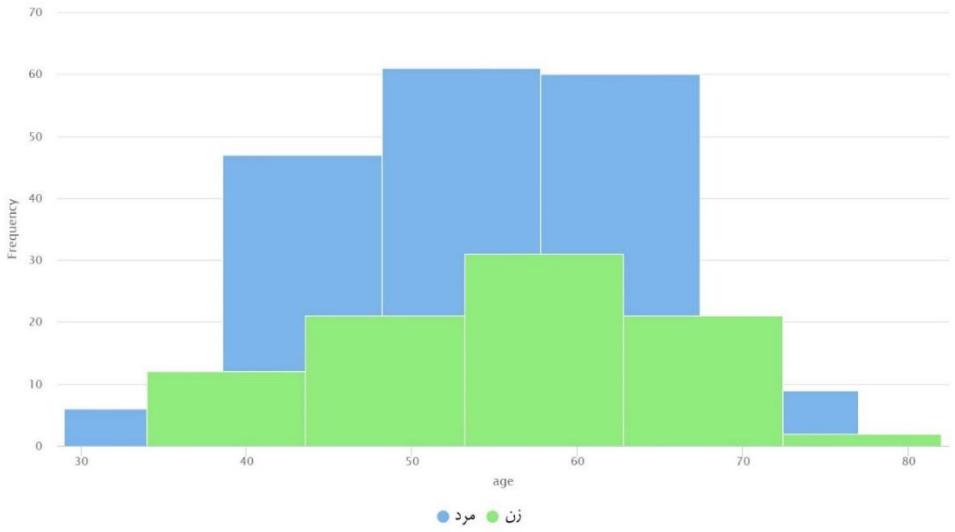
جدول ۵: متغیرهای مجموعه داده

نام صفت	توضیحات	مقادیر	دتوضیحات
Age	سن بیمار	بین ۲۹ الی ۷۶	میانگین ۵۴,۴
Sex	جنسیت	ممرد۱، زن ۰	۱۸۳1 مرد، ۸۷ زن
Chest pain type	درد قفسه سینه	۱ آنژین صدری معمولی ۲ درد قلبی ۳ بدون آنژین ۴ بدون علامت	۲۰ مورد آنژین صدر معمولی، ۴۲ مورد درد قلبی، ۷۹ مورد بدون آنژین ۱۲۹، مورد بدون علامت
Resting blood pressure	فشارخون در زمان استراحت	بین ۹۴ تا ۲۰۰	مقادیر پیوسته
Serum cholesterol	کلسترول(چربی بد خون)	۱۲۶ تا ۵۶۴	مقادیر پیوسته mg/dl
Fasting blood sugar> 120 mg/dl	قند خون ناشتا	۰ بله ۱ خیر	۲۳۰ مورد بله، ۴۰ مورد خیر
Resting electrocardiographic results	نتایج نوار قلب در حال استراحت که شامل 3 مقدار نرمال، موج غیر قلبی و نشاندهندهی افزایش مقطعی	۰ عادی ۱ موج غیر قلبی ۲ موج افزایش مقطعی	۱۳۱ مورد عادی، ۲ مورد موج غیر قلبی، ۱۳۷ مورد موج افزایش مقطعی

ISSN: 2745-3673
http://www.Rcsj.ir

	یا احتمالی ضخامت بطن چپ است.		
Maximum heart rate achieved	ماکزیمم ضربان قلب به دست آمده	بین ۷۱ تا ۲۰۲	مقادیر پیوسته
Exercise induced angina	آنژین ناشی از ورزش (فعالیت)	۰ خیر ۱ بله	۱۸۱ مورد خیر، ۸۹ مورد بله
St depression induced by exercise relative	افسردگی ST ناشی از تست ورزش نسبت به استراحت	بین ۰ تا ۶,۲	مقادیر پیوسته
The slope of the peak exercise ST segment	بیان کننده شیب قطعه St در زمان حداکثر ورزش	۱ شیب بالا ۲ مسطح ۳ شیب به پایین	۱۳۰ مورد شیب بالا، ۱۲۲ مورد مسطح، ۱۸ مورد شیب به پایین
Number of major vessels colored by Fluoroscopy	تعداد عروق اصلی رنگی شده توسط فلورسکوپی	شامل مقادیر ۰،۱،۲،۳	۱۶۰مورد صفر، ۵۸ مورد یک، ۳۳ مورد دو، ۱۹ مورد سه
Thal	اسکن تالیوم	۳ معمولی ۶ نقص ثابت ۷ نقص برگشت پذیر	۱۵۲ دو مورد معمولی، ۱۴ مورد نقص ثابت، ۱۰۴ مورد نقص برگشت پذیر
صفت تشخیصی	تشخیص بیماری قلبی	۱ سالم ۲ بیماری قلبی	۱۵۰ مورد سالم، ۱۲۰ مورد بیمار قلبی

در شکل۱ بازه سنی، تعداد شرکت کنندگان و جنسیت افراد در این مجموعه داده نمایش داده شده است که از۱۸۳ مرد ،۵۳ نفر در بازه۲۹ تا۴۸ سال و۶۱ نفر بین۴۸ تا۵۸ سال و۶۹ نفر بین سنین۵۸ تا ۷۷ سال هستند که از این تعداد۸۳ نفر سالم و۱۰۰ نفر دارای بیماری قلبی هستند. همچنین از۸۷ زن،۳۳ نفر بین سنین۳۴ تا۵۳ سال و۳۱ نفر بین سنین۵۳ تا۶۳ سال و۲۳ نفر بین سنین۶۳ تا ۷۶ سال هستند که۶۳ نفر از این تعداد سالم و۲۰ نفر دارای بیماری قلبی هستند. انحراف معیار از میانگین متغیر سن۹,۱۰۹ است و میانگین۵۴,۴۳۳ سن شرکت کنندگان است.



شکل۱: بررسی تعداد، بازه سنی و جنسیت

ISSN: 2745-3673
http://www.Rcsj.ir

پیش‌پردازش و آماده‌سازی داده‌ها

در روش‌های کاهش ابعاد به بررسی ارتباط و همبستگی یا استفاده از روش‌های کلاسه‌بندی در پیش‌پردازش داده‌ها مثل درخت-تصمیم استفاده می‌شود که به ترتیب در این مطالعه بررسی و نتایج آن ذکر گردیده است.

آآماده‌سازی و پیش‌پردازش داده‌ها یکی از مراحل اساسی در داده‌کاوی محسوب می‌شود به صورتی که نامفهوم بودن داده‌ها یا استفاده نادرست از ابزار داده‌کاوی می‌تواند این فرآیند را در مسیری نادرست قرار دهد. از سویی پایگاه داده‌های امروزی به دلیل حجم بالا مستعد داده‌های نادرست و ناسازگار هستند[12]. با توجه به نوع متغیرهای این مجموعه داده از روش آماری کای‌دو برای متغیرهای نامینال و برای متغیرهای پیوسته نیز از ضریب خطی همبستگی پیرسون استفاده گردید. در ادامه مراحل پیش‌پردازش و آماده‌سازی داده‌ها تشریح می‌گردد.

نرمال‌سازی داده‌ها

ننرمال‌سازی داده‌ها وقتی داده‌ها در یک دامنه نیستند انجام می‌گردد تا در یک دامنه مشابه قرار گیرد. ویژگی‌های با مقادیر بزرگ ممکن است اثر بسیار زیادتری در تابع هزینه نسبت به ویژگی‌های با مقادیر کم داشته باشند. این مشکل با نرمالیزه نمودن ویژگی‌ها طوری که مقادیرشان در دامنه‌های مشابه قرار گیرند برطرف خواهد شد تا اهمیت داده‌ها به واحد اندازه‌گیری آن‌ها بستگی نداشته باشد. استفاده از داده‌های نرمال نشده ممکن است روی نتایج حاصل از تحلیل‌ها اثر نامناسبی داشته باشد.

در اینجا می‌توانیم از هر تبدیلی برای نرمال کردن داده‌ها استفاده کنیم از جمله تبدیل‌های خطیZ-یاMin-Max ،T- Score Score، که در این مقاله برای متغیر سن به دلیل مشخص بودن محدوده از تبدیلMax-Min و برای نرمال کردن متغیرهای نتایج نوار قلب در حال استراحت، کلسترول(چربی بد خون)، ماکزیمم ضربان قلب به دست آمده، افسردگیST ناشی از تست ورزش نسبت به استراحت، از تبدیلZ- Score استفاده می‌گردد.

ننرمال‌سازی Min-Max

با داشتن حداقل و حداکثر مقادیر موجود می‌توانیم از این روش استفاده کنیم و به هر محدوده جدید دلخواهی نرمال‌سازی کنیم.

$$v = \frac{v - Min}{Max - Min}(NewMax - NewMin) + NewMin \quad (۱)$$

کهv عددی است که می‌خواهیم نرمال شود، Minحداقل موجود در مجموعه و Maxحداکثر موجود برای صفت خاصه مورد نظر است. NewMinو NewMax محدوده جدید است. این روش یک تغییر شکل خطی بر روی داده‌های اولیه ایجاد می‌کند.

نرمال‌سازی Z- Score

این روش برای مواقعی مناسب است که در نمونه‌های صفت‌خاصه مقدار خارج از محدوده وجود داشته باشد و همینطور مقدار حداکثر و حداقل داده‌ها مشخص نیستند. که از میانگین و انحراف استاندارد صفات‌خاصه استفاده می‌کند، برای تبدیل مقدار V_{old} به مقدار جدید V_{new} از فرمول زیر در این مقاله استفاده گردیده است.

$$V_{new} = \frac{V_{old} - Mean}{St\_Dev} \quad (۲)$$

که در آنMean میانگین مقادیر صفات‌خاصه مورد نظر وSt_ Dev انحراف استاندارد همان مقادیر است.

روش آماری کای‌دو

تکای‌دو یک روش خودکار گسسته‌سازی است که با کمک کلاس‌های نمونه‌ها، شباهت‌های بین توزیع داده‌ها در دو بازه همسایه را تعیین می‌کند. این آزمون به بررسی وجود ارتباط بین دو متغیر می‌پردازد و برای رابطه‌هایی بکار می‌رود که هر دو متغیر ناپارامتری باشد. اگر در یک نمونه هیچ رابطه سیستماتیک بین دو متغیر وجود نداشته باشد، می‌توان نتیجه گرفت که دو متغیر از یکدیگر مستقل هستند.

$$x^2 = \sum_{t=1}^{m} \frac{(O_t - E_t)^2}{E_t} \quad (۳)$$

ISSN: 2745-3673
http://www.Rcsj.ir

که در آنO فراوانی‌های مشاهده شده وE فراوانی‌های مورد انتظار است. فراوانی‌های مورد انتظار نباید در هیچ مقوله‌ای صفر باشد. مجموع مقوله‌هایی که مقدار مشاهدات مربوط به آنها کمتر از۵ است، نباید بیش از۲۰ درصد کل مقوله‌ها باشد. برای اطمینان از عدم همبستگی بین داده‌های مندرج و عدم در نظر گرفته شدن ستون‌های با همبستگی قبل از فرآیند داده‌کاوی می‌باید از این امر اطمینان حاصل کنیم. که با انجام تست کای‌دو روی ستون‌های انتخاب شده، به این نتیجه رسیدیم که نشان دهنده عدم وجود وابستگی بین آن‌ها است. در جدول شماره۱،۲ و۳ نتایج ارزیابی بعضی از متغیرها نشان داده شده است. که با توجه به اینکه تمامی نتایج مربوط به Asymptotic Significance زیر۰،۰۵ قرار دارد، فرض صفر یا همبستگی بین متغیرهای نامینال منتفی خواهد بود.

جدول ۱: نتایج تست کای‌دو بین متغیر درد قفسه سینه و آنژین ناشی از ورزش

	Value	df	Asymptotic Significance (2-sided)
Pearson Chi-Square	51.409[a]	3	.000
N of Valid Cases	270		

جدول ۲: نتایج تست کای‌دو بین متغیر number of major vessels colored by flourosopy * exercise induced angina
تعداد عروق اصلی رنگی شده توسط فلورسکوپی و آنژین ناشی از ورزش (فعالیت)

	Value	df	Asymptotic Significance (2-sided)
Pearson Chi-Square	11.898[a]	3	.008
N of Valid Cases	270		

جدول ۳: نتایج تست کای‌دو بین متغیر تعداد اسکن تالیوم و آنژین ناشی از ورزش (فعالیت)

	Value	df	Asymptotic Significance (2-sided)
Pearson Chi-Square	27.899[a]	2	.000
N of Valid Cases	270		

برای متغیرهای غیرنامینال، ضریب همبستگی بین متغیرها ارزیابی گردید که نتایج آن در جدول۴ قابل مشاهده است.

محاسبه ضریب‌همبستگی

در بررسی صفات گوناگون، در یک جامعه آماری، تغییرات دو متغیر مستقیماً بر یکدیگر موثر هستند و به عبارتی، یکی از صفت‌ها، در بزرگی و کوچکی اندازه صفت دیگر دخالت دارد. شاخص تعیین کننده همبستگی را ضریب‌همبستگی گوییم. اگر با افزایش متغیری، متغیر دیگر افزایش پیدا کند یا با کاهش متغیری متغیر دیگر نیز کاهش پیدا کند، همبستگی مثبت یا مستقیم است. اگر با کاهش یک متغیر، متغیر دیگر افزایش پیدا کند، یا بلعکس، همبستگی را منفی یا معکوس می‌نامیم. همبستگی حالت دیگری نیز دارد که افزایش یا کاهش یک متغیر در دیگر متغیر بدون تاثیر است که آن را همبستگی صفر می‌نامیم. در این مقاله شدت همبستگی متغیرها نیز با استفاده از ضریب همبستگی خطی پیرسون با فرمول زیر محاسبه شده است.

$$\frac{n(\sum xy)-(\sum x)(\sum y)}{[n\sum x^2-(\sum x)^2][n\sum y^2-(\sum y)^2]} \quad (۴)$$

که در آنx وy دو متغیر وn تعداد ردیف‌ها است و ضریب‌همبستگی مقداری بین۱ و۱- دارد.

ISSN: 2745-3673
http://www.Rcsj.ir

جدول ۴: نتایج محاسبه ضریب همبستگی برای متغیرها

متغیر اول	متغیر دوم	مقدار همبستگی
سن	ماکزیمم ضربان قلب به دست آمده	-0.402
ماکزیمم ضربان قلب به دست آمده	افسردگی ST ناشی از تست ورزش نسبت به استراحت	-0.349
فشارخون در زمان استراحت	ماکزیمم ضربان قلب به دست آمده	-0.039
کلسترول(چربی بد خون)	ماکزیمم ضربان قلب به دست آمده	-0.019
کلسترول(چربی بد خون)	افسردگی ST ناشی از تست ورزش نسبت به استراحت	0.028
فشارخون در زمان استراحت	کلسترول(چربی بد خون)	0.173
سن	افسردگی ST ناشی از تست ورزش نسبت به استراحت	0.194
سن	کلسترول(چربی بد خون)	0.220
فشارخون در زمان استراحت	افسردگی ST ناشی از تست ورزش نسبت به استراحت	0.223
سن	فشارخون در زمان استراحت	0.273

درخت تصمیم

دروشی نظارت شده است که توانایی استفاده از قوانین تصمیم‌گیری جهت رسیدن به متغیر هدف را فراهم می‌سازد. ساختار آن درختی بوده از ریشه شروع و به گره برگ می‌رسد. مزایای این روش سادگی، درک و قابلیت تفسیر ساده و امکان هرس‌کردن تصمیم‌هایی که قابل تعمیم نیستند می‌باشد[13]. روش کار الگوریتم بصورت سلسله مراتبی است که بر اساس داده‌های آموزشی با انتخاب یکی از صفات خاصه در هر مرحله شروع به‌کار می‌کند. در ادامه با تقسیم‌بندی هر یک از صفات ادامه می‌دهد تا زمانیکه تمام داده‌ها به اطلاعات دارای برچسب واحد کلاس شوند[14].

برای محاسبه آنتروپی که محاسبه میزان خالص بودن است. در هر نود، خصوصیتی که بیشترین کاهش را در آنتروپی نمونه‌ها ایجاد می‌کند، انتخاب می‌شود[15]. که برای ویژگی با مقدار بولی با فرمول زیر قابل محاسبه است:

$Entropy(S) = -p_{\oplus} log_2 p_{\oplus} - p_{\ominus} log_2 p_{\ominus}$	ف(۵)

که $p_{\oplus}$ نسبت مثال‌های مثبت به کل مثال‌ها و $p_{\ominus}$ نسبت مثال‌های منفی به کل مثال‌ها است. برای محاسبه آنتروپی برای ویژگی که دارای c مقدار مختلف باشد آنتروپی S نسبت به این دسته‌بندی c گانه غیر بولی به صورت زیر تعریف می‌شود:

$$Entropy(S) = \sum_{i=1}^{c} -p_i \, log_2 p_i \quad (۶)$$

که در آن p_i نسبتی از S است که به دسته i تعلق دارند. Log همچان در مبنای دو در نظر گرفته می‌شود که حداکثر آنتروپی می‌تواند log_2^c باشد. برای محاسبه بهره‌اطلاعاتی یک ویژگی که عبارت است از مقدار کاهش آنتروپی که بواسطه جداسازی مثال‌ها از طریق این ویژگی حاصل می‌شود از فرمول زیر استفاده گردیده است.

$Gain(S.A) = Entropy(S) - \sum_{v \in Values(A)} \frac{\lvert S_v \rvert}{\lvert S \rvert} Entropy(S_v)$	(۷)

ISSN: 2745-3673
http://www.Rcsj.ir

که در بهره‌اطلاعاتی Gain(S, A) برای یک ویژگی نظیر A نسبت به مجموعه مثال‌های S است. Values (A) مجموعه همه مقدار ویژگی‌های A بوده و S_v زیر مجموعه‌ای از S است که برای آن A دارای مقدار V است.

بیز ساده

الگوریتم بیزساده بر پایه تئوری بیز عمل می‌کند. این الگوریتم برای کار با دیتاست‌های بزرگ به دلیل سادگی مناسب بوده و نتایج پیش‌بینی آن مناسب و از نوع یادگیری با نظارت است[16]. این طبقه‌بندی کننده ساده و شناخته شده است و در مواقعی که تعداد مشاهدات کمی در دسترس باشد نیز عملکرد خوبی دارد. روش بیز روشی برای دسته‌بندی پدیده‌ها، بر پایه احتمال وقوع یا عدم وقوع یک پدیده‌است[17].

مجموعه نمونه‌های‌آموزشی با برچسب کلاس و یک نمونه تست E با n مقدار مشخصه $(a_1.a_2.\dots.a_n|c)$ را در نظر گرفته و طبقه‌بندی‌کننده بیزین را برای طبقه‌بندی کننده E به صورت زیر تعریف می‌کنیم:

$$c(E) = arg_{c\epsilon C}\ max\ P(c)P(a_1.a_2.\cdots.a_n|c) \quad (۸)$$

که فرض پایه‌ای بیزساده این است که در هر کلاس مقادیر مشخصه‌ها از یکدیگر مستقل هستند که با قانون احتمالی استقلال داریم:

$P(a_1.a_2.\cdots.a_n\|c) = P(a_1\|c)\ P(a_2\|c)\cdots P(a_n\|c)$ $P(a_1.a_2.\cdots.a_n\|c) = \prod\prod_{i=1}^{n} p(a_i\|c)$	(۹)

که با جایگذاری فرمول۹ در فرمول۸ طبقه‌بند بیزساده به صورت زیر خواهد بود:

$c_{NB}(E) = arg_{c\epsilon C}\max P(c)\prod\prod_{i=1}^{n} P(a_i\ \|c)$	(۱۰)

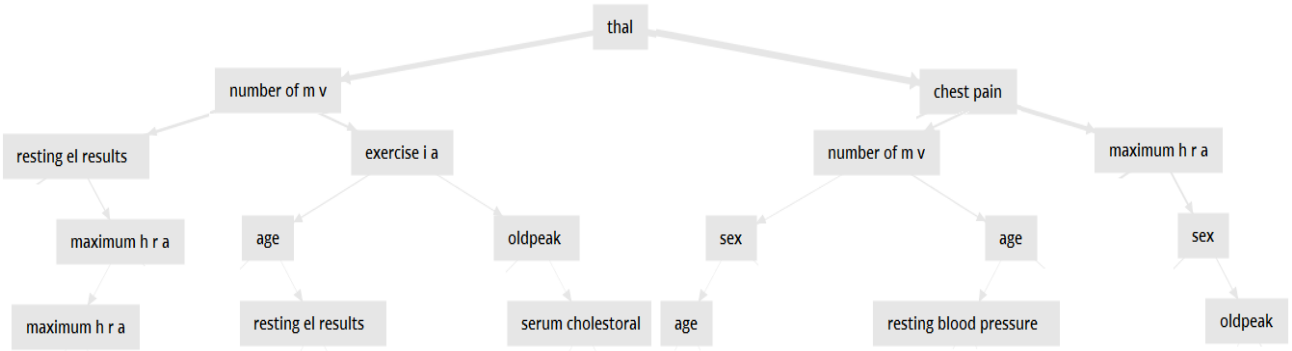
که $c_{NB}(E)$ به معنی طبقه‌بند بیزساده بر روی نمونه تست E است. که با کمک این فرمول تمامی احتمالات می‌توانند مستقیماً از روی داده‌های آموزشی تعیین شوند[18]. فرض مستقل بودن متغیرها فرض محدود کننده این روش است که در صورت عدم رعایت آن بر کارایی این الگوریتم تاثیر منفی دارد.

کاهش‌متغیرها

درشکل۲ نتایج مدل‌سازی درخت‌تصمیم بدون نمایش برگ‌ها است. در این مقاله بجای استفاده از۱۴ متغیر که در بیشتر روش‌های پیشین بکار رفته است. تنها از۵ متغیر جهت مدل‌سازی استفاده گردیده است به صورتی که گره‌های سه سطح اول درخت انتخاب و جهت مدل‌سازی در مرحله بعد بکار برده می‌شود. متغیرهای انتخابی که در بالاترین سطوح درخت‌تصمیم دارای بهره اطلاعاتی بیشتری هستند در جدول۶ نشان داده شده است.

متغیرهای اسکن تالیوم، درد قفسه سینه، تعداد عروق اصلی رنگی شده توسط فلورسکوپی، آنژین ناشی از ورزش(فعالیت) و ماکزیمم ضربان قلب به دست آمده دارای بهره‌اطلاعاتی بیشتری نسبت به سایر متغیرها هستند و در نتیجه استقلال و توانایی بیشتری نسبت به برآورده کردن فرض اولیه بیزساده در تصادفی و مستقل بودن متغیرها را دارا هستند. جدول۵ نتایج محاسبه مقدار بهره‌اطلاعاتی نشان داده شده است.

ISSN: 2745-3673
http://www.Rcsj.ir



شکل۲: درخت تصمیم مدل‌سازی شده

جدول ۵: نتایج محاسبه بهره اطلاعاتی

نام متغیر	مقدار بهره اطلاعاتی
اسکن تالیوم	0.208
درد قفسه سینه	0.192
تعداد عروق اصلی رنگی شده توسط فلورسکوپی	0.175
آنژین ناشی از ورزش (فعالیت)	0.130
ماکزیمم ضربان قلب به دست آمده	0.120
افسردگی ST ناشی از تست ورزش نسبت به استراحت	0.119
بیان کننده شیب قطعه St در زمان حداکثر ورزش	0.110
جنسیت	0.067
سن	0.057
کلسترول(چربی بد خون)	0.027
نتایج نوار قلب در حال استراحت	0.024
فشارخون در زمان استراحت	0.016
قند خون ناشتا	00.00

مدل پیشنهادی

هدف از این مقاله حذف متغیرهای مازاد و نامرتبط است که با استفاده از روش بهره‌اطلاعاتی و آنتروپی متغیرهای مستقل با وزن و اطلاعات بیشتر که برای طبقه‌بندی مفیدتر است انتخاب و متغیرهای دیگر حذف می‌گردد. استفاده از بهره‌اطلاعاتی یکی از روش‌های محقق شدن هدف حذف متغیرهای نامرتبط است و آنتروپی به طور عمده در مقیاس تئوری اطلاعات مورد استفاده قرار می‌گیرد که میزان خالص بودن مجموعه دلخواهی از نمونه‌ها را توصیف می‌کند. فرض اولیه در دسته‌بندی کننده بیزساده این است که متغیرها به صورت تصادفی و مستقل هستند ولی معمولاً این فرض در عمل نقض می‌گردد. در صورتی که دو یا چندین متغیر وابستگی بالایی به‌هم داشته باشند و فرض اولیه نقض گردد با توجه به اینکه آن مشخصه‌ها وزن بسیار زیادی در تصمیم‌گیری تعلق یک نمونه به یک کلاس می‌گیرند باعث کم شدن کارایی و صحت پیشگویی در دامنه‌هایی با مشخصه‌های وابسته در این روش می‌گردد. این مشکل در درخت تصمیم وجود ندارد، زیرا در صورتی‌که دو متغیر وابسته به‌هم باشند، فقط یکی از آن‌ها برای جدا کردن مجموعه‌داده آموزشی استفاده می‌شود. از این‌رو با استفاده از معیارهای انتخاب گره در درخت‌تصمیم، می‌توان متغیرهای وابسته را شناسایی و برای رفع مشکل وابستگی متغیرها در الگوریتم بیز ساده آن‌ها را حذف نمود، درخت‌تصمیم وظیفه پیدا کردنِ ویژگی‌های دارای اطلاعات زیادتر را دارد و آن‌ها را باید در سطوحِ بالاتری از درخت قرار دهد. هر بار که یک ویژگی در سطحی از درخت انتخاب شد، زیرْ درخت‌های آن نیز دقیقا به همان صورت(ویژگی‌هایی با اطلاعات بالا) انتخاب می‌شوند و در سطوح و گره‌های بعدی قرار می‌گیرند. در این روش با اضافه کردن یک فاز شناسایی، متغیرهای دارای استقلال و

ISSN: 2745-3673
http://www.Rcsj.ir

اطلاعات بیشتر با محاسبه بهره‌اطلاعاتی در سه سطح بالاتر درخت انتخاب و به صورتBackward Selection مجموعه داده برای فرآیند مدل‌سازی با دسته‌بندی کننده بیز ساده کاهش ابعاد می‌یابد. این مرحله از پیش‌پردازش علاوه بر کاهش ابعاد، باعث انتخاب متغیرهای مستقل جهت حل مشکل بیزساده در مواجه با متغیرهای غیر مستقل و در نتیجه باعث بهبود کارایی این الگوریتم می‌شود. در مرحله مدل-سازی بجای استفاده از۱۳ ویژگی فقط از۵ ویژگی برای مدل‌سازی در بیز ساده استفاده می‌شود. نتایج روش پیشنهادی در مقایسه با سایر الگوریتم‌ها در جدول۶ نشان داده شده است.

معیارهای ارزیابی روش

در این روش از سه معیار ارزیابی صحت، دقت و فراخوان برای مقایسه روش‌ها استفاده شده است. که نحوه محاسبه هر یک به صورت زیر است.

$$\text{Precision} = \frac{TP}{TP + FP} \quad (۱۱)$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (۱۲)$$

$$Recall = \frac{TP}{TP + FN} \quad (۱۳)$$

TN، تعداد رکوردهایی است که دسته واقعی آنها منفی بوده و الگوریتم دسته‌بندی نیز دسته آنها را به درستی منفی تشخیص داده است. TPتعداد رکوردهایی است که دسته واقعی آنها مثبت بوده و الگوریتم دسته‌بندی نیز دسته آنها را بدرستی مثبت تشخیص داده است. FP تعداد رکوردهایی است که دسته واقعی آنها منفی بوده و الگوریتم دسته‌بندی دسته آنها را به اشتباه مثبت تشخیص داده است. FNتعداد رکوردهایی است که دسته واقعی آنها مثبت بوده و الگوریتم دسته‌بندی دسته آنها را به اشتباه منفی تشخیص داده است.

ارزیابی روش پیشنهادی

ر این مطالعه از۷۰ درصد داده‌ها برای مدل‌سازی و۳۰ درصد به عنوان تست استفاده گردید. در این پژوهش الگوریتم‌های بیزساده، درخت‌تصمیم و شبکه‌عصبیRBF و همچنین روش پیشنهادی مبتنی بر درخت‌تصمیم و بیز ساده مدل‌سازی و مقایسه گردید.

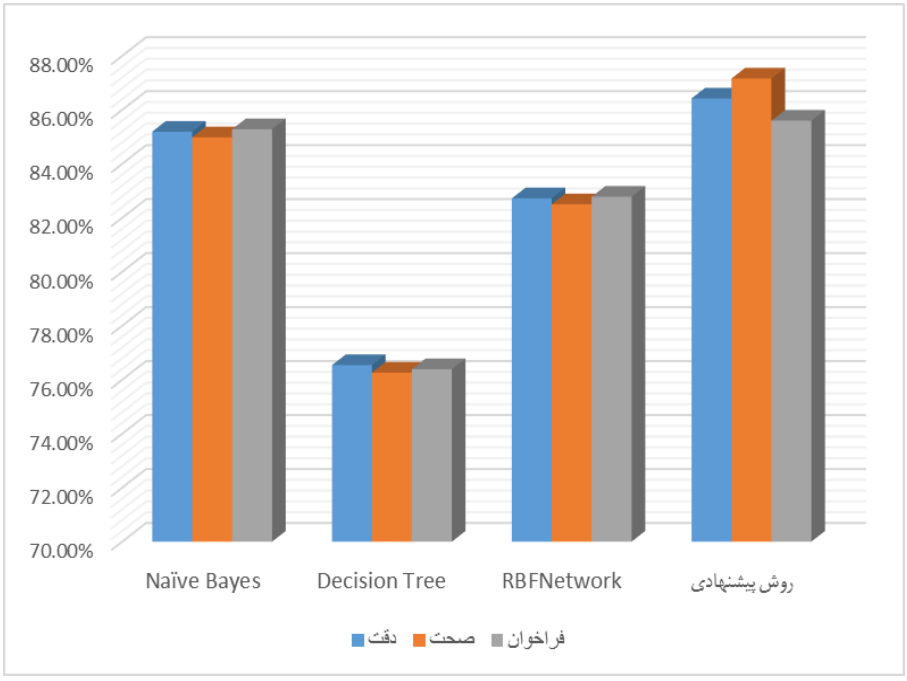
نتایج

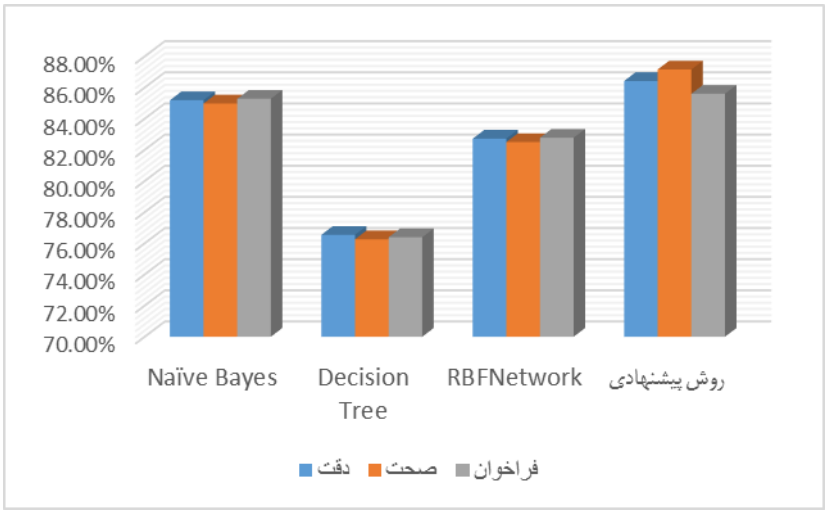
ددر جدول۶ و شکل۲ نتایج سه معیار دقت، صحت و فراخوان در روش‌های مختلف مقایسه شده است که روش پیشنهادی نسبت به سایر روش‌ها دارای عملکرد بهتری است.

جدول ۶: نتایج محاسبه بهره اطلاعاتی

نام الگوریتم	دقت	صحت	فراخوان
Naïve Bayes	٪۸۵,۱۹	٪۸۴,۹۸	٪۸۵,۲۸
Decision Tree	٪۷۶,۵۴	٪۷۶,۲۶	٪۷۶,۳۹
ششبکه‌عصبیRBF Netword	٪۸۲,۷۲	٪۸۲,۵۰	٪۸۲,۷۸
روش پیشنهادی	٪۸۶,۴۲	٪۸۷,۱۶	٪۸۵,۶۰

ISSN: 2745-3673
http://www.Rcsj.ir





شکل۲: مقایسه روش پیشنهادی با سایر الگوریتم‌ها

نتیجه‌گیری

در این مقاله روشی جهت بهبود عملکرد بیزساده در رفع مشکل متغیرهای وابسطه با استفاده از اضافه کردن یک فاز در مرحله پیش‌پردازش داده‌ها جهت کاهش تعداد متغیرها با استفاده از درخت‌تصمیم برای پیش‌بینی وضعیت بیماران قلبی ارائه گردید. نتایج بدست آمده نشان می‌دهد این روش علاوه‌بر کاهش تعداد متغیرها دارای دقت۸۶,۴۲٪ است که کارایی قابل قبولی نسبت به روش‌های به کار رفته در مطالعات پیشین دارد. در Miranda [19] و همکاران، یک طبقه‌بندی مبتنی بر بیزساده ارائه دادند که می‌تواند بیماری‌های قلبی و سطح ریسک در بزرگسالان را شناسایی کند. متغیر هدف شامل سه سطح ریسک۱ که حداقل یکی از موارد چربی، دیابت، عروق کرونر بالاتر از حد عادی است، سطح ریسک۲ که حداقل دو ویژگی بالاتر از حد نرمال استاندارد و سطح ریسک۳ که هر سه ویژگی بالاتر از نرمال است. در این مدل مجموعه‌داده بکار برده شده شامل۳۸ متغیر که با کاهش ابعاد به صورتBackward Selection مدل‌سازی با۱۶ متغیر صورت گرفت که دارای دقت بیشتر از۸۰ درصد برای پیش‌بینی هر سه سطح خطر است. در[20] Chen و همکاران، در این پژوهش روش‌های بیزساده، Kنزدیک‌ترین همسایه، ماشین بردارپشتیبان و شبکه عصبی مدل‌سازی و مقایسه شدند که الگوریتم بیز ساده با دقت۸۳ درصد نسبت به الگوریتم‌های مذکور کارایی بالاتری در پیش‌بینی بیماران عروق کرونر داشت. نتایج این مقاله نشان می‌دهد که بیماری‌های مرتبط(مثلا

ISSN: 2745-3673
http://www.Rcsj.ir

دیابت و فشارخون بالا)، سن و وضعیت سیگار کشیدن، مهم ترین عوامل پیش‌بینی بیماری عروق کرونر هستند در حالی که پلی مورفیسم های ژنتیکی باعث تأثیرات پیچیده‌تری می‌شوند. لنگری‌زاده و همکاران در[21] با هدف مرور نظام‌مند کاربرد شبکه بیزین‌ساده در بیماری‌ها بر روی ۹۰ مقاله پرداختند، مطالعات این محققین نشان داد در۹۲ درصد از پژوهش‌ها که از سال۲۰۰۷ تا۲۰۱۷ صورت گرفته است، الگوریتم بیزساده کارایی بهتری داشته و پیشنهاد گردیده است. این مطالعه بر روی مقالاتی با موضوعات نارسایی بطن راست، آسم، بیماری‌های قلبی و عروق کرونر، انواع سرطان، غده سمی، HIV، فشارخون و پارکینسون صورت گرفته است.

نتایج بدست آمده نشان می‌دهد روش پیشنهادی در مقایسه با سایر مطالعات صورت گرفته از نظر دقت، صحت و فراخوان عملکرد بهتری داشته و با توجه به رعایت فرض استقلال در بیزساده علاوه بر کاهش ابعاد تا حد مناسبی پارامترهای مختلف این الگوریتم را بهبود داده است. هر چند اضافه شدن یک فاز درخت‌تصمیم باعث کاهش سرعت مدل‌سازی می‌گردد ولی این روش نیازی به طراحی کامل درخت-تصمیم نداشته و در مراحل اولیه می‌توان ویژگی‌های مناسب را جداسازی کرد که باعث افزایش سرعت و هزینه محاسبات کمتر نسبت به سایر مدل‌های پیشین می‌گردد. برای تحقیقات آتی پیشنهاد می‌شود، روش پیشنهادی با استفاده از روش‌هایBagging یاBoosting تقویت شود.

ISSN: 2745-3673
http://www.Rcsj.ir

منابع و مراجع

[1] Chen, A. H., Huang, S. Y., Hong, P. S., Cheng, C. H., & Lin, E. J. HDPS: Heart disease prediction system. *Computing in Cardiology. 2011*; IEEE : 557-560.
[2] 2-Amani F, Kazemnejad A, Habibi R, Hajizadeh E.Pattern of mortality trend in Iran during 1970-2009. JGorgan Univ Med Sci. 2011;12(4):85-90. Persian
[3] 3-Kohi, F., Salehinia, H., Mohammadian-Hafshejani, A. Trends in mortality from cardiovascular disease in Iran from 2006-2010. Journal of Sabzevar University of Medical Sciences. 2015; 22(4): 630-638. Persian
[4] 4- Lan K, et al, Mehta R, Govrdhan A, A Survey of Data Mining and Deep Learning in Bioinformatics, Journal of medical systems. 2018; 42(8): 139.
[5] 5.SAFDARI R, et al. A model for predicting myocardial infarction using data mining techniques. Iranian journal of medical informatics. 2013; 2(4):1-6.
[6] 6-Mahmoodi MS. Designing a Heart Disease prediction System using Support Vector Machine. Journal of Health and Biomedical Informatics. 2017; 4(1): 1-10. Persian
[7] 7-Sabbagh Gol H. Detection of Coronary Artery Disease Using C4.5 Decision Tree. Journal of Health and Biomedical Informatics. 2017; 3(4): 287-299. Persian
[8] 8-Suganya S, Tamije Selvy P. A proficient heart disease prediction method using fuzzy-CART algorithm. International Journal of Scientific Engineering and Applied Science. 2016; 2(1): 1-6.
[9] 9- Srinivas K, Kavihta Rani B, Govrdhan A, Applications of Data Mining Techniques in Healthcare and Prediction of Heart Attacks, International Journal on Computer Science and Engineering. 2010; 2(2): 250-255.
[10] 10- Hosseinkhani Z, et al. Distribution of Cardiovascular Disease (CVD) Risk Factors in Adults in Qazvin City. Medical Journal of Mashad University of Medical Sciences. 2013; 56(5): 275-282. Persian
[11] 11- UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science. 2017
[12] 12- Gholami M. Data Mining for all. 1th ed. St. Tehran: Naghoos; 2017. Persian.
[13] 13- Han J, Pey J, Kamber M . Data mining: concepts and techniques. . Elsevier, 2011
[14] 14-Venkatalakshmi. B, Shivsankar, M .Heart Disease Diagnosis Using Predictive Data mining, International Journal of Innovative Research in Science, Engineering and Technology. 2014;. 3(3): 585-600.
[15] 15- Sayyed Muzammil A, Tuteja R. Data Mining Techniques, International Journal of Computer Science and Mobile Computing. 2014, 5(4): 879-883.
[16] 16-SHAIKH A, MAHOTO N, KHUHAWAR F, MEMON M .Performance Evaluation of Classification Methods for Heart Disease Dataset, SINDH UNIVERSITY RESEARCH JOURNAL (SCIENCE SERIES). 2015; 47(3): 389-394.
[17] 17- Kefelegn S, Kamat P, Prediction and Analysis of Liver Disorder Diseases by using Data Mining Technique: Survey, International Journal of Pure and Applied Mathematics. 2018; 118(9): 765-770.
[18] 18- Goyal A, Mehta R, Govrdhan A, Performance Comparison of Naïve Bayes and J48 Classification Algorithms, International Journal of Applied Engineering Research. 2012; 7(11): 1-5.
[19] 19- Miranda E, Irwansyah E, Amelga AY, Maribondang MM, Salim M. Detection of Cardiovascular Disease Risk's Level for Adults Using Naive Bayes Classifier. Healthc Inform Res. 2016; 22(3): 196-205.
[20] 20- Chen Q, Li G, Leong TY, Heng CK. Predicting coronary artery disease with medical profile and gene polymorphisms data. Stud Health Technol Inform. 2007; 129(2):1219-24.
[21] 21- Langarizadeh M, Moghbeli F, Alibeik MR. Using Naïve Bayesian Network in Predicting Diseases: A Systematic Review. Journal of Health and Biomedical Informatics. 2017; 3(4): 327-319. Persian