مقاله پژوهشی

مجله انفورماتیک سلامت و زیست پزشکی
مرکز تحقیقات انفورماتیک پزشکی
دوره سوم شماره چهارم. ص ۲۹۹-۲۸۷. ۱۳۹۵

تشخیص بیماری عروق کرونر قلبی با استفاده از درخت تصمیم C4.5

حامد صباغ گل[1]*

• **دریافت مقاله:** ۹۵/۹/۲۵ • **پذیرش مقاله:** ۹۵/۱۲/۱۷

مقدمه: یکی از شایع‌ترین بیماری‌ها و علل مرگ و میر در دنیای امروز بیماری‌های قلبی است. استفاده از تکنیک‌های داده‌کاوی برای ایجاد مدل‌های پیشگویی کننده، جهت شناسایی افراد در معرض خطر برای کاهش عوارض ناشی از بیماری بسیار کمک کننده است. در این پژوهش با استفاده از درخت تصمیم C4.5 به روش‌های پیشگیری و تشخیص این بیماری پرداخته می‌شود.

روش: این پژوهش از نوع کاربردی و توصیفی می‌باشد. در این پژوهش از داده‌های استاندارد UCI و مجموعه داده Cleveland استفاده نمودیم. این پایگاه داده شامل ۲۹۷ رکورد می‌باشد. تجزیه و تحلیل به کمک نرم‌افزار Weka با به‌کارگیری متدولوژی CRISP3 انجام شده است. در بخش مدل‌سازی درخت تصمیم C4.5 با به‌کارگیری متغیرهای ورودی و تعیین متغیر هدف ایجاد شد.

نتایج: با توجه به مدل استفاده شده مشخص شد که به ترتیب متغیرهای سطح بالای کلسترول، جنسیت، سن بالا، بالا بودن ماکزیمم ضربان قلب، اسکن تالیوم بالاتر از ۳ و نوار قلب غیرنرمال بیشترین تأثیر را در ابتلا به بیماری عروق کرونر قلبی دارا هستند. همچنین به کمک درخت تصمیم ایجاد شده، قوانینی استخراج شده است که می‌تواند به عنوان الگویی در جهت پیشگویی احتمال ابتلا افراد به بیماری عروق کرونر قلبی استفاده شود. صحت مدل ایجاد شده با استفاده از درخت تصمیم بیش از ۸۰ درصد بوده است.

نتیجه‌گیری: با توجه به محاسبات انجام شده، نرخ دسته‌بندی برابر با ۷۲/۶٪ و دقت الگوریتم C4.5 برابر با ۸۰/۲٪ به‌دست آمد که در مقایسه با نتایج مطالعات انجام شده در حوزه داده‌کاوی بیماری قلبی، دقت به‌دست آمده الگوریتم پیشنهادی قابل قبول است.

کلید واژه‌ها: داده‌کاوی، بیماری عروق کرونر قلبی، درخت تصمیم C4.5

• **ارجاع:** صباغ گل حامد. تشخیص بیماری عروق کرونر قلبی با استفاده از درخت تصمیم C4.5. مجله انفورماتیک سلامت و زیست پزشکی ۱۳۹۵؛ ۳(۴): ۲۹۹-۲۸۷.

۱. کارشناسی ارشد مهندسی کامپیوتر گرایش نرم‌افزار، مربی، عضو هیئت علمی گروه کامپیوتر، دانشگاه پیام نور بیرجند، بیرجند ایران

* **نویسنده مسئول:** بیرجند، انتهای بلوار شهید آوینی، دانشگاه پیام نور خراسان جنوبی، مرکز بیرجند

• **شماره تماس:** ۰۵۶۳۲۲۰۲۰۲۵

• **Email:** Sabbagh.h@pnu.ac.ir

مقدمه

بیماری قلبی- عروقی یا بیماری قلبی دسته‌ای از بیماری‌ها است که در قلب یا رگ‌ها (سرخرگ‌ها، مویرگ‌ها و سیاهرگ‌ها) رخ می‌دهد. بیماری قلبی-عروقی به هر گونه بیماری که دستگاه گردش خون را تحت تأثیر قرار دهد اشاره دارد که شامل بیماری‌های قلبی، بیماری‌های عروقی مغز و کلیه و بیماری‌های شریانی می‌شود [۱]. بیماری‌های قلبی یکی از عوامل اصلی مرگ‌ومیر در دنیا، به خصوص ایران است و بهترین درمان آن تشخیص به موقع و پیشگیری آن است. در ایران، سکته قلبی اولین علت مرگ افراد بالاتر از ۳۵ سال می‌باشد. بیماری قلبی که معمولاً از آن به عنوان بیماری شریان‌های اکلیلی (Coronary Artery Disease)CAD نام برده می‌شود؛ واژه‌ای با دامنه وسیع است که به هر نوع شرایطی که قلب را تحت تأثیر قرار می‌دهد، اطلاق می‌شود. CAD بیماری مزمنی است که طی آن شریان اکلیلی به‌تدریج سفت و باریک می‌شود و همچنین رایج‌ترین بیماری قلبی-عروقی که موجب حملات قلبی می‌شود. به عنوان مثال طبق آمار سالانه ۵۰۰۰۰۰ حملات قلبی منجر به مرگ در ایالات متحده رخ می‌دهد که آمار قابل تأملی برای یک کشور توسعه‌یافته است. در حالی‌که بیشتر مردم مبتلا به بیماری قلبی نشانه‌هایی همچون درد قفسه سینه و خستگی دارند، ولی حدود ۵۰ درصدشان تا زمان حمله قلبی هیچ نشانه‌ای ندارند [۲].

حجم داده‌های پزشکی روز به روز در حال افزایش است و پزشکان معمولاً اطلاعات ارزشمندی را در خصوص بیماری‌ها و ارتباط آن‌ها با دیگر عوامل ایجاد کننده بیماری‌ها به دست می‌آورند [۳]؛ اما این مجموعه داده‌های خام به خودی خود ارزشی ندارند، برای معنی بخشیدن به این داده‌ها باید آن‌ها را تحلیل و تبدیل به اطلاعات یا بهتر از آن‌ها دانش کرد [۴]. با توجه به شیوع بیماری‌های قلبی –عروقی در سراسر جهان، استفاده از روش‌های جدید در تحقیقات زیست پزشکی بسیار مورد توجه قرار گرفته است. داده‌کاوی ابزاری است که برای حصول به چنین دانشی ما را یاری می‌کند. یکی از زمینه‌های پرکاربرد داده‌کاوی در علم پزشکی است؛ استفاده از تکنیک‌های داده‌کاوی برای ایجاد مدل‌های پیش‌گویی کننده، جهت شناسایی افراد در معرض خطر برای کاهش عوارض ناشی از بیماری بسیار کمک کننده است[۳].

محیط مراقبت سلامت غنی از اطلاعات و ضعیف از دانش است [۵]. داده‌کاوی، پتانسیل خوبی برای ایجاد یک محیط غنی از دانش دارد که می‌توانند کمک قابل توجهی به کیفیت تصمیمات بالینی نمایند [۲].

داده‌کاوی پزشکی دارای پتانسیل زیادی برای کشف الگوهای پنهان موجود در داده‌ها داراست که این الگوها می‌توانند برای تشخیص‌های بالینی مورد استفاده قرار گیرد [۶]. امروزه استفاده از روش‌های متنوع داده‌کاوی برای شناسایی الگوها و ارتباطات میان متغیرهای مختلف در تولید مدل‌های پیش‌بینی‌کننده در علوم پزشکی بسیار مورد توجه قرار گرفته است [۷]. کاربرد روش‌های داده‌کاوی در حوزه‌های مختلف پزشکی مانند تشخیص، پیش‌گویی و حتی درمان به اثبات رسیده است [۸].

یکی از عملکردهای پیش‌گویانه در داده‌کاوی، دسته‌بندی است [۹]. از مهم‌ترین روش‌های رایج دسته‌بندی درخت تصمیم می‌باشد و از میان الگوریتم‌های مورد استفاده در ساخت درخت تصمیم، مهم‌ترین آن‌ها الگوریتم C4.5 است [۱۰].

Colombet و همکاران در مطالعه خود نشان داده‌اند که درختان تصمیم نسبت به مدل رگرسیون لجستیک، قابلیت بیشتری در جهت پیشگویی امکان ابتلا افراد به بیماری‌های قلبی دارا هستند [۱۱]. در مطالعه انجام شده توسط Bellaachia و Guven جهت پیش‌گویی امکان زنده ماندن بیماران مبتلا به سرطان سینه الگوریتم C4.5 با میزان دقت ۷/۶۸ درصد به عنوان بهترین مدل پیش‌گویی کننده معرفی شده است [۱۲]. در مطالعه انجام شده توسط Stärk و Pfeiffer بیان شده است که مدل‌های درخت تصمیم برای ID3، C4.5، CART و CHAID از جمله انجام تحلیل‌های اکتشافی در پایگاه‌های داده بزرگ نسبت به مدل‌های دیگر موفق‌تر عمل می‌کنند [۱۳]. Ordonze و همکاران از الگوریتم درخت تصمیمC4.5 و الگوریتم قوانین همبستگی با استفاده از ۲۵ ریسک فاکتور جهت پیش‌گویی بیماری قلبی استفاده کردند و به این نتیجه رسیدند که قوانین همبستگی عموماً قواعد پیش‌بینی ساده‌تری نسبت به درختان تصمیم ایجاد می‌کند [۱۴].

هدف اصلی ما در این پژوهش استفاده از درخت تصمیمC4.5 بر روی مجموعه داده استاندارد بیماران قلبی Cleveland برای پیش‌بینی و تشخیص بیماری عروق کرونر قلبی و ارائه یک مدل به منظور انجام غربالگری پیشگیرانه جهت کاهش عوارض ناشی از بیماری است که در ادامه به کم و کیف آن می‌پردازیم.

می‌توان با استفاده از نتایج به دست آمده از این پژوهش، پیشنهاداتی به متخصص بالینی جهت تشخیص سریع‌تر و در

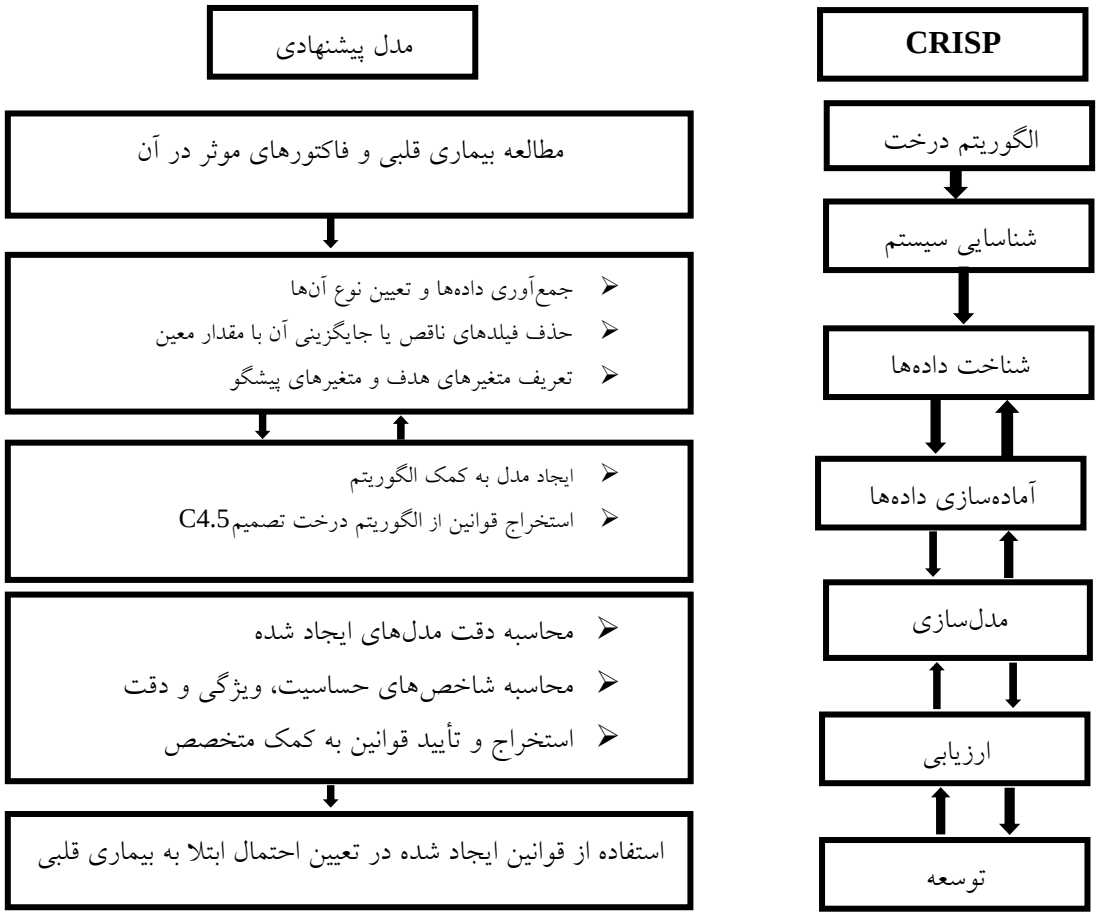
صورت لزوم کاهش آسیب‌های ناشی از روش‌های تهاجمی تشخیص بیماری‌های عروق کرونر ارائه نمود.

درخت تصمیم

درخت تصمیم یکی از روش‌های قوی و متداول برای دسته‌بندی و پیش‌بینی است. در واقع درخت‌های تصمیم بالا به پایین رایج‌ترین تکنیک دسته‌بندی هستند و از مهم‌ترین دلایل رایج بودنشان می‌توان شفاف بودن، قابل فهم، انعطاف‌پذیری و پردازش نسبتاً سریع ساختار آن‌ها را نام برد. پیش‌بینی به‌دست آمده از درخت در قالب یک سری قواعد توضیح داده می‌شود. در این درخت هر گره داخلی شامل سؤالی بر مبنای یک صفت مشخص و یک فرزند برای هر پاسخ ممکن بوده و هر برگ با یکی از کلاس‌های ممکن برچسب‌گذاری می‌شود [۱۵]. درخت تصمیم جهت دسته‌بندی یک نمونه با شروع از ریشه مسیری را بر اساس سؤالات مطرح شده در گره‌های داخلی و پاسخ‌های آن دنبال می‌کند تا زمانی که به یک برگ برسد در نهایت برچسب مربوطه کلاس نمونه موردنظر خواهد بود. اغلب الگوریتم‌های یادگیری درخت تصمیم بر پایه یک عمل جستجوی بالا به پایین عمل می‌کنند [۱۶].

روش

روش‌های متعددی برای اجرای پروژه‌های داده‌کاوی وجود دارد که یکی از روش‌های قدرتمند در این زمینه متدولوژی کریسپ (CRISP) می‌باشد [۱۷]. این پژوهش نیز بر اساس این متدولوژی تنظیم شده است. شکل (۱) در ادامه به بررسی هر یک از این مراحل در جهت رسیدن به مدلی برای پیش‌گویی احتمال ابتلا به بیماری قلبی می‌پردازد.



شکل ۱: گام‌های متدولوژی کریسپ و چارچوب استفاده شده در این مطالعه

الف) شناخت سیستم

به کارگیری موفق داده‌کاوی مستلزم شناخت حوزه‌ای است که قرار است داده‌کاوی در آن به کار برده شود و علاوه بر آن

شناخت کافی از روش‌ها و ابزارهای داده‌کاوی نیز لازم است. به‌طورکلی تیم داده‌کاوی بایستی دانش کافی در حوزه‌ای که قرار است بررسی شود داشته باشند. در گام اول پژوهش با مشورت پزشک متخصص قلب و عروق و نیز با مطالعه بر روی بیماری عروق کرونر قلبی و تعیین فاکتورهای مؤثر در ابتلا و همچنین روش‌های تشخیصی و درمانی و روش‌های پیشگیری از ابتلا به بیماری، سعی در شناخت کافی حوزه مورد بررسی می‌باشد.

ب) آماده‌سازی داده‌ها

در این پژوهش از مجموعه داده تشخیص بیماری قلبی Cleveland مخزن داده‌ای سایت مرجع (archive.ics.uci.edu/ml/datasets/Heart+Diseas) UCI استفاده شده است. علائم زیادی از بیماری قلبی وجود دارد، یافتن الگوهایی از داده بیماری قلبی در تشخیص دلایل آتی این بیماری کمک می‌کند. پایگاه داده بیماری قلبی توسط مرکز پزشکی Cleveland Clinic Foundation، Long Beach و V.A در سال ۱۹۹۸ ایجاد شده است [۱۶].

پایگاه داده شامل ۳۰۳ نمونه که در برگیرنده ۲۹۷ نمونه کامل و شش نمونه با مقادیر از دست رفته است. این پایگاه داده ۷۶ صفت خام دارد در حالی‌که همه آزمایش‌ها فقط بر روی ۱۴ صفت از آن‌ها انجام شده است؛ بنابراین، این پایگاه داده شامل ۱۳ علامت بیماری و یک صفت تشخیص است که فیلد هدف به وجود بیماری قلبی بر اساس علائم موجود در بیمار اشاره دارد که یک مقدار عددی ۰ (کمتر از ۵۰٪ تنگی عروق) یا ۱ (بیشتر از ۵۰٪ تنگی عروق) است که در ادامه مفهوم هر کدام از علائم بیان می‌شود.

در این مرحله داده‌هایی که در حال حاضر در دسترس هستند و داده‌هایی که برای ساخت مدل نیاز بود، تعیین شدند. برای شروع پژوهش از طریق مطالعات کتابخانه‌ای و مشاوره بالینی و بر اساس داده‌های مجموعه داده Cleveland، فاکتورهایی که بیشترین تأثیر را در ابتلای افراد به بیماری دارا بودند تعیین شد که در جدول ۱ مشاهده می‌شوند.

همچنین متغیرهای تعیین شده برای ایجاد مدل به دو دسته متغیرهای هدف و متغیرهای پیشگو دسته‌بندی شدند که متغیر هدف ابتلا یا عدم ابتلا و سایر متغیرها به عنوان متغیر پیشگو مورد استفاده قرار گرفتند.

طبق جدول ۱ در حالت معمولی نتایج یک اسکن تالیوم به شرح زیر می‌باشد:

- اگر نتایج اسکن در حالت استراحت و پس از ورزش طبیعی باشند، نتیجه گرفته می‌شود که میزان جریان خون سرخرگ‌های کرونری به عضلات قلب کافی است.
- اگر جریان خون در حالت استراحت طبیعی باشد، ولی در طول ورزش طبیعی نباشد) یعنی نقص در خون‌رسانی یا پرفوزیون)، نتیجه گرفته می‌شود که قسمتی از عضلات قلب در طی فعالیت شدید، خون کافی دریافت نمی‌کند و احتمالاً یکی از سرخرگ‌های کرونری مسدود شده است.
- اگر جریان خون قسمتی از عضلات قلب، در حالت استراحت کاهش یابد و در طول ورزش بدتر هم بشود، نتیجه گرفته می‌شود که بخشی از عضلات قلب، در تمام اوقات با کاهش خون‌رسانی مواجه است.
- اگر هیچ‌گونه تالیومی در هر دو حالت استراحت و پس از ورزش در عضله قلب مشاهده نشود (که اصطلاحاً نقص تثبیت شده نامیده می‌شود Fixed-Defect)، نتیجه گرفته می‌شود که احتمالاً قبلاً یک حمله قلبی اتفاق افتاده و قسمتی از بافت قلب مرده است و به جای آن بافت اسکار تشکیل شده است [۱۸].

جدول ۱: صفات اطلاعاتی مورد استفاده

ردیف	نام صفت	معادل به کاررفته	نام توضیحات	مجموعه مقادیر	درصد فراوانی
۱	Age	Age	سن بیمار	۷۶-۲۹	کمتر از ۵۳ سال ۵۰٪/۲۳ بیشتر از ۵۳ سال ۴۹٪/۷۷
۲	Sex	Sex	جنسیت بیمار	۱ مرد ۰ زن	۰ ۳۶٪/۹۹ زن ۱ ۶۳٪/۰۱ مرد
۳	Chest pain type	cp	بیان کننده درد قفسه سینه که شامل ۴ مقدار است	۱ آنژین صدری معمولی ۲ درد قلبی ۳ بدون درد ۴ بدون علامت	۱ ۲۲٪ ۲ ۲۱٪/۴۶ ۳ ۳۵٪/۱۶ ۴ ۳۳٪/۷۹
۴	Resting blood pressure	trestbps	فشارخون در زمان استراحت	۱۹۲-۹۴	≥۱۴۰ ۷۰٪/۳۲ <۱۴۰ ۲۹٪/۶۸
۵	Serum cholesterol	chol	کلسترول (چربی بد خون)	۵۶۴-۱۲۶	مطلوب (زیر ۲۰۰) ۱۶٪/۸۹ حاشیه‌ای (بین ۲۰۰-۲۳۹) ۳۲٪/۸۸ خطرناک (بالای ۲۴۰) ۵۰٪/۲۳
۶	Fasting blood sugar	fbs	قند خون ناشتا	۱ دارد ۰ ندارد	۱ ۸۷٪/۶۷ ۰ ۱۲٪/۳۳
۷	Resting electrocardiographic results	restecg	نتایج نوار قلب در حال استراحت که شامل ۳ مقدار نرمال، موج غیر قلبی و نشان‌دهنده‌ی افزایش مقطعی یا احتمالی ضخامت بطن چپ است.	۰ نرمال ۱ موج غیر قلبی ۲ موج افزایش مقطعی	۰ ۵۳٪/۹ ۱ ۰٪/۵ ۲ ۴۵٪/۶
۸	Maximum heart rate achieved	thalach	ماکزیمم ضربان قلب به دست آمده	۲۰۲-۸۸	>۱۰۰ ۹۸٪/۶۳ <۱۰۰ ۱٪/۳۷
۹	Exercise induced angina	exang	آنژین ناشی از ورزش که شامل مقادیر بله و خیر است.	۰ خیر ۱ بله	۰ ۷۸٪ ۱ ۲۲٪
۱۰	St depression induced by exercise relative	oldpeak	افسردگی ایجاد شده St موقع تست ورزش	۴-۰/۲	≥۰/۸ ۱٪/۶۷ >۰/۸ ۳۲٪/۹
۱۱	The slope of the peak exercise ST segment	slope	بیان کننده شیب قطعه st در زمان حداکثر ورزش که شامل مقادیر: بالا رفتن، صاف و پایین آمدن قطعه st است.	۱ بالا رفتن ۲ صاف ۳ پایین آمدن	۱ ۸۵٪/۵ ۲ ۳۶٪ ۳ ۵٪/۵
۱۲	Number of major vessels colored by fluoroscopy	Ca	این صفت بیانگر تعداد رگ‌هایی که در فلوروسکوپی دیده می‌شود.	۳-۰	۰ ۷۱٪/۲ ۱ ۱۸٪/۳ ۲ ۶٪/۵ ۳ ۳٪
۱۳	Thal	thal	اسکن تالیوم است که شامل ۳ مقدار ضایعه ثابت، نرمال و ضایعه قابل برگشت است.	۳ ضایعه (نقص) ثابت ۶ ضایعه (نقص) نرمال ۷ ضایعه (نقص) قابل برگشت	۳ ۶۹٪/۴ ۶ ۴٪/۲ ۷ ۲۶٪/۴
۱۴	Num (صفت تشخیص)	num	تشخیص بیماری قلبی (وضعیت آنژیوگرافی)	٪۵۰:۰ > تنگی قطر ٪۵۰:۱> تنگی قطر	۰ ۲۵٪ ۱ ۷۵٪

ج) مدل‌سازی

روش‌های داده‌کاوی متنوعی برای مدل‌سازی وجود دارد. در این مرحله با استفاده از الگوریتم درخت تصمیم C4.5 به ارائه مدل پیش‌گویانه پرداخته شد. مدل‌سازی با استفاده از نرم‌افزار Weka 3.6 انجام شد. در این پروژه از ۲۲۰ رکود معتبر آن استفاده شده است. در ادامه الگوریتم درخت تصمیم C4.5 با به کارگیری متغیرهای ورودی و تعیین متغیر هدف ایجاد شد. برای ساخت مدل درخت تصمیم متغیرهای سن بیمار، جنسیت، درد قفسه سینه، فشارخون، کلسترول، قند خون، نتایج نوار قلب، ماکزیمم ضربان قلب، آنژین ناشی از ورزش، افسردگی st ایجاد شده موقع تست ورزش، تعداد رگ‌های فلوروسکوپی، اسکن تالیوم و شیب قطعه st به عنوان متغیرهای پیشگو تعیین شد و متغیر ابتلا یا عدم ابتلا به بیماری نیز به عنوان متغیر هدف تعیین گردید. در مرحله بعد داده‌ها به دو دسته آموزش (۸۰ درصد) و آزمون (۲۰ درصد) تقسیم شدند. داده‌های بخش آموزش مدل را می‌سازند و داده‌های بخش آزمون مدل ایجاد

شده را مورد ارزیابی قرار می‌دهند. یک درخت تصمیم ترکیب تعدادی استلزام منطقی (قانون اگر-آنگاه) است. معمولاً مجموعه قوانین استخراج شده از درخت تصمیم، مهم‌ترین اطلاعاتی است که از آن‌ها به دست می‌آید. در مدل ایجاد شده در این نرم‌افزار به منظور تقسیم شاخص‌ها از شاخص جینی استفاده شده است [۱۹]. دلیل انتخاب این مدل نیز به این جهت بود که با محاسبه شاخص‌های موردنظر دارای بالاترین دقت در بین مدل‌های اجرا شده بود. نحوه محاسبه شاخص‌ها در بخش تجزیه‌وتحلیل درخت تصمیم ارائه شده است. در جدول ۲ تعدادی از قوانین ایجاد شده براساس مشاوره بالینی و نتایج حاصله توسط مدل C4.5 بیان شده است.

جدول ۲: تعدادی از قوانین ایجاد شده توسط الگوریتم C4.5

ردیف	قوانین	احتمال ابتلا
۱	اگر اسکن تالیوم نرمال و بر اساس فلوروسکوپی دارای گرفتگی عروق قلب نباشد، آنگاه احتمال ابتلا کمتر از ۵۰٪ تنگی عروق برابر است با	۷۴٪
۲	اگر اسکن تالیوم نرمال و براساس فلوروسکوپی دارای گرفتگی عروق قلب باشد و جنسیت زن و قطعه st با شیب رو به بالا در تست ورزش باشد آنگاه احتمال ابتلا کمتر از ۵۰٪ تنگی عروق برابر است با	۵۶٪
۳	اگر اسکن تالیوم نرمال و براساس فلوروسکوپی دارای گرفتگی عروق قلب باشد و جنسیت زن و قطعه st با شیب رو به پایین یا مسطح در تست ورزش باشد و ماکزیمم ضربان قلب کوچک‌تر یا مساوی ۱۶۰ باشد آنگاه احتمال ابتلا کمتر از ۵۰٪ تنگی عروق برابر است با	۵٪
۴	اگر اسکن تالیوم نرمال و براساس فلوروسکوپی دارای گرفتگی عروق قلب باشد و جنسیت زن و قطعه st با شیب رو به پایین یا مسطح در تست ورزش و ماکزیمم ضربان قلب بزرگ‌تر از ۱۶۰ باشد آنگاه احتمال ابتلا بیشتر از ۵۰٪ تنگی عروق برابر است با	۴٪
۵	اگر اسکن تالیوم نرمال و براساس فلوروسکوپی دارای گرفتگی عروق قلب باشد و جنسیت مرد و درد قفسه سینه دارای علامت باشد آنگاه احتمال ابتلا بیشتر از ۵۰٪ تنگی عروق برابر است با	۲٪/۲۷
۶	اگر اسکن تالیوم نرمال و بر اساس فلوروسکوپی دارای گرفتگی عروق قلب باشد و جنسیت مرد و درد قفسه سینه بدون علامت و سن کمتر از ۵۵ سال باشد آنگاه احتمال ابتلا کمتر از ۵۰٪ تنگی عروق برابر است با	۵٪
۷	اگر اسکن تالیوم نرمال و براساس فلوروسکوپی دارای گرفتگی عروق قلب باشد و جنسیت مرد و درد قفسه سینه بدون علامت و سن بیشتر از ۵۵ سال باشد آنگاه احتمال ابتلا بیشتر از ۵۰٪ تنگی عروق برابر است با	۲٪/۲۷
۸	اگر اسکن تالیوم نقص ثابت یا برگشت‌پذیر باشد و درد قفسه سینه دارای علامت باشد آنگاه احتمال ابتلا کمتر از ۵۰٪ تنگی عروق برابر است با	۱۱٪
۹	اگر اسکن تالیوم نقص ثابت و درد قفسه سینه بدون علامت و نوار قلب در حال استراحت دارای موج نرمال یا غیر قلبی آنگاه احتمال ابتلا کمتر از ۵۰٪ تنگی عروق برابر است با	۰٪/۹
۱۰	اگر اسکن تالیوم نقص ثابت و درد قفسه سینه بدون علامت و نوار قلب در حال استراحت دارای موج نرمال یا غیرقلبی و افسردگی قطعه st ایجاد شده موقع تست ورزش بزرگ‌تر از ۰/۸ باشد آنگاه احتمال ابتلا بیشتر از ۵۰٪ تنگی عروق برابر است با	۲٪/۷۲
۱۱	اگر اسکن تالیوم نقص ثابت و درد قفسه سینه بدون علامت و نوار قلب در حال استراحت دارای موج نرمال یا غیرقلبی و افسردگی قطعه st ایجاد شده موقع تست ورزش کوچک‌تر از ۰/۸ باشد و فشارخون در زمان استراحت بزرگ‌تر از ۱۳۶ باشد آنگاه احتمال ابتلای بیشتر از ۵۰٪ تنگی عروق برابر است با	۰٪/۹
۱۲	اگر اسکن تالیوم نقص ثابت و درد قفسه سینه بدون علامت و نوار قلب درحال استراحت دارای موج نرمال یا غیرقلبی و افسردگی قطعه st ایجاد شده موقع تست ورزش کوچک‌تر از ۰/۸ باشد و فشارخون در زمان استراحت کوچک‌تر یا مساوی از ۱۳۶ باشد آنگاه احتمال ابتلا کمتر از ۵۰٪ تنگی عروق برابر است با	۲٪/۷۲
۱۳	اگر اسکن تالیوم نقص ثابت یا برگشت‌پذیر باشد و درد قفسه سینه بدون علامت و نوار قلب دارای موج افزایش مقطعی و قطعه st با شیب رو به بالا در تست ورزش و فشارخون در زمان استراحت کمتر یا مساوی ۱۴۰ باشد آنگاه احتمال ابتلا بیشتر از ۵۰٪ تنگی عروق برابر است با	۲٪/۷۲
۱۴	اگر اسکن تالیوم نقص ثابت یا برگشت‌پذیر باشد و درد قفسه سینه بدون علامت و نوار قلب دارای موج افزایش مقطعی و قطعه st با شیب رو به بالا در تست ورزش و فشارخون در زمان استراحت بزرگ‌تر از ۱۴۰ باشد آنگاه احتمال ابتلا کمتر از ۵۰٪ تنگی عروق برابر است با	۰٪/۹
۱۵	اگر اسکن تالیوم نقص ثابت یا برگشت‌پذیر باشد و درد قفسه سینه بدون علامت و نوار قلب دارای موج افزایش مقطعی و جنسیت مرد و کلسترول بالاتر از ۲۰۰ باشد آنگاه احتمال ابتلا بیشتر از ۵۰٪ تنگی عروق برابر است با	۶۴٪/۰۵

د) ارزیابی مدل

در این مرحله پس از ایجاد مدل به ارزیابی مدل ایجاد شده می‌پردازیم. برای بررسی صحت مدل داده‌ها به دو دسته آموزش (۸۰ درصد) و تست (۲۰ درصد) تقسیم شدند. داده‌های بخش آموزش مدل را می‌سازند و داده‌های بخش آزمون (تست) مدل ایجاد شده را مورد ارزیابی قرار می‌دهند. جهت ارزیابی مدل‌ها می‌توان از شاخص‌های حساسیت، ویژگی، دقت، ارزش اخباری مثبت و ارزش اخباری منفی استفاده کرد. جهت محاسبه این شاخص‌ها از ماتریس تداخلی ایجاد شده در محیط نرم‌افزار استفاده شد.

در ادامه پژوهش؛ دقت، حساسیت، ویژگی، مقدار پیش‌بینی مثبت و مقدار پیش‌بینی منفی را بررسی و محاسبه می‌کنیم.

دقت: عبارت است از تعداد نمونه‌هایی که به درستی در کلاس مورد نظر تشخیص داده می‌شوند نسبت به کل نمونه‌ها.

$$\text{دقت} = \frac{\text{منفی های واقعی+مثبت های واقعی}}{\text{تعداد کل داده ها}}$$

حساسیت: عبارت است از تعداد نمونه‌هایی که به درستی عدم وجود ناراحتی قلبی را نشان داده نسبت به تعداد کل نمونه‌هایی که واقعاً ناراحتی قلبی ندارند.

$$\text{حساسیت} = \frac{\text{مثبت های واقعی}}{\text{مثبت های واقعی+منفی های کاذب}}$$

ویژگی: تعداد نمونه‌هایی که به درستی وجود بیماری قلبی را نشان داده نسبت به تعداد کل نمونه‌هایی که واقعاً بیماری قلبی دارند.

$$\text{ویژگی} = \frac{\text{منفی های واقعی}}{\text{منفی های واقعی +مثبت های کاذب}}$$

مقدار پیش‌بینی مثبت: تعداد نمونه‌هایی که به درستی عدم وجود بیماری را نشان داده نسبت به تعداد کل نمونه‌هایی که پیش‌بینی شده بیماری ندارند.

$$\text{مقدار پیش‌بینی مثبت} = \frac{\text{مثبت های واقعی}}{\text{مثبت های واقعی+مثبت های کاذب}}$$

مقدار پیش‌بینی منفی: تعداد نمونه‌هایی که به درستی وجود بیماری را نشان داده نسبت به کل نمونه‌هایی که پیش‌بینی شده بیماری دارند.

$$\text{مقدار پیش‌بینی منفی} = \frac{\text{منفی های واقعی}}{\text{منفی های واقعی+منفی های کاذب}}$$

که هرکدام به صورت زیر محاسبه می‌شوند:

$$\text{Classification rate} = \frac{TP+TN}{TP+FN+FP+TN}$$

$$\text{Precision} = \frac{TP}{TP+FP}$$

$$\text{TNR} = \frac{TN}{TN+FN}$$

$$\text{TPR} = \frac{TP}{TP+FP}$$

با توجه به داده‌ها

$$TP=138,\ TN=21,\ FN=26,\ FP=34$$

$$\text{Classification rate} = \frac{138+21}{138+26+34+21} = 0.726$$

$$\text{Precision} = \frac{138}{138+34} = 0.802$$

$$\text{Recall} = \frac{138}{138+26} = 0.841$$

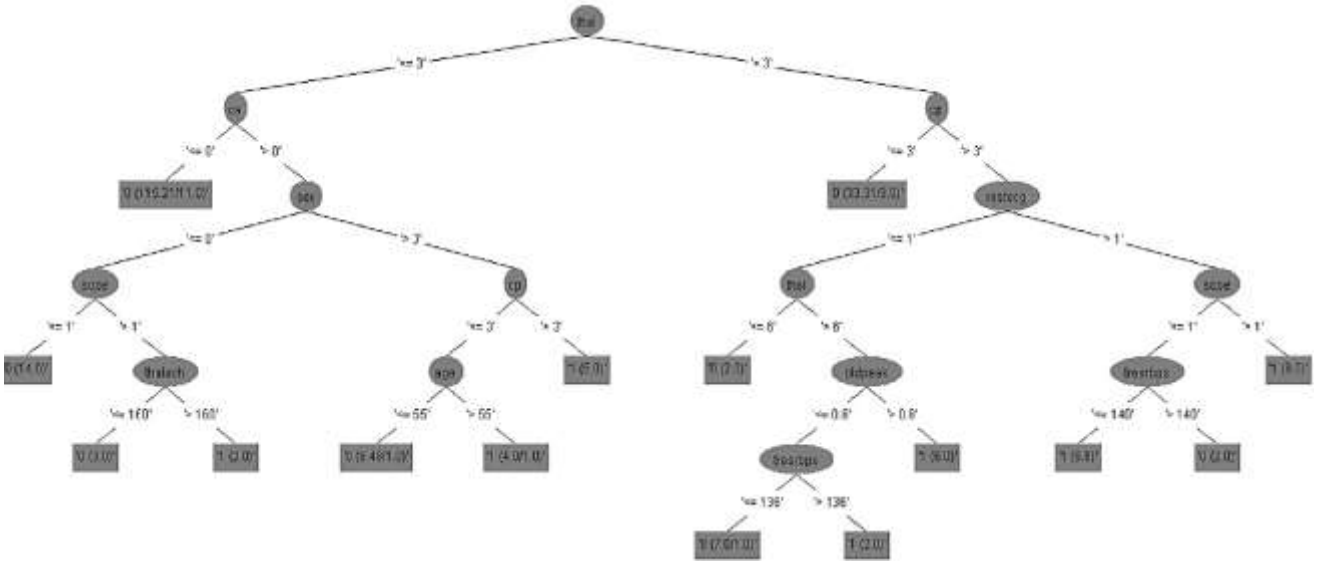
$$\text{TP Rate} = \frac{138}{26+138} = 0.841$$

$$\text{FP Rate} = \frac{34}{34+21} = 0.618$$

نتایج

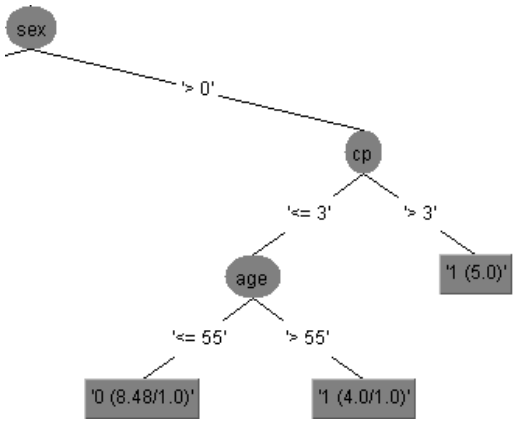
با توجه به مدل‌های استفاده شده مشخص شد که به ترتیب متغیرهای سطح بالای کلسترول، جنسیت، سن بالا، بالا بودن ماکزیمم ضربان قلب، اسکن تالیوم بالاتر از ۳، نوار قلب غیرنرمال و میزان بالای افسردگی قطعه st ایجاد شده در تست ورزش، بیشترین تأثیر را در ابتلا به بیماری عروق کرونر قلبی دارا هستند. به کمک درخت تصمیم ایجاد شده، قوانینی استخراج شده است که می‌تواند به عنوان الگویی در جهت پیشگویی احتمال ابتلا افراد به بیماری قلبی استفاده شود. درخت تصمیم حاصل شده در شکل ۲ نشان داده شده است.

در ادامه به تجزیه و تحلیل درخت تصمیم و بحث در مورد قسمت‌های مختلف آن می‌پردازیم.



شکل ۲: بخشی از درخت تصمیم ایجاد شده

الف)

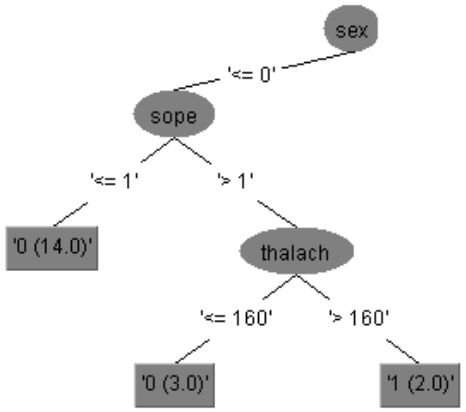


شکل ۳: ساختار درخت تصمیم بر اساس جنسیت مرد

براساس ساختار درخت تصمیم براساس جنسیت مرد، همان‌طور که در شکل ۳ مشاهده می‌شود:

- مردانی که در ناحیه قفسه سینه خود درد بدون علامت احساس می‌کنند احتمال ابتلا بیماری قلبی در آن‌ها مشاهده می‌شود. (درصد بیماران قلبی مرد نسبت به زن بیشتر است).
- مردانی که در ناحیه قفسه سینه خود درد بدون علامت احساس می‌کنند و سن بالای ۵۵ سال دارند احتمال ابتلا بیماری قلبی در آن‌ها مشاهده می‌شود.
- مردانی که در ناحیه قفسه سینه خود درد بدون علامت احساس می‌کنند و سن کمتر از ۵۵ سال دارند احتمال ابتلا بیماری قلبی کمتر در آن‌ها مشاهده می‌شود.

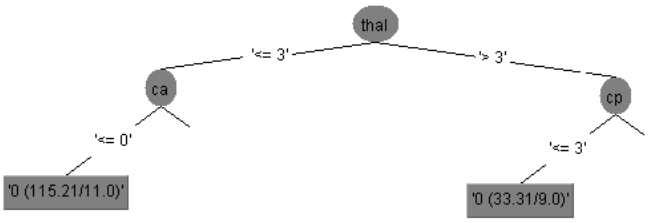
ب)



شکل ۴: ساختار درخت تصمیم بر اساس جنسیت زن

همچنین براساس ساختار درخت تصمیم براساس جنسیت زن، همانطور که در شکل ۴ مشاهده می‌شود:

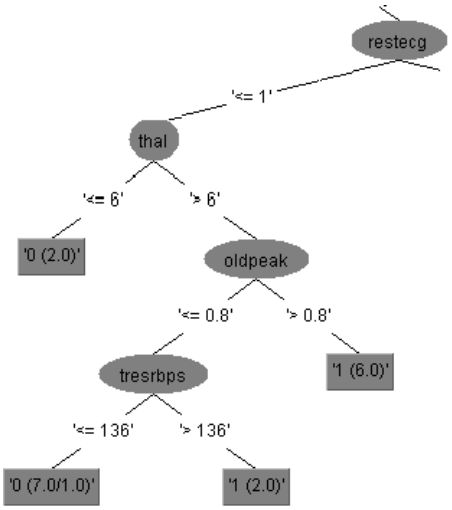
- زنانی که در هنگام ورزش تپش قلب زیادی احساس نمی‌کنند احتمال ابتلا بیماری قلبی در آن‌ها کمتر مشاهده می‌شود.
- زنانی که ماکزیمم ضربان قلبشان بیشتر از ۱۶۰ است احتمال ابتلا بیماری قلبی در آن‌ها بیشتر مشاهده می‌شود.



شکل ۵: ساختار درخت تصمیم بر اساس اسکن تالیوم

شکل ۵ ساختار درخت تصمیم براساس اسکن تالیوم را نشان می‌دهد. همانطور که در شکل مشاهده می‌شود:

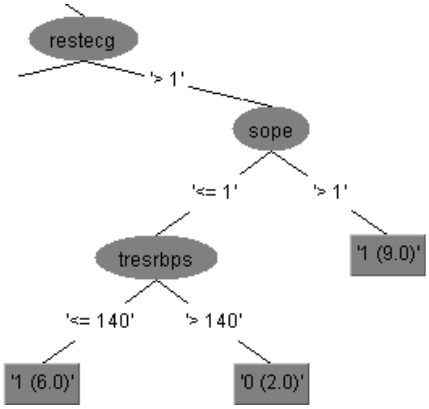
- افرادی که اسکن تالیوم آن‌ها نرمال تشخیص داده شده است و تعداد رگ‌های فلوروسکوپی کمی مشاهده می‌شود احتمال ابتلا بیماری قلبی در آن‌ها کمتر دیده می‌شود.
- افرادی که اسکن تالیوم آن‌ها ضایعه ثابت و یا ضایعه قابل برگشت تشخیص داده شده است و درد قفسه سینه، درد معمولی مشاهده شده احتمال ابتلا بیماری قلبی در آن‌ها کمتر دیده می‌شود.



شکل ۶: ساختار درخت تصمیم بر اساس نتایج نوار قلب (موج نرمال)

همچنین براساس ساختار درخت تصمیم براساس نتایج نوار قلب (موج نرمال)، همانطور که در شکل ۶ مشاهده می‌گردد:

- افرادی که نتایج نوار قلب‌شان موج نرمال داشته‌اند و نتایج اسکن تالیوم آن‌ها نرمال و یا ضایعه ثابت بوده احتمالاً کمتر به بیماری قلبی مبتلا می‌شوند.
- افرادی که نتایج نوار قلب‌شان موج نرمال داشته‌اند و نتایج اسکن تالیوم آن‌ها ضایعه قابل برگشت بوده و افسردگی قطعه st ایجاد شده ناشی از تست ورزش بالاتر از ۰/۸ باشد، بیشتر به بیماری قلبی مبتلا می‌شوند.
- افرادی که نتایج نوار قلب‌شان موج نرمال داشته و نتایج اسکن تالیوم آن‌ها ضایعه قابل برگشت بوده و افسردگی st ایجاد شده ناشی از تست ورزش کمتر از ۰/۸ باشد و فشار خون نرمال یا کمتر از ۱۳۶ میلی‌جیوه دارند کمتر به بیماری قلبی مبتلا می‌شوند.
- افرادی که نتایج نوار قلب‌شان موج نرمال داشته و نتایج اسکن تالیوم آن‌ها ضایعه قابل برگشت بوده و افسردگی st ایجاد شده ناشی از تست ورزش کمتر از ۰/۸ باشد و فشارخون بالاتر از ۱۳۶ میلی جیوه دارند بیشتر به بیماری قلبی مبتلا می‌شوند.



شکل ۷: ساختار درخت تصمیم بر اساس نتایج نوار قلب (موج غیر قلبی/ موج افزایش مقطعی)

ساختار درخت تصمیم بر اساس نتایج نوار قلب (موج غیرقلبی/ موج افزایش مقطعی) در شکل ۷ نشان می‌دهد:

- افرادی که نتایج نوار قلب‌شان موج غیر قلبی و یا موج افزایش مقطعی داشته‌اند و در هنگام ورزش تپش قلب زیادی احساس می‌کنند بیشتر به بیماری قلبی ابتلا می‌شوند.
- افرادی که نتایج نوار قلب‌شان موج غیر قلبی و یا موج افزایش مقطعی داشته‌اند و فشارخون کمتر از ۱۴۰ میلی جیوه دارند بیشتر به بیماری قلبی ابتلا می‌شوند.
- افرادی که نتایج نوار قلب‌شان موج غیرقلبی و یا موج افزایش مقطعی داشته‌اند و فشارخون بیشتر از ۱۴۰ میلی جیوه دارند کمتر به بیماری قلبی ابتلا می‌شوند.

همان‌طور که مشاهده شد الگوریتم مورد استفاده در این مطالعه، الگوریتم C4.5 بود که دارای میزان دقت قابل قبولی (۸۰٪/۲) می‌باشد. در مرحله ارزیابی نظر متخصص موردنظر نیز در مورد قوانین ایجاد شده اعمال می‌گردد. به این ترتیب که قوانین به دست آمده به متخصص موردنظر ارائه شده و قوانینی که از نظر بالینی معتبر باشند به عنوان قوانین نهایی ارائه گردیدند.

همچنین ریسک فاکتورهای: سطح بالای کلسترول، جنسیت، سن بالا، بالا بودن ماکزیمم ضربان قلب، اسکن تالیوم بالاتر از ۳، نوار قلب غیرنرمال و میزان بالای افسردگی st در تست ورزش، به ترتیب بیشترین تأثیر در ابتلا به بیماری عروق کرونر قلبی را دارا هستند؛ و فاکتورهای درد قفسه سینه و میزان قند خون ناشتا کمترین تأثیر را دارند. این در حالی است که براساس مقایسه‌های انجام شده براساس اولویت‌بندی متغیرها توسط الگوریتم‌های موردبررسی نیز این متغیرها جزء فاکتورهای اول قرار گرفته‌اند که نشان از اهمیت این متغیرها دارد. برخی از مهم‌ترین موارد این قوانین به شرح ذیل می‌باشد:

۱. در ۸۷/۳٪ از بیماران کلسترول خون بالای ۲۰۰ واحد و ابتلا به بیماری عروق کرونر قلبی با هم مشاهده گردید.
۲. در ۸۴٪ از بیماران با جنسیت مذکر ابتلا به بیماری عروق کرونر قلبی مشاهده گردید.
۳. در ۶۴٪ از بیماران بالای ۵۵ سال ابتلا به بیماری عروق کرونر قلبی مشاهده گردید.
۴. همچنین در ۹۲٪ از بیماران بالای ۵۰ سال که دارای کلسترول بالاتر از ۲۰۰ واحد بودند ابتلا به بیماری عروق کرونر قلبی مشاهده گردید.
۵. در ۶۱/۸٪ از بیماران با ماکزیمم ضربان قلب بالای ۱۴۰ ابتلا به بیماری عروق کرونر قلبی مشاهده گردید.
۶. در ۶۰/۶٪ از بیماران با اسکن تالیوم دارای نقص ثابت و قابل برگشت ابتلا به بیماری عروق کرونر قلبی مشاهده گردید.
۷. در ۵۸٪ از بیماران با نوارقلب غیرنرمال ابتلا به بیماری عروق کرونر قلبی مشاهده گردید.
۸. در ۵۴/۵٪ از بیماران با میزان افسردگی st در تست ورزش بالای ۰/۸ ابتلا به بیماری عروق کرونر قلبی مشاهده گردید.
۹. در ۷/۲٪ از بیماران با دارا بودن قند خون ناشتا ابتلا به بیماری عروق کرونر قلبی مشاهده گردید.

بحث و نتیجه‌گیری

در این پژوهش، با استفاده از الگوریتم درخت تصمیم C4.5 به ارائه مدل و استخراج قوانین آن در راستای پیشگویی احتمال ابتلا به بیماری عروق کرونر پرداختیم. از درخت تصمیم C4.5 نتایج قابل قبولی به دست آمد که دقت آن ۸۰/۲٪ بود و حساسیت آن یعنی تعداد نمونه‌هایی که به درستی عدم وجود ناراحتی قلبی را نشان داده‌اند نسبت به کل نمونه‌هایی که واقعاً بیماری عروق کرونر قلبی ندارند ۸۴/۱٪ می‌باشد.
همچنین با توجه به محاسبات انجام شده نرخ دسته‌بندی برابر است با ۶/۷۲٪ و مقدار پیش بینی مثبت ۱/۸۴٪ و مقدار پیش‌بینی منفی ۸/۶۱٪ به دست آمده است که در مقایسه با نتایج مطالعات انجام شده در حوزه داده‌کاوی بیماری قلبی با الگوریتم درخت تصمیم، دقت به دست آمده الگوریتم پیشنهادی، قابل قبول است. این نتایج در جدول ۳ مشاهده می‌شود.

جدول ۳: مقایسه معیارهای الگوریتم درخت تصمیم C4.5 با سایر الگوریتم‌های درخت تصمیم

معیارها	C4.5(مدل پیشنهادی)	C&RT	QUEST	CHAID
حساسیت	۸۴٪/۱	۸۲٪/۶۷	۵۶٪	۸۳٪
ویژگی	۷۲٪/۶	۷۶٪/۵	۸۴٪/۵	۸۷٪
ارزش اخباری مثبت	۸۴٪/۱	۷۲٪/۵	۷۳٪	۶۲٪/۵
ارزش اخباری منفی	۶۱٪/۸	۸۵٪/۵	۷۱٪/۹	۵۸٪/۳۳
دقت	۸۰٪/۲	۷۹٪/۱	۷۲٪/۲۸	۸۵٪/۷

همچنین با بررسی دقیق‌تر یافته‌های حاصل از مدل موردنظرمان بیشترین فاکتورهای تأثیرگذار در ابتلا به‌ترتیب متغیرهای: سطح بالای کلسترول، جنسیت، سن بالا، بالا بودن ماکزیمم ضربان قلب، اسکن تالیوم بالاتر از ۳، نوار قلب غیرنرمال و بالا بودن میزان افسردگی ایجاد شده st ناشی از تست ورزش می‌باشد. با استفاده از قوانین به دست‌آمده برای یک فرد جدید با داشتن متغیرهای مشخص، می‌توان تعیین کرد که احتمال ابتلای وی به بیماری کرونر قلبی چقدر خواهد بود. در جدول ۴ به مقایسه نتایج پژوهش مشابه با پژوهش حاضر می‌پردازیم.

جدول ۴: مقایسه نتایج مطالعات انجام شده در حوزه داده‌کاوی در بیماری قلبی

نویسندگان و سال ارائه تحقیق	الگوریتم‌های مورداستفاده	نوع بیماری	دقت مدل نهایی (نوع)	یافته‌ها	متغیرهای پیشگویی کننده
Christine (۱۹۹۸) [۲۱]	رگرسیون لجستیک، درخت طبقه بندی	انفارکتوس قلبی	۸۱٪ (درخت طبقه‌بندی)	عملکرد بهتر درخت تصمیم در پیشگویی ابتلا به انفارکتوس قلبی	سن، سابقه خانوادگی بیماری قلبی، مصرف سیگار، درد در ناحیه قفسه سینه، فشارخون بالا، دیابت، تعریق شبانه، استفراغ، جنسیت و ...
Kajabadi (۲۰۰۹) [۲۲]	درخت تصمیم	بیماری عروق کرونر	محاسبه نشده	عوامل تأثیرگذار عمده بر بروز بیماری قلبی مشخص شده‌اند	چربی، فاکتورهای خونی، فاکتورهای چاقی، متغیرهای قندی، متغیرهای عمومی، ...
Karaolis (۲۰۰۹) [۱۹]	درخت تصمیم C4.5	انفارکتوس قلبی، پیوند عروق کرونر قلبی	۶۶٪ (درخت تصمیم C4.5)	عوامل تأثیرگذار عمده بر بروز انفارکتوس قلبی مشخص نشده‌اند	جنسیت، سن، فشارخون بالا، چربی خون بالا، مصرف سیگار، سطح کلسترول، دیابت و ...
Jyoti (۲۰۱۱) [۲۰]	شبکه بیز درخت تصمیم، شبکه عصبی مصنوعی	بیماری قلبی	۸۹٪(درخت تصمیم)	ایجاد قوانینی جهت یافتن ارتباط بین متغیرها	جنسیت، سن، درد قفسه سینه، فشارخون بالا، قندخون ناشتا، سطح کلسترول، مصرف سیگار و ...
مدل پیشنهادی	درخت تصمیم C4.5	بیماری عروق کرونر قلبی	۸۰/۲٪	عوامل تأثیرگذار بر بروز بیماری، ایجاد قوانینی جهت پیشگویی و تشخیص بیماری	جنسیت، سن، سطح کلسترول، درد قفسه سینه، فشارخون بالا و ...

مطابق مطالعات گذشته، عملکرد مدل‌های طبقه‌بندی کننده ممکن است بر روی پایگاه‌های داده مختلف نتایج متفاوتی داشته باشد. برای مثال Karaolis و همکاران برای تشخیص انفارکتوس قلبی و پیوند عروق کرونر قلبی با استفاده از درخت تصمیم‌گیری، از "الگوریتم درخت تصمیم برای ارزیابی ریسک فاکتورهای بیماری عروق کرونر استفاده کردند"[۱۹]. یافته‌های آن مشابه با قوانین استخراج شده از الگوریتم درخت تصمیم در مطالعه حاضر بوده، ضمن اینکه درخت تصمیم ارائه شده در این مطالعه از دقت بالاتری برخوردار بوده است.

در بررسی انجام شده توسط Jyoti جهت پیش‌بینی احتمال ابتلا به بیماری قلبی مدل ارائه شده توسط درخت تصمیم دارای دقت ۸۹٪ بوده است [۲۰]. در این پژوهش از شبکه عصبی مصنوعی نیز استفاده شده است.

در پژوهش Christine به مقایسه عملکرد رگرسیون لجستیک و چند الگوریتم از درخت تصمیم در تعیین ابتلا به انفارکتوس قلبی پرداخته شده است و مدل درخت تصمیم با حساسیت ۸۱٪ به عنوان مدل مناسبی جهت پیشگویی معرفی شده است [۲۱]. همان طور که در جدول ۴ مشاهده شد دقت ارائه شده در این پژوهش در مقایسه با سایر پژوهش‌های انجام شده بالا و قابل قبول بوده و می‌تواند در طراحی مدل‌های مناسب جهت پیشگویی امکان ابتلای افراد به بیماری‌های قلبی استفاده شود. همچنین می‌تواند در برنامه‌های غربالگری جهت شناسایی افراد در معرض خطر استفاده شود. ضمناً با توجه به این که نتایج تحقیق بر روی داده‌های استاندارد صورت گرفت، این نتایج می‌تواند به عنوان مبنای ارزیابی برای پژوهش‌های آتی قرار گیرد. همچنین پیشنهاد می‌شود که این مدل با مجموعه داده-های واقعی بیشتر و در بازه زمانی طولانی اجرا شده و پس از رسیدن به سطح دقت مطلوب در برنامه‌های غربالگری مورد استفاده قرار گیرد. آنگاه پس از ایجاد تغییرات ضروری به عنوان مدل مناسب جهت پیشگویی بیماری عروق کرونر قلبی مورد استفاده قرار گیرد.

References

1. Bridget B. Kelly; Institute of Medicine; Fuster, Valentin. Promoting Cardiovascular Health in the Developing World. A Critical Challenge to Achieve Global Health. Washington, D.C; National Academies Press; 2010.

2. Das R, Turkoglu I, Sengur A. Effective diagnosis of heart disease through neural networks ensembles. Expert Systems with Applications 2009;36(4):7675-80.

3. Soni J, Ansari U, Sharma D, Soni S. Predictive Data Mining for Medical Diagnoses: An Overview of Heart Disease Prediction. International Journal of Computer Applications 2011;17(8):85-93.

4. SubbalakshmiG, Ramesh K, Chinna Rao M. Decision Support in Heart Disease Prediction System using Naive Bayes. Indian Journal of Computer Science and Engineering 2011;2(2):183-95.

5. Boo S, Froelicher ES. Cardiovascular Risk Factors and 10-year Risk for Coronary Heart Disease in

Korean Women. Asian Nursing Research 2012;6(1):1-8.
6. Mohammadi F, Taherian A, Hosseini M A, Rahgozar M. Effect of Home-Based Cardiac Rehabilitation Quality of Life in the Patients with Myocardical Infarction. Archives of Rehabilitation 2006; 7 (3):11-9. Persian
7. Fayyad M, Piatetsky-Shapiro G, Smyth P, Uthurusamy R, Smyth Advances in Knowledge Discovery and Data Mining. American Association for Artificial Intelligence USA: 1996.
8. Lavrac N Selected techniques for data mining in medicine. Artif Intell Med 1999;16(1):3-23.
9. Huang MJ, Chen MY, Lee SC. Integrating data mining with case-based reasoning for chronic diseases prognosis and diagnosis. Expert Systems with Applications 2007;32(3):856-67.
10. Witten IH, Frank E Hall MA. Data Mining: Practical Machine Learning Tools and Techniques. 4th ed. Morgan Kaufmann: Elsevier Science; 2016.
11. Colombet I, Ruelland A, Chatellier G, Gueyffier F, Degoulet P, Jaulent MC. Models to predict cardiovascular risk: comparison of CART, multilayer perceptron and logistic regression. Proc AMIA Symp 2000; 156–60.
12. Bellaachia A, Guven E. Predicting Breast Cancer Survivability Using Data Mining Techniques. Washington University 2005; 2(2): 53-62.
13. Stärk KDC, Pfeiffer DU. The application of non-parametric techniques to solve classification problems in complex data sets in veterinary epidemiology – An example. Intelligent Data Analysis 1999;3(1):23-35.
14. Ordonez C. Comparing Association rules and Decision Trees for Disease Prediction. Intelligent Data Analysis 2011; 15(2): 3-9.
15. Roiger RJ. Data mining: A tutorial-based primer. 2th ed U.S. Florida: CRC Press; 2017.
16. Ghazanfari M. Data Mining and Knowledge Discovery. Tehran: Publication of Elmo Sanaat; 2008. Persian
17. Chen J, Xing Y, Xi G, Chen J, Yi J, Zhao D, et al. A Comparison of Four Data Mining Models: Bayes, Neural Network, SVM and Decision Trees in Identifying Syndromes in Coronary Heart Disease. In: 18. Liu D, Fei S, Hou ZG, Zhang H, Sun C, editors. Advances in Neural Networks – ISNN 2007: 4th International Symposium on Neural Networks, ISNN 2007, Nanjing, China, Jun 3-7, 2007, Proceedings, Part I. Berlin, Heidelberg: Springer Berlin Heidelberg; 2007. p. 1274-9.
19. Karaolis M, Moutiris JA, Papaconstantinou L, Pattichis CS. Association rule analysis for the assessment of the risk of coronary heart events. Conf Proc IEEE Eng Med Biol Soc 2009;2009:6238-41.
20. Jyoti S, Ujma A, Dipesh S, Sunita S. Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction. International Journal of Computer Applications 2011; 17(8): 35-43.
21. Tsien CL1, Fraser HS, Long WJ, Kennedy RL. Using classification tree and logistic regression methods to diagnose myocardial infarction. Stud Health Technol Inform 1998;52 Pt 1:493-7.
22. Kajabadi A, Saraee M, Asgari S. Medical Data Mining: An Approach to Discovery Relationships among Cardiovascular Risk Factors. 3rd Iran Datamining Conference, Iran, Tehran; 2009; Available from: www.civilica.Com/Paper-IDMC03-IDMC03_065.

Original Article

Journal of Health and Biomedical Informatics
Medical Informatics Research Center
2017; 3(4): 287-299

Detection of Coronary Artery Disease Using C4.5 Decision Tree

Sabbagh Gol Hamed[1]*

• **Received:** 15 Dec, 2016 • **Accepted:** 7 Mar, 2017

Introduction: Today, one of the most common diseases and causes of death in the world is heart diseases. Data mining techniques are very useful to create predictive models for identifying people at risk and decreasing the disease complications. In this study, using C4.5 decision tree method, the prevention and diagnosis of this disease are discussed.

Methods: This was an applied descriptive study. UCI standard data and Cleveland data collection were used. The database contains 297 records. Analysis was performed through Weka software and using CRISP3 methodology. The C4.5 decision tree model, using input variables and determining the target variable, was created.

Results: According to the applied model, it was found that high levels of cholesterol, sex, age, high maximum heart rate, scan thallium higher than 3 and abnormal ECG have the greatest impact on the risk of coronary heart disease. Furthermore, by using the created decision tree, some rules were extracted that can be used as a model to predict the risk of coronary heart disease. The accuracy of the model created by using decision tree was over 80 percent.

Conclusion: According to our calculations, the rate of categorization was 72.6% and the accuracy of C4.5 algorithm was 80.2% that in comparison with the results of studies in the field of data mining of heart diseases, the obtained accuracy for the suggested algorithm is acceptable.

Keywords: Data mining, Coronary artery disease, C4.5 Decision tree.

• **Citation:** Sabbagh Gol H. Detection of Coronary Artery Disease Using C4.5 Decision Tree. Journal of Health and Biomedical Informatics 2017; 3(4): 287-299.

1. M.S.c in Computer Engineering, Faculty of Computer, Department of Computer Engineering, Payame Noor University (PNU), Birjand Branch. Birjand, Iran

***Correspondence**: End of Shahid Avini Blvd, South Khorasan Payame Noor University, Birjand Branch. Birjand

• **Tel:** 09151643216 • **Email:** sabbagh.h@pnu.ac.ir