مقاله پژوهشی

مجله انفورماتیک سلامت و زیست پزشکی
مرکز تحقیقات انفورماتیک پزشکی
دوره چهارم شماره دوم. ص ۱۲۱-۱۰۸. ۱۳۹۶

تشخیص بیماری تب کریمه‌کنگو با استفاده از درخت تصمیم C4.5

رضا اسماعیلی گوهری[۱]، الهام اسماعیلی گوهری[۲]*، مهدی شفیعی[۳]

• دریافت مقاله: ۹۶/۵/۱ • پذیرش مقاله: ۹۶/۶/۲۷

مقدمه: با شروع فصل تابستان، بیماری بین انسان و حیوان، یعنی تب کریمه‌کنگو به سرعت شیوع پیدا می‌کند. تشخیص این بیماری با استفاده از آزمایش‌های لازم، در کمترین حالت زمانی حدود یک هفته به طول می‌انجامد. روش‌های داده‌کاوی و یادگیری ماشین متعددی برای ایجاد مدل‌های پیشگویی‌کننده جهت شناسایی افراد در معرض خطر وجود دارد. در این پژوهش از درخت تصمیم C4.5 به دلیل سادگی و کارآمدی‌اش به منظور تشخیص این بیماری استفاده شده است.

روش: این پژوهش از نوع کاربردی و توصیفی است. در این پژوهش از داده‌های مربوط به افراد مظنون به بیماری تب کریمه‌کنگو استفاده شد. این داده‌ها در یک دوره ۴ ساله از سال ۱۳۹۳ از مراکز درمانی کشور جمع‌آوری شد. این پایگاه داده شامل ۹۶۵ رکورد و ۲۸ ویژگی است. ابتدا با استفاده از روش انتخاب ویژگی برنامه‌نویسی درجه دو، متغیرهای مؤثر و تأثیرگذار بر مدل انتخاب و سپس درخت تصمیم C4.5 با به کارگیری متغیرهای ورودی و تعیین متغیر هدف ایجاد گردید. تجزیه‌وتحلیل داده‌ها به کمک نرم‌افزار Matlab صورت گرفت.

نتایج: با توجه به مدل مشخص شد که متغیرهایی همچون تب، خون‌ریزی، شروع ناگهانی علائم، افزایش آنزیم‌های کبدی، افزایش بیلی روبین توتال، کاهش هموگلوبین، Hematuria، Leukocytosis، Proteinuria و Leukopenia بیشترین تأثیر را در تشخیص به این بیماری دارند.

نتیجه‌گیری: نتایج نشان می‌دهد که معیار حساسیت مدل پیشنهادی، ۹۵٪ و معیار تشخیص آن ۵۰٪ است که در مقایسه با مطالعات انجام‌شده دیگر در حوزه داده‌کاوی پزشکی، از اثربخشی قابل قبولی در تشخیص این بیماری برخوردار است.

کلید واژه‌ها: سیستم تصمیم‌یار پزشکی، تشخیص بیماری، تب کریمه‌کنگو، درخت تصمیم C4.5

•ارجاع: اسماعیلی گوهری رضا، اسماعیلی گوهری الهام، شفیعی مهدی. تشخیص بیماری تب کریمه‌کنگو با استفاده از درخت تصمیم C4.5. مجله انفورماتیک سلامت و زیست پزشکی ۱۳۹۶؛ ۴(۲): ۱۲۱-۱۰۸.

۱. کارشناس ارشد مهندسی کامپیوتر، گروه مهندسی کامپیوتر، مؤسسه آموزش عالی بهمنیار، کرمان، ایران.
۲. کارشناس ارشد مهندسی کامپیوتر، گروه مهندسی کامپیوتر، دانشکده فنی و مهندسی، دانشگاه یزد، یزد، ایران.
۳. دکترای حرفه‌ای-MPH، گروه بیماری‌های معاونت بهداشتی، دانشگاه علوم پزشکی کرمان، کرمان، ایران.
*** نویسنده مسئول:** یزد، صفائیه، دانشگاه یزد، دانشکده فنی و مهندسی، گروه مهندسی کامپیوتر.
• شماره تماس: ۰۹۱۳۳۴۲۷۴۳۵
• **Email:** g.elhamesmaeeli@gmail.com

مقدمه

تب کریمه کنگو (Crimean-Congo Hemorrhagic Fever(CCHF یک بیماری خونریزی‌دهنده ویروسی است که عامل آن ویروسی از گروه Arbovirus و جنس Nairovirus و از خانواده Bunyavirdae می‌باشد. این بیماری برای اولین بار در سال ۱۹۴۲ در کریمه روسیه دیده شد. در سال ۱۹۴۴ در خلال جنگ جهانی دوم بیماری‌ای در شبه جزیره کریمه شایع و باعث مرگ بیش از ۲۰۰ نفر از روستاییان و سربازان گردید. سپس در سال ۱۹۵۶ موارد مشابه آن در کنگو (زئیر) مشاهده گردید و ویروس عامل بیماری از افراد مبتلا جداسازی شد و به عنوان ویروس کنگو نام‌گذاری شد. در سال ۱۹۶۹ مشخص شد که عامل ایجادکننده تب خونریزی‌دهنده کریمه مشابه عامل بیماری است که در سال ۱۹۵۶ در کنگو شناخته شده است و به همین دلیل نام تب خونریزی‌دهنده کریمه کنگو برای این بیماری ویروسی نامیده شد [۱]. در سال ۱۹۷۰، این بیماری برای اولین بار در ایران گزارش شد. در این بیماری، بیمار در طی مدتی که در بیمارستان بستری است به شدت برای دیگران آلوده‌کننده است. دوره کمون این بیماری ۹-۴ روز است. شایع‌ترین علائم بیماری عبارت‌اند از تب، سردرد، درد عضلانی، حالت تهوع، درد شکم، اسهال، سرگیجه و حساسیت به نور، خروج خون از منافذ بدن، خونریزی‌های پتشی زیرپوستی (لکه‌های خونی زیرپوستی) [۱].

تشخیص صحیح بیماری‌ها در بسیاری از موارد، امری مهم و دشوار است. محیط مراقبت سلامت، سرشار از داده‌های خام مفید است درحالی که ما نیازمند دانش در این حیطه هستیم [۲]. حجم داده‌های پزشکی روزبه‌روز در حال افزایش است و پزشکان معمولاً اطلاعات ارزشمندی را در خصوص بیماری‌ها، ارتباط آن‌ها با یکدیگر و عوامل ایجادکننده آن‌ها به دست می‌آورند [۳]؛ اما این مجموعه داده‌های خام به خودی‌خود ارزشی ندارند و برای معنی بخشیدن به این داده‌ها باید آن‌ها را تحلیل و به اطلاعات و یا به عبارت بهتر به دانش تبدیل کرد [۴]. امروزه استفاده از داده‌کاوی در تحقیقات زیست پزشکی بسیار مورد توجه قرار گرفته است. داده‌کاوی می‌تواند ارتباطات و وابستگی‌های جدید و بدیعی را کشف کند که برای پزشکان مفید هستند؛ به عبارت دیگر، از داده‌کاوی می‌توان به عنوان یک روش معتبر، حساس و قابل اعتماد برای کشف الگوها و روابط بین آن‌ها استفاده کرد [۵].

ابزارهای داده‌کاوی به طور گسترده در زمینه‌های مختلف به کار می‌روند، برای مثال می‌توان از کاربرد داده‌کاوی در مواردی چون شناسایی الگوهای بازاریابی، پیش‌بینی رفتار مشتری، تشخیص و پیش‌بینی بیماری‌ها و شناسایی تقلب نام برد [۶]. امروزه، استفاده از روش‌های متنوع داده‌کاوی و استخراج دانش برای شناسایی الگوها و ارتباطات میان متغیرهای مختلف در تولید مدل‌های پیش‌بینی‌کننده در علوم پزشکی بسیار مورد توجه قرار گرفته است [۷]. یکی از کاربردهای داده‌کاوی در حیطه پزشکی، استفاده از آن در جهت شناسایی و تشخیص بیماری‌ها، دسته‌بندی بیماران و پیدا کردن الگوهایی برای تشخیص سریع‌تر بیماری و جلوگیری از بروز عوارض در آن‌ها است. داده‌کاوی پزشکی دارای پتانسیل زیادی برای کشف الگوهای پنهان موجود در داده‌ها است که این الگوها می‌توانند برای تشخیص‌های بالینی مورد استفاده قرار گیرند [۸، ۴]. افزایش دقت تشخیص، کاهش هزینه‌ها و کاهش منابع انسانی از جمله مزایای استفاده از داده‌کاوی در تجزیه و تحلیل پزشکی است [۱۰، ۹]. به طور کلی تکنیک‌های داده‌کاوی به دو دسته تقسیم می‌شوند: توصیف‌کننده و پیشگویی‌کننده. تکنیک‌های توصیفی، خواص عمومی داده‌ها را مشخص می‌کنند و هدف آن‌ها پیدا کردن الگوهای قابل تفسیر توسط انسان از داده‌ها است. تکنیک‌های پیشگویانه، پیش‌بینی رفتار آینده داده‌ها را برعهده دارند و هدف آن‌ها به کارگیری چند متغیر در پایگاه داده برای پیش‌بینی مقادیر آینده یا ناشناخته دیگر متغیرها است [۱۱]. یکی از روش‌های پیش‌بینی، دسته‌بندی است. دسته‌بندی یک تابع یادگیری است که یک قلم داده را به یکی از دسته‌های از قبل تعریف‌شده نگاشت می‌کند.

یکی از روش‌های متداول دسته‌بندی، درخت تصمیم است. ساختار درخت تصمیم در یادگیری ماشین، یک مدل پیش‌بینی‌کننده است که حقایق مشاهده شده در مورد یک پدیده را به استنتاج‌هایی در مورد مقدار هدف آن پدیده نگاشت می‌کند. یادگیری درخت تصمیم یکی از رایج‌ترین روش‌های داده‌کاوی است که به دلیل سادگی و کارآمدی‌اش، علی رغم مشکلاتی از جمله امکان وجود صفات دارای نویز و یا صفات فاقد مقدار، به شکل گسترده‌ای در مسائل مختلفی استفاده می‌شود. پرکاربردترین الگوریتم مورد استفاده در ساخت درخت تصمیم، الگوریتم C4.5 است [۱۳، ۱۲]. در ادامه به بیان تعدادی از پژوهش‌هایی که با استفاده از الگوریتم‌های یادگیری ماشین به پیش‌بینی و تشخیص بیماری‌های مختلف پرداخته‌اند، می‌پردازیم.

Tanner و همکاران از دسته‌بند درخت تصمیم C4.5 به

منظور تشخیص بیماری تب دنگی استفاده کردند. آن‌ها از پارامترهای بالینی و خون‌شناسی برای این منظور بهره گرفتند. روش آن‌ها قادر است تا افراد سالم و بیمار را در ۷۲ ساعت اول بیماری تشخیص دهد. آن‌ها با استفاده از معیارهای حساسیت و تشخیص به ارزیابی مدل پرداختند. نتایج آزمایش‌های آن‌ها بر روی ۱۲۰۰ بیمار، نشان‌دهنده دقت بالای الگوریتم پیشنهادی آن‌ها در تشخیص این بیماری است [۱۴]. Shaukat و همکاران، برای پیش‌بینی بیماری تب دنگی، از چندین تکنیک دسته‌بندی از جمله الگوریتم بیز، درخت REP، درخت تصمیم J48 و SMO استفاده کردند. نتایج ارزیابی آن‌ها نشان می‌دهد که از میان روش‌های موجود، الگوریتم بیز با ۹۲٪ صحت بیشتری را نسبت به سایر الگوریتم‌ها داشته است [۱۵]. Saha و همکاران برای دسته‌بندی ویژگی‌های مؤثر در تب دنگی، از شبکه عصبی پرسپترون چندلایه و ماشین بردار پشتیبان استفاده کرده‌اند. نتایج ارزیابی آن‌ها نشان می‌دهد که ماشین بردار پشتیبان صحت بهتری در دسته‌بندی نمونه‌های منفی و مثبت دارد [۱۶]. در تحقیق دیگری که توسط Saikia و همکاران انجام شد، از سیستم خبره فازی برای تشخیص بیماری تب دنگی استفاده شده است. آن‌ها با استفاده از علائم بیماری و آزمایش‌های پزشکی فرد، سیستم استنتاج فازی‌ای ایجاد کردند که بتواند در مراحل اولیه تب دنگی را تشخیص دهد [۱۷].

در تحقیقی که توسط Olanow و همکاران انجام شد، از درخت تصمیم در تشخیص بیماری پارکینسون استفاده شد. تشخیص بیماری پارکیسنون در مراحل ابتدایی بیماری کاری دشوار است. آن‌ها تلاش کردند که با استفاده از درخت تصمیم و استخراج ویژگی‌های ظاهری و آزمایشگاهی به تشخیص این بیماری بپردازند. نتایج آزمایش‌ها نشان از دقت بالای الگوریتم آن‌ها در تشخیص این بیماری دارد [۱۸]. در تحقیق Cai و همکاران، برای پیش‌بینی بیماری پارکینسون، از الگوریتم ماشین بردار پشتیبان استفاده شده است. آن‌ها برای بهینه‌سازی پارامترهای مورد استفاده خود از الگوریتم بهینه‌سازی غذایابی باکتری استفاده کردند. صحت دسته‌بند آن‌ها در بهترین حالت برابر با ۹۷٪ به دست آمده است [۱۹]. Chen و همکاران برای تشخیص اولیه بیماری پارکینسون از الگوریتم ELM و KELM استفاده کردند. نتایج نشان می‌دهد که استفاده از KELM آسان‌تر از ELM است. علاوه‌بر این، KELM نتایج بهتری نسبت به ELM دارد [۲۰].

Wang و همکاران با استفاده از ترکیب درخت تصمیم و شبکه عصبی بر روی ۲۹۶۴ زن به پیش‌بینی خطر ابتلا به پوکی استخوان در زنان پرداختند. آن‌ها از ۳۳ ویژگی برای این منظور استفاده کردند. میانگین دقت آن‌ها در بهترین مدل ۷۰/۵ درصد بوده است [۲۱]. در تحقیق دیگری که توسط Gao و همکاران بر روی ۱۴ فرد سالم و ۱۴ فرد بیمار مبتلا به پوکی استخوان انجام شد، آن‌ها موفق شدند با استفاده از درخت تصمیم (الگوریتم C4.5) با دقت مناسبی به تشخیص بیماری پوکی استخوان بپردازند [۲۲].

در مطالعه انجام‌شده توسط Tsien و همکاران، جهت پیش‌بینی امکان زنده ماندن بیماران مبتلا به سرطان سینه الگوریتم C4.5 با میزان دقت ۸۶/۷ درصد به عنوان بهترین مدل پیشگویی‌کننده معرفی شده است [۲۳]. در مطالعه انجام‌شده توسط Bellaachia و همکاران، بیان شده است که مدل‌های درخت تصمیم (از جمله ID3 و C4.5) برای انجام تحلیل‌های اکتشافی در پایگاه داده‌های بزرگ نسبت به مدل‌های دیگر موفق‌تر عمل می‌کنند [۲۴].

Soni و همکاران با استفاده از سه الگوریتم شبکه بیز، درخت تصمیم و شبکه عصبی مصنوعی به تشخیص بیماری قلبی و افراد در معرض خطر پرداختند. آن‌ها با استفاده از ویژگی‌هایی از جمله جنسیت، سن، درد قفسه سینه، فشارخون بالا، قندخون ناشتا، سطح کلسترول و مصرف سیگار به ایجاد قوانینی جهت یافتن ارتباط بین متغیرها پرداختند. نتایج ارزیابی‌ها نشان داد که درخت تصمیم با ۸۹٪ بیشترین دقت را در میان روش‌های دیگر دارد [۳].

Su و همکاران توانستند بر پایه روش‌های شبکه عصبی مصنوعی، درخت تصمیم، رگرسیون و قواعد وابستگی بر پایه عکس‌های سه بعدی و دوبعدی بدن با دقت ۸۹٪ پیش‌بینی کنند که آیا فرد موردنظر به بیماری دیابت مبتلا است یا خیر؟ [۲۵].

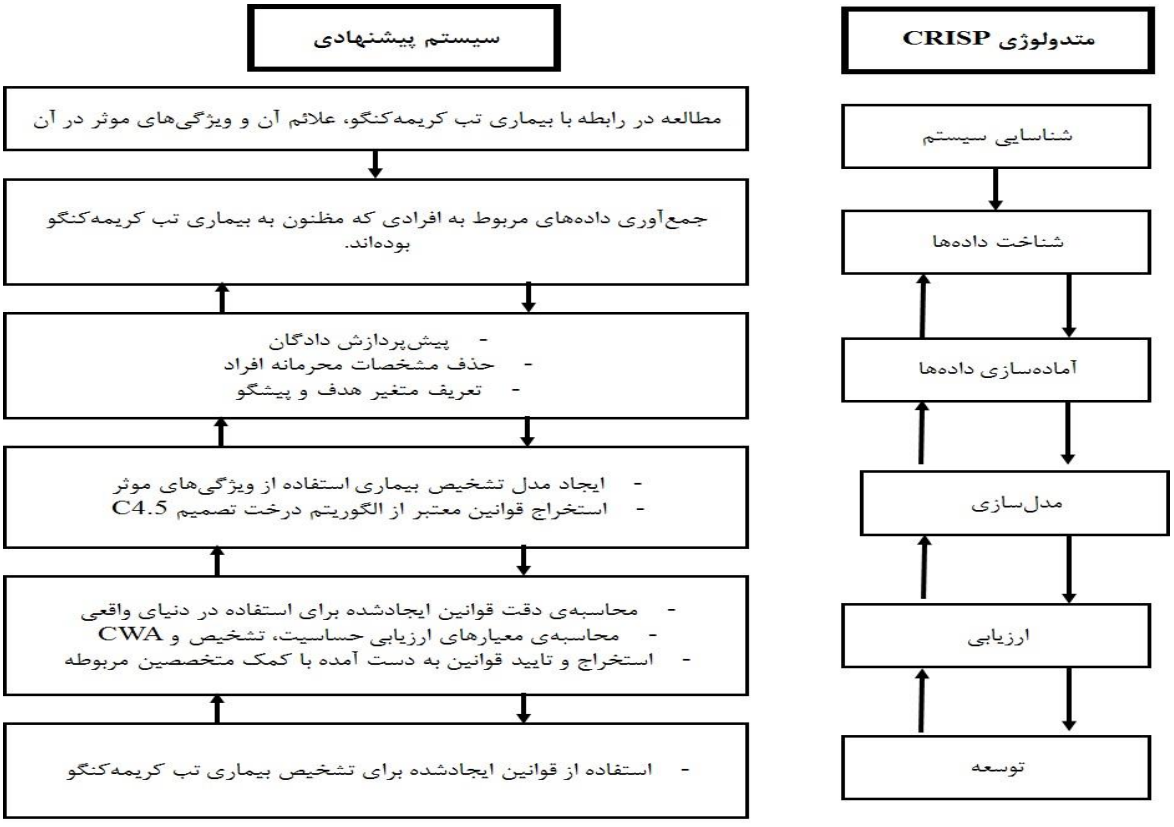
یکی از حوزه‌های فعال تحقیق در زمینه‌های انفورماتیک پزشکی و سیستم‌های تصمیم‌یار پزشکی، پزشکی اضطراری است. به بیان ساده‌تر برای بیماران مبتلا به مریضی‌های واگیردار از جمله تب کریمه‌کنگو بایستی روند تشخیص و درمان به موقع و سریع انجام شود. از طرف دیگر تشخیص اشتباه و یا دیر، ممکن است باعث از دست رفتن بیمار یا سرایت آن بیماری به افراد دیگر شود. در این زمینه به دلیل طولانی بودن روند انجام آزمایش‌ها و برگرداندن نتایج، تشخیص به

موقع برای بیماران یکی از چالش‌هایی است که متخصصان با آن روبه‌رو هستند.

با توجه به توسعه سیستم‌های تصمیم‌یار در حوزه پزشکی، در این مطالعه، هدف ارائه یک سیستم تصمیم‌یار برای تشخیص بیماری تب کریمه‌کنگو و یاری رساندن به پزشک برای تشخیص اقدامات لازم بعدی است. بدین منظور از درخت تصمیم C4.5 برای دسته‌بندی افراد بیمار و سالم استفاده شده است. داده‌های مورد استفاده در این پژوهش، شامل داده‌های مربوط به بیمارانی است که با علائم بیماری تب کریمه‌کنگو در طول سال‌های ۱۳۹۳ تا ۱۳۹۶ در سراسر کشور به مراکز درمانی مراجعه کرده‌اند.

روش

این مطالعه از نوع کاربردی-توصیفی است. تاکنون روش‌های متعددی برای اجرای پروژه‌های داده‌کاوی ارائه شده است، روش کریسپ (CRoss Industry Standard Process for Data Mining) CRISP-DM یکی از روش‌های رایج و قدرتمند در این زمینه است. این روش شناسی یک روش صنعتی اثبات شده برای هدایت تلاش‌های داده‌کاوی است [۲۶]. در این پژوهش نیز از این روش بهره برده شده است. در شکل ۱، به بررسی هر یک از مراحل روش کریسپ در جهت رسیدن به مدلی برای تشخیص بیماری تب کریمه‌کنگو پرداخته شده است.



شکل ۱: مراحل روش کریسپ برای ساخت مدلی برای تشخیص بیماری تب کریمه‌کنگو

الف) شناخت سیستم

استفاده موفق از مفاهیم داده‌کاوی، مستلزم شناخت حوزه‌ای است که قصد استفاده از داده‌کاوی را در آن زمینه داریم. همچنین بایستی از روش‌ها و ابزارهای داده‌کاوی نیز شناخت کافی وجود داشته باشد. برای کسب دانش کافی در حوزه تشخیص بیماری تب کریمه‌کنگو، با مشورت پزشکان متخصص و کارشناسان مراکز درمانی، بررسی مستندات موجود در زمینه تب کریمه‌کنگو، آشنایی با روش‌های تشخیصی و درمانی و ویژگی‌های مؤثر در ابتلا به این بیماری، سعی شد تا شناخت کافی و جامعی در حوزه مورد بررسی کسب شود. همان طور که بیان شد، در این تحقیق، هدف، تشخیص بیماری تب کریمه‌کنگو است؛ بنابراین با نگاشت افراد به برچسب سالم/بیمار، تشخیص این بیماری به یک مسئله دسته‌بندی تبدیل خواهد شد. از طرف دیگر می‌توان به این مسئله به گونه

دیگری نیز نگریست. پزشک برای تصمیمات بعدی خود برای درمان و بررسی وضعیت فعلی بیمار، آزمایش‌هایی را تجویز می‌کند، اگر بتوان سیستمی ارائه کرد که بتواند مقدار این آزمایش‌ها را برای پزشک پیش‌بینی کند، در این صورت، این سیستم می‌تواند یک سیستم تصمیم‌یار پزشک باشد. آزمایش تب کریمه‌کنگو مقدار آنتی‌بادی IgG را تشخیص می‌دهد، در صورتی که این آنتی‌بادی در خون موجود باشد، مقدار آنتی‌بادی مثبت است، در نتیجه فرد دارای بیماری تب کریمه‌کنگو نیست؛ اما در صورتی که مقدار این آنتی‌بادی در خون منفی باشد، فرد به عنوان بیمار تب کریمه‌کنگو معرفی می‌شود. به همین خاطر نتایج آزمایش تب کریمه‌کنگو به صورت منفی یا مثبت بیان می‌شود. پس نتایج آزمایش به صورت مقداری دودویی است؛ بنابراین می‌توان در این تحقیق از دو رویکرد مختلف استفاده کرد. در رویکرد اول، مسئله به صورت یک مسئله دسته‌بندی مطرح شده است و در رویکرد دوم، پیش‌بینی مقدار آزمایش، هدف اصلی قرار گرفته است. از آنجایی که مقادیر این آزمایش به صورت دودویی، مثبت یا منفی است، می‌توان این دو مسئله را به یک مسئله واحد نگاشت کرد؛ بنابراین بایستی با استفاده از راهنمایی پزشکان و کارشناسان به تعیین فاکتورهای مؤثر در تشخیص بیماری تب کریمه‌کنگو پرداخته و سپس با به کارگیری الگوریتم‌های لازم از جمله درخت تصمیم، به تشخیص این بیماری پرداخت.

ب) آماده‌سازی داده‌ها

در این تحقیق از داده‌های مربوط به افراد مظنون به بیماری تب کریمه‌کنگو استفاده شده است. این داده‌ها در یک دوره ۴ ساله از سال ۱۳۹۳ از مراکز درمانی سراسر کشور جمع‌آوری شده است.

این پایگاه داده حاوی دسته‌های متمایزی از داده‌ها است؛ از جمله داده‌ها و علائم بالینی، داده‌های آزمایشگاهی، یادداشت‌های پزشکان و پرستاران، داروهای تجویزی و اطلاعات عمومی فرد از جمله شغل و محل زندگی. محتوای پایگاه داده از داده‌های واقعی به دست آمده است؛ بنابراین داده‌ها حاوی اطلاعات محافظت‌شده پزشکی هستند. به همین دلیل ابتدا داده‌ها غیرقابل شناسایی شده‌اند؛ به عبارت دیگر نام، نام خانوادگی، آدرس، شماره تلفن منزل و شماره تلفن همراه فرد مظنون تغییر یافته است. عملیاتی که در این رابطه صورت گرفت، عبارت‌اند از:

۱) تمام تاریخ‌ها به تاریخ‌هایی در آینده تبدیل شده‌اند. مثلاً "۰۷-۱۲-۱۴۰۰ ۰۰:۰۰:۰۰" تاریخ و زمان بستری یک مظنون در بیمارستان (تاریخ مراجعه به مراکز درمانی) است. ماه و روز به همان صورت اصلی حفظ شده‌اند و سال بستری به سالی در آینده تبدیل شده است.

۲) برای هر فرد، تمام تاریخ‌های مراجعه به صورت یکسانی تبدیل شده‌اند تا فواصل زمانی حفظ شوند.

۳) روزهای هفته و فصل‌های سال به همان صورت باقی مانده‌اند.

۴) افرادی که در طول بستری ۹۰ ساله بودند، از پایگاه داده حذف شدند.

۵) افرادی که بیش از ۸۹ سال سن داشتند، در اولین بستری، سنشان به صورت ۲۰۰ سال ذخیره شده است. یکی از علت‌های چنین تبدیلی، این است که برای افراد مظنون در این بازه سنی با کنار هم قرار دادن اطلاعاتی که برای آن‌ها ذخیره شده است، احتمال بیشتری وجود دارد که شناسایی شوند. در واقع این تبدیل برای محفوظ ماندن هویت این افراد انجام شده است.

۶) تبدیل تاریخ‌ها به صورت تصادفی انجام می‌شود.

۷) نام، نام خانوادگی، کدملی، استان و شهر محل سکونت، آدرس و شماره تلفن افراد حذف شده است و هر مظنون به وسیله یک کد شناساگر تعریف می‌شود.

هر مظنون ممکن است چندین مرتبه در تاریخ‌های مختلف با علائم بیماری تب کریمه‌کنگو به مرکز درمانی مراجعه کند. از آنجایی که اطلاعات افراد در پایگاه‌داده باید با کد شناساگر به هم مرتبط شوند، کد شناساگر برای هر فرد مظنون به صورت زیر تعریف می‌شود:

کد شناساگر Subject_ID: این کد برای شناسایی فرد مظنون است و در واقع همان شماره رکورد پزشکی وی است.

داده‌های مورد استفاده، شامل اطلاعات مربوط به ۹۶۵ فرد مظنون به این بیماری است که از این تعداد ۲۰۹ نفر، بیمار قطعی مبتلا به تب کریمه‌کنگو هستند. در این پایگاه‌داده ویژگی‌های مختلفی وجود دارد که در ادامه به شرح هر یک از آن‌ها و مقادیر قابل قبول برای هر کدام پرداخته می‌شود:

- علائم بالینی که شامل تب، اختلال هوشیاری یا کما، تهوع، استفراغ، اسهال، خون‌ریزی، سردرد، درد شکم، درد عضلات و شروع ناگهانی علائم هستند. خونریزی می‌تواند به شکل‌های مختلف خونریزی بینی، خونریزی لثه، خونریزی حفره شکم، خونریزی واژینال، راش‌های پتشی (لکه‌های خونی زیرپوستی)، محل تزریق سرم، خونریزی وسیع پوستی، استفراغ خونی، خلط خونی، ملنا (مدفوع سیاه)، خونریزی پشت حلق،

محل درآوردن کاتتر، جراحی شکم G. I. B، خونریزی از دهان، اکیموز پشت ساق پا، خونریزی و خون‌مردگی دور چشم، خونریزی از گوشه چشم، خون‌ریزی و خون‌مردگی دست، خونریزی ریه، خونریزی زیر پوستی، هماتوم پشت پا، هموتوراکس و خونریزی معده باشد.

- یافته‌های اولیه آزمایشگاهی که شامل افزایش آنزیم‌های کبدی، افزایش بیلی روبین توتال، کاهش هموگلوبین، Hematuria، Leukocytosis، Proteinuria و Leukopenia می‌باشد. جنسیت که می‌توان به دو صورت "مرد" یا "زن" باشد.
- شغل یکی از عوامل مؤثر در بیماری تب کریمه‌کنگو است. علت آن هم این است که افرادی که بیشتر با دام و طیور در ارتباط هستند، بیشتر در معرض ابتلا به این بیماری قرار دارند. شغل‌های مؤثر در این بیماری عبارت‌اند از "آشپز، پزشک، تکنسین دامپزشکی، چوپان، دامپزشک، دامدار، کشاورز، دباغ، راننده حمل گوشت و محصولات آن، قصاب، کارگر کشتارگاه و بسته‌بندی، راننده حمل دام و طیور، کارمند آزمایشگاه، کارمند بهداشتی–درمانی و کادر پرستاری". کسانی که شغلی به جزء مشاغل بالا دارند، کمتر در معرض ابتلا به این بیماری قرار دارند؛ بنابراین شغل به عنوان یک فاکتور مؤثر در این بیماری در نظر گرفته می‌شود.
- محل زندگی یکی دیگر از عوامل مؤثر در ابتلا به این بیماری است. به عبارت دیگر افرادی که در محیط‌های روستایی زندگی می‌کنند (یا به تازگی به آنجا سفر کرده‌اند)، احتمال ابتلای آن‌ها به این بیماری بیشتر می‌شود؛ بنابراین کسانی که در مناطق "عشایری" و "روستایی" زندگی می‌کنند و یا در دو هفته گذشته به این مناطق سفر داشته‌اند، احتمال ابتلای آن‌ها به این بیماری نسبت به افرادی که در مناطق "شهری" زندگی می‌کنند، افزایش پیدا می‌کند. براساس پایگاه داده مورد بررسی، بیشتر افراد مظنون به این بیماری به ترتیب از استان‌های سیستان بلوچستان، خراسان رضوی و کرمان بودند.

جدول ۱ ویژگی‌های مورد استفاده برای دسته‌بندی بیماران را نشان می‌دهد که صفت‌های ۱ تا ۱۸، ویژگی‌های بالینی و صفت‌های ۱۹ تا ۲۵ یافته‌های اولیه آزمایشگاهی هستند.

جدول ۱: ویژگی‌های اطلاعاتی مورد استفاده

ردیف	نام صفت	معادل به کار رفته	توضیحات	مجموعه مقادیر
۱	تب	fv	تب بیش از ۳۸ درجه سانتی‌گراد حداقل برای یک بار	۱ = دارد ۰ = ندارد
۲	خون‌ریزی	bld	خونریزی یا تمایل به خونریزی	۱ = دارد ۰ = ندارد
۳	شروع ناگهانی علائم	sm		۱ = دارد ۰ = ندارد
۴	تهوع	ns		۱ = دارد ۰ = ندارد
۵	استفراغ	vmt		۱ = دارد ۰ = ندارد
۶	درد عضلات	msp		۱ = دارد ۰ = ندارد
۷	سردرد شدید	shd		۱ = دارد ۰ = ندارد
۸	اختلال هوشیاری یا کما	dcs		۱ = دارد ۰ = ندارد
۹	اسهال	drh		۱ = دارد ۰ = ندارد
۱۰	درد شکم	stmch		۱ = دارد ۰ = ندارد
۱۱	کبودی ساق پا	bruise		۱ = دارد ۰ = ندارد
۱۲	سوزش ادراری	urthrts		۱ = دارد ۰ = ندارد
۱۳	دیسترس تنفسی	resdis		۱ = دارد ۰ = ندارد
۱۴	تعریق–لرز	swt		۱ = دارد ۰ = ندارد
۱۵	کاهش اشتها	rdapt		۱ = دارد ۰ = ندارد

جدول ۱: ویژگی‌های اطلاعاتی مورد استفاده (ادامه)

۱۶	افت فشارخون	lbprs		۱ = فشارخون سیستولیک کمتر از 9mmHg ۰ = فشارخون سیستولیک بیش از 9mmHg
۱۷	ضعف و بی‌حالی	wkax		۱ = دارد ۰ = ندارد
۱۸	کاهش وزن	wls		۱ = بیش ۵۰۰ گرم در هفته ۰ = کمتر از ۵۰۰ گرم در هفته
۱۹	افزایش آنزیم‌های کبدی	lenzy		۱ = هشت تا ده برابر مقدار طبیعی ۰ = کمتر از ۸ برابر مقدار طبیعی
۲۰	افزایش بیلی روبین توتال	tblrb	یکی از رنگ‌دانه‌های زردرنگ صفراوی است که عامل رنگ زرد ادرار و رنگ قهوه‌ای مدفوع است	۱ = طبیعی (۱/۲-۰/۱ در لیتر) ۰ = غیر طبیعی (بیش از ۱/۲ در لیتر)
۲۱	کاهش هموگلوبین	hmglob		۱ = هموگلوبین کمتر از ۷ گرم در دسی‌لیتر ۰ = هموگلوبین بیش از ۷ گرم در دسی‌لیتر
۲۲	Hematuria	hemtr	ادرار خونی	۱ = دارد ۰ = ندارد
۲۳	Leukocytosis	lkcts	افزایش تعداد گلبول‌های سفید در گردش خون	۱ = Leukocytosis بیش از ۹۰۰۰ در میلی‌متر مکعب ۰ = Leukocytosis کمتر از ۹۰۰۰ در میلی‌متر مکعب
۲۴	Proteinuria	prtnr	وجود مقادیر غیر طبیعی پروتئین در ادرار	۱= بیش از یک گرم در روز ۰ = کمتر از یک گرم در روز
۲۵	Leukopenia	kpn	کاهش تعداد گلبول‌های سفید در گردش خون	۱ = Leukopenia کمتر از ۳۰۰۰ در میلی‌متر مکعب ۰ = Leukopenia بیش از ۳۰۰۰ در میلی‌متر مکعب
۲۶	جنسیت	sex	می‌توان به دو صورت "مرد" یا "زن" باشد.	۱ = مرد ۰ = زن
۲۷	شغل	Job	در صورتی که یکی از شغل‌های فوق‌الذکر از جمله آشپز، چوپان و دام‌پزشک داشته باشد به آن عدد ۱ و در غیر این صورت عدد ۰ به آن تعلق می‌گیرد.	۱ = شغل‌های فوق‌الذکر ۰ = سایر مشاغل
۲۸	محل زندگی یا مسافرت	loc	محل زندگی و یا سفر که می‌تواند به دو صورت شهری و یا روستایی/عشایری باشد.	۱ = عشایری یا روستایی ۰ = شهری

ج) مدل‌سازی

بسترهای داده‌ای که دارای ابعاد (تعداد ویژگی) زیادی هستند، علی‌رغم فرصت‌هایی که به وجود می‌آورند، چالش‌های محاسباتی زیادی را ایجاد می‌کنند. یکی از مشکلات داده‌های با ابعاد زیاد این است که در بیشتر مواقع تمام ویژگی‌های داده‌ها برای ساختن دانشی که در داده‌ها نهفته است، مهم و حیاتی نیستند. به همین دلیل در بسیاری از زمینه‌ها کاهش ابعاد داده یکی از مباحث مهم است. در اغلب موارد بسیاری از ویژگی‌های کاندید برای کار یادگیری، نامربوط یا زائد هستند و کارایی به کارگیری الگوریتم یادگیری را خراب خواهند کرد و به مشکل بیش برازش منجر می‌شود. به عبارتی، دقت یادگیری و سرعت آموزش ممکن است به میزان درخور توجهی با این ویژگی‌ها زائد بدتر شود؛ بنابراین، انتخاب ویژگی‌های مرتبط و ضروری در مرحله پیش‌پردازش از اهمیتی بنیادین برخوردار است. یک انتخاب ویژگی مناسب کارایی یک مدل استنتاجی را افزایش می‌دهد. در این پژوهش نیز هدف ما، انتخاب زیرمجموعه‌ای از ویژگی‌ها جهت استفاده از آن‌ها در ساخت درخت تصمیم است، به عبارت دیگر هدف، پیدا کردن یک زیرمجموعه‌ای از ویژگی‌ها با حداقل اندازه ممکن است که برای تشخیص افراد دارای بیماری تب کریمه‌کنگو دقت بالایی داشته باشند؛ بنابراین از روش‌های مبتنی بر انتخاب ویژگی استفاده می‌شود. در این پژوهش از روش انتخاب ویژگی برنامه‌ریزی درجه دوم که در [۲۷] آمده است استفاده می‌شود. نتایج ارزیابی نشان می‌دهد که این روش در مقایسه با بسیاری از روش‌های مطرح دیگر هم از نظر پیچیدگی زمانی و هم از نظر دقت از بهبود خوبی برخوردار است.

در این روش، انتخاب ویژگی به یک مسئله در حوزه برنامه‌ریزی درجه دو تبدیل می‌شود. در واقع یک تابع هدف با یک قسمت درجه دو و یک قسمت خطی تعریف می‌شود و هدف کمینه‌سازی این تابع است [۲۷، ۲۸]؛ بنابراین در حوزه برنامه‌ریزی درجه دو تابعی به فرم

$$\min_{x} \{\frac{1}{2} x^T Q X - F^T x\} \qquad (۱)$$

تعریف می‌شود. فرض می‌کنیم که تعداد نمونه‌ها N و تعداد ویژگی‌ها M باشد. در رابطه (۱)، x یک بردار M بعدی و Q یک ماتریس $M \times M$ با خصوصیات تقارن، مثبت و نیمه‌معین می‌باشد. بردار F نیز یک بردار M بعدی با درایه‌های غیرمنفی است. در مسئله انتخاب ویژگی، Q شباهت بین ویژگی‌ها و F میزان همبستگی هر ویژگی با کلاس هدف را نشان می‌دهد. بعد از این‌که بهینه‌سازی تابع انجام شد، مؤلفه‌های x وزن هر ویژگی را مشخص می‌کنند. ویژگی‌های با وزن بیشتر، متغیرهای بهتری برای ساخت مدل می‌باشند. از آنجایی که x_iها وزن هر ویژگی را نشان می‌دهند؛ بنابراین باید مقادیری مثبت باشند و حاصل جمع همه آن‌ها برابر یک باشد.

از آنجایی‌که دو قسمت خطی و درجه دو در رابطه (۱) ممکن است در مسائل مختلف، اهمیت متفاوتی داشته باشند؛ بنابراین فرم عمومی‌تر این عبارت، به صورت

$$\min_{x} \{\frac{1}{2}(1-\alpha)x^T Q x - \alpha F^T x\} \qquad (۲)$$

تعریف می‌شود. در رابطه (۲)، x، Q و F به همان صورت قبل تعریف می‌شوند و α مقداری بین صفر و یک است [۲۷]. در این تحقیق انتخاب ویژگی بر روی مجموعه داده‌های علائم بالینی افراد مظنون به بیماری تب کریمه‌کنگو انجام می‌شود. بردار F هر درایه برابر با فراوانی هر ویژگی می‌باشد و برای ماتریس Q از ضریب همبستگی پیرسون استفاده می‌شود.

فرمول پیرسون به صورت

$$p_{ij} = \frac{\mathrm{cov}(v_i, v_j)}{\sqrt{\mathrm{var}(v_i)\,\mathrm{var}(v_j)}} \qquad (۳)$$

تعریف می‌شود که در آن cov کوواریانس ویژگی‌ها و var واریانس هر ویژگی است [۲۷]. با توجه به رابطه (۳) هر درایه q_{ij} ماتریس Q به صورت مقدار قدرمطلق p_{ij} تعریف می‌شود.

پس از به دست آوردن وزن هر ویژگی، ۱۰ ویژگی بالینی با بالاترین وزن برای مرحله بعد انتخاب می‌شوند که به همراه ویژگی‌های آزمایشگاهی، جنسیت، شغل و محل زندگی، ۲۰ ویژگی ما را شکل می‌دهند. این ویژگی‌ها پس از مشاوره‌های بالینی با متخصصان مربوطه و تأیید آن‌ها، برای استفاده در مرحله بعد مورد استفاده قرار می‌گیرند.

در مرحله بعد الگوریتم درخت تصمیم C4.5 پیاده‌سازی شده است. برای ساخت مدل درخت تصمیم متغیرهای تب، خون‌ریزی، شروع ناگهانی علائم، حالت تهوع، استفراغ، درد عضلات، سردرد، اختلال هوشیاری یا کما، اسهال، درد شکم، افزایش آنزیم‌های کبدی، افزایش بیلی روبین توتال، کاهش هموگلوبین، Hematuria، Leukocytosis، Proteinuria، Leukopenia، جنسیت، شغل و محل زندگی به عنوان متغیرهای پیش‌گو تعیین شدند و متغیر سالم/بیمار نیز به عنوان متغیر هدف در نظر گرفته شد. سپس داده‌ها به نسبت ۸۰–۲۰ به ترتیب به دو مجموعه آموزشی و آزمون تقسیم شدند. از داده‌های مجموعه آموزشی به منظور ساخت مدل استفاده می‌شود و از داده‌های بخش آزمون، برای ارزیابی مدل استفاده می‌شود. در جدول ۲ تعدادی از قوانین ایجادشده توسط مدل پیشنهادی بیان شده است:

جدول ۲: تعدادی از قوانین ایجادشده توسط الگوریتم C4.5

ردیف	قوانین	احتمال ابتلا
۱	اگر فرد دارای تب و خونریزی باشد و شروع ناگهانی علائم نداشته باشد، آنگاه احتمال بیمار بودن وی برابر است با:	۶۴٪/۰۵
۲	اگر فرد شروع ناگهانی علائم داشته باشد، ولی کاهش هموگلوبین نداشته باشد، همچنین شغل وی در دسته شغل‌های پرخطر باشد، آنگاه احتمال بیمار بودن وی برابر است با:	۷۰٪/۱
۳	اگر فردی افزایش بیلی روبین توتال داشته باشد و Hematuria و تب نداشته باشد، آنگاه احتمال بیمار بودن وی برابر است با:	۲۰٪/۹۵
۴	اگر فردی افزایش آنزیم‌های کبدی داشته باشد، همچنین تهوع و استفراغ نیز داشته باشد، آنگاه احتمال بیمار بودن وی برابر است با:	۷۸٪/۳۲
۵	اگر فردی دارای Leukopenia و Proteinuria باشد و در مناطق روستایی زندگی کند، آنگاه احتمال بیمار بودن وی برابر است با:	۸۹٪/۱۲
۶	اگر فردی دارای Hematuria باشد و در مناطق شهری زندگی کند. همچنین سردرد نیز نداشته باشد، آنگاه احتمال بیمار بودن وی برابر است با:	۳۲٪/۱۴
۷	اگر فردی اسهال باشد و آنزیم‌های کبدی وی افزایش پیدا کرده باشد، آنگاه احتمال بیمار بودن وی برابر است با:	۶۵٪/۵۶
۸	اگر فردی درد عضلات داشته باشد و بیلی روبین توتال وی افزایش پیدا کند، آنگاه احتمال بیمار بودن وی برابر است با:	۶۳٪/۸۵
۹	اگر فردی Hematuria داشته باشد و بیلی روبین توتال وی نرمال باشد. همچنین در مناطق روستایی زندگی کند، آنگاه احتمال بیمار بودن وی برابر است با:	۵۶٪/۹۴
۱۰	اگر آنزیم‌های کبدی و بیلی روبین توتال فردی افزایش پیدا کند و شغلی به جز شغل‌های پرخطر داشته باشد، آنگاه احتمال بیمار بودن وی برابر است با:	۶۹٪/۵۴
۱۱	اگر هموگلوبین فردی کاهش یافته باشد و شغل پرخطر داشته باشد، آنگاه احتمال بیمار بودن وی برابر است با:	۸۴٪/۲۵
۱۲	اگر فرد مذکری شغل پرخطر داشته باشد و Leukopenia هم داشته باشد، آنگاه احتمال بیمار بودن وی برابر است با:	۸۶٪/۳۵
۱۳	اگر فردی دارای Proteinuria باشد و در روستا زندگی کند، آنگاه احتمال بیمار بودن وی برابر است با:	۶۴٪/۲۱
۱۴	اگر فردی مذکری دارای Leukocytosis باشد و در مناطق شهری زندگی کند، احتمال بیمار بودن وی برابر است با:	۳۷٪/۲۵

د) ارزیابی مدل

پس از ایجاد مدل پیشنهادی، در این بخش به ارزیابی آن می‌پردازیم. برای محاسبه معیارهای ارزیابی، داده‌ها به دو مجموعه آموزشی و آزمون تقسیم شده‌اند. حال با استفاده از داده‌های مجموعه آزمون به ارزیابی اثربخشی مدل می‌پردازیم. جهت این منظور از معیارهای حساسیت، تشخیص، (Class Weighted Accuracy) CWA، مقدار پیش‌بینی مثبت و مقدار پیش‌بینی منفی استفاده شده است.

معیار حساسیت: معیار حساسیت در واقع قابلیت دسته‌بند را در تشخیص درست فرد بیمار نشان می‌دهد که رابطه آن به صورت:

$$Sensitivity = \frac{TP}{TP + FN} \quad (۴)$$

تعریف می‌شود [۲۹]. در رابطه (۴)، مثبت صحیح (True Positive) TP تعداد مواردی است که دسته‌بند به درستی فرد بیمار را بیمار تشخیص داده است و منفی کاذب (False Negative) FN تعداد مواردی است که دسته‌بند یک فرد بیمار را به اشتباه، سالم تشخیص داده است.

معیار تشخیص: معیار تشخیص، قابلیت دسته‌بند در تشخیص درست افراد سالم را نشان می‌دهد که رابطه آن به صورت:

$$Specificity = \frac{TN}{TN + FP} \quad (۵)$$

تعریف می‌شود [۲۹]. در رابطه (۵)، منفی صحیح (True Negative) TN تعداد مواردی است که دسته‌بند به درستی فرد سالم را سالم تشخیص داده است و مثبت کاذب (False Positive) FP تعداد مواردی است که دسته‌بند یک فرد سالم را به اشتباه بیمار تشخیص داده است.

معیار CWA: با این حال، در سیستم‌های پزشکی، ترکیب وزن‌دار معیار حساسیت و تشخیص اهمیت دارد [۲۹]؛ به عبارت دیگر، از آنجاکه هدف، تشخیص بیماری در افراد مظنون به بیماری تب کریمه‌کنگو است، مدل پیشنهادی باید به گونه‌ای باشد که FN به حداقل برسد؛ زیرا در این مسئله، برچسب زدن اشتباه افراد بیمار به سالم بسیار خطرناک‌تر از زدن برچسب بیمار به فرد سالم است، بنابراین از معیار CWA نیز برای ارزیابی دسته‌بندها استفاده شده است. این معیار، ترکیبی وزن‌دار از حساسیت و تشخیص است که در [۲۹] ارائه شده است. برای دسته‌بندی دودویی معیار CWA به فرم

$$CWA = w \times Sensitivity + (1 - w) \times Specificity \quad (۶)$$

تعریف می‌شود [۲۹]. در رابطه (۶)، w وزنی است که برای کلاس مثبت یا همان برچسب بیمار مبتلا به تب کریمه‌کنگو در نظر گرفته می‌شود. وزن w مقداری حقیقی بین صفر و یک است. در این تحقیق برای w مقدار ۰/۹ در نظر گرفته شده است.

مقدار پیش‌بینی منفی: این معیار، نسبت تعداد افرادی را که دسته‌بند به درستی سالم تشخیص داده است به تعداد کل افرادی که (خواه به درستی و خواه به اشتباه) سالم تشخیص داده شده‌اند، نشان می‌دهد.

$$TNR = \frac{TN}{TN + FN} \quad (۷)$$

مقدار پیش‌بینی مثبت: این معیار، نسبت تعداد افرادی را که دسته‌بند به درستی بیمار تشخیص داده است به تعداد کل افرادی که (خواه به درستی و خواه به اشتباه) بیمار تشخیص داده شده‌اند، نشان می‌دهد.

$$TPR = \frac{TP}{TP + FP} \quad (۸)$$

با توجه به داده‌های آزمایشی داریم:

$$TP = 199, FN = 10, TN = 378, FP = 378$$

نتایج ارزیابی روش پیشنهادی بر روی داده‌های افراد مظنون به بیماری در جدول ۳ آمده است.

جدول ۳: نتایج دسته‌بندی داده‌های آزمایشی

دسته‌بند	معیارهای مورد بررسی				
	حساسیت	تشخیص	CWA ($w = 0.9$)	مقدار پیش‌بینی منفی	مقدار پیش‌بینی مثبت
درخت تصمیم	۹۵٪	۵۰٪	۹۰٪/۵	۹۷٪	۳۴٪

همچنین زمان لازم برای ساخت درخت و دسته‌بندی نمونه‌های آزمایشی نیز برابر با ۵/۳۴ ثانیه بود. مشخصات سیستم آزمایش را در ادامه مشاهده می‌کنید:

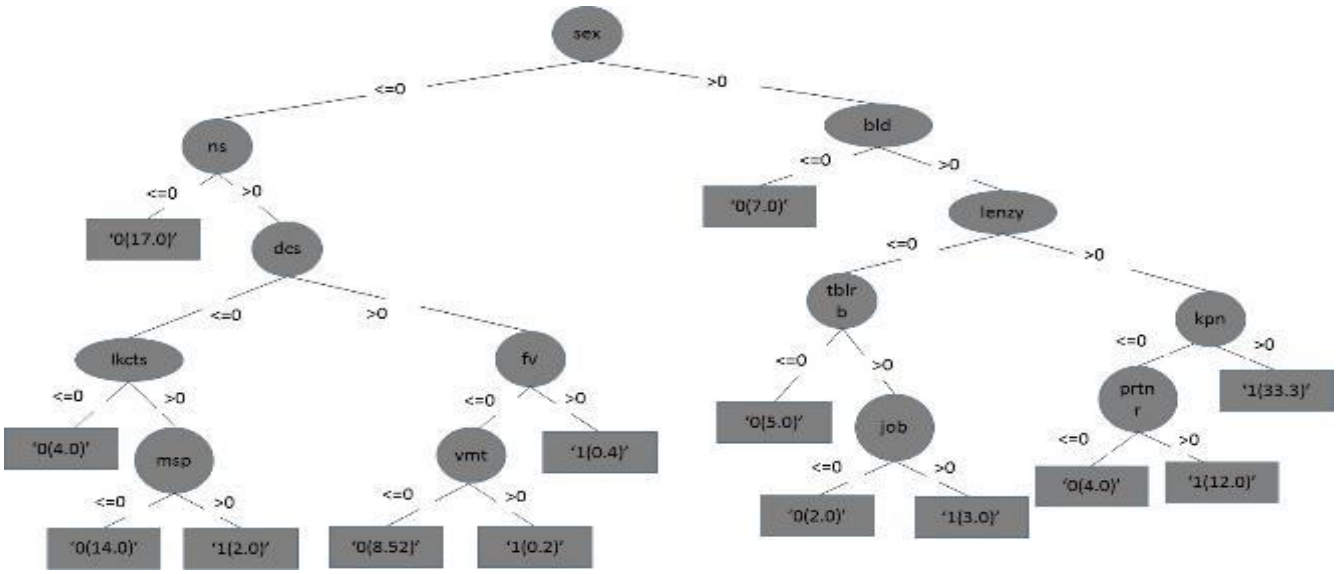
برای اجرای برنامه، از سیستمی با مشخصات زیر استفاده شده است:

- پردازنده Intel® Xeon® Processor X5650 با ۷ هسته پردازشی با توان ۲/۷ گیگاهرتز و دوازده نخ سخت‌افزاری
- ۳۲ گیگابایت حافظه اصلی
- ۲ ترابایت فضای ذخیره سازی دیسک
- سیستم‌عامل ویندوز ۱۰

برای پیاده‌سازی دسته‌بند نیز از ابزارهای Matlab، نسخه R2016a استفاده شده است.

نتایج

با توجه به مدل استفاده شده مشخص شد که متغیرهای تب، خون‌ریزی، شروع ناگهانی علائم، افزایش آنزیم‌های کبدی، افزایش بیلی روبین توتال، کاهش هموگلوبین، Hematuria، Leukocytosis، Proteinuria و Leukopenia بیشترین تأثیر را در تشخیص درست این بیماری دارند و متغیرهای سردرد، اسهال، تهوع و استفراغ کم‌ترین تأثیر را دارند. به کمک درخت تصمیم ایجادشده، قوانینی استخراج شده است که می‌تواند به عنوان الگویی در جهت تشخیص بیماری تب کریمه‌کنگو استفاده شود. قسمتی از درخت تصمیم به دست آمده در شکل ۲ نشان داده شده است.



شکل ۲: قسمتی از درخت تصمیم ایجادشده

براساس ساختار درخت تصمیم همان طور که در شکل ۲ آمده است، تعدادی از قوانین به صورت زیر هستند:

- مردانی که خونریزی داشته باشند، بیلی روبین توتال آن‌ها در حد طبیعی باشد و شغل آن‌ها یکی از شغل‌های مستعد بیماری باشد، احتمال ابتلای آن‌ها به بیماری تب کریمه‌کنگو وجود دارد.
- مردانی که آنزیم‌های کبدی آن‌ها ۸ تا ۱۰ برابر مقدار طبیعی خود باشد، Proteinuria آن‌ها بیش از ۱ گرم در روز باشد، احتمال ابتلای آن‌ها به بیماری تب کریمه‌کنگو وجود دارد.
- زنانی که حالت تهوع نداشته باشند، احتمال ابتلای آن‌ها به بیماری تب کریمه‌کنگو کمتر است.

- زنانی که حالت تهوع و تب داشته باشند و اختلال هوشیاری داشته باشند، احتمال ابتلای آن‌ها به بیماری تب کریمه‌کنگو وجود دارد.

برای انتخاب قوانین معتبر برای استفاده در دنیای واقعی، نظر متخصصان نیز در رابطه با قوانین استخراج‌شده، اعمال می‌شود؛ بنابراین این قوانین به متخصص و کارشناس مربوطه ارائه شدند و پس از مشاوره‌های بالینی، قوانینی که از نظر بالینی معتبر و کاربردی هستند، به عنوان قوانین نهایی برای استفاده معرفی شدند. برخی از مهم‌ترین این قوانین به شرح زیر است:

۱- در ۶۴٪ افراد بیمار، افزایش آنزیم‌های کبدی و Leukocytosis با هم مشاهده شده است.

۲- در ۶۲٪ بیماران مذکر، Leukopenia دیده شده است.

۳- ۸۷٪ از بیمارانی که در مناطق روستایی زندگی کرده‌اند، شغل‌های پرخطر داشته‌اند.

۴- در ۹۴٪ از بیمارانی که افزایش آنزیم‌های کبدی و Proteinuria وجود داشته است، ابتلا به بیماری تب کریمه‌کنگو دیده شده است.

۵- در ۵۴٪ از بیماران مونث که Proteinuria داشته‌اند، احتمال ابتلا به بیماری دیده شده است.

۶- در ۶/۵٪ از بیمارانی که سردرد داشته‌اند، احتمال ابتلا به بیماری دیده شده است.

۷- در ۱۰٪ بیماران مونث دارای Hematuria، احتمال ابتلا به بیماری دیده شده است.

بحث و نتیجه‌گیری

کاربرد کامپیوتر در زمینه‌های مباحث زیست‌پزشکی و به خصوص در تشخیص و پیش‌بینی بیماری‌ها اثبات شده است. سرعت بالای کامپیوتر در استخراج اطلاعات مفید از حجم عظیمی از داده‌ها و امکان پردازش آن‌ها در سریع‌ترین زمان ممکن، یکی از مهم‌ترین دلایل استفاده از کامپیوتر و مباحث مرتبط با آن در این زمینه است. از طرف دیگر این روزها با گسترش اینترنت و فضای وب، روز به روز به اطلاعات موجود افزوده می‌شود، کامپیوترها این امکان را دارند که بتوانند این حجم از داده‌ها را ذخیره کنند و از آن‌ها برای استخراج دانش استفاده کنند.

ما نیز در این تحقیق، پیرو پژوهش‌های پیشین در زمینه زیست پزشکی، به تشخیص بیماری تب کریمه‌کنگو با استفاده از الگوریتم‌های یادگیری ماشین پرداختیم. در ابتدا داده‌های مورد بررسی، براساس آزمایش‌های نهایی به برچسب سالم/بیمار نگاشت داده شدند. با این کار مسئله تشخیص بیماری در فرد مظنون به یک مسئله دسته‌بندی تبدیل شده و می‌توان از روش‌های دسته‌بندی برای مدل‌سازی آن استفاده کرد. سپس با استفاده از انتخاب ویژگی برنامه‌نویسی درجه دو ویژگی‌هایی که اثرگذاری بیشتری بر مدل دارند، انتخاب می‌شوند. در گام آخر از دسته‌بند درخت تصمیم به منظور دسته‌بندی داده‌ها استفاده شده است. نتایج ارزیابی نشان می‌دهد که استفاده از درخت تصمیم C4.5 در تشخیص بیماری تب کریمه‌کنگو دقت قابل قبولی داشته است. به صورت دقیق‌تر، معیار حساسیت مدل پیشنهادی برابر با ۹۵٪ و معیار تشخیص آن برابر با ۵۰٪ است که در حوزه داده‌کاوی پزشکی با الگوریتم درخت تصمیم، اثربخشی قابل قبولی است. همچنین مقدار پیش‌بینی مثبت مدل پیشنهادی برابر با ۳۴٪ و مقدار پیش‌بینی منفی برابر با ۹۷٪ است. در نتیجه می‌توان از این مدل به خوبی در غربالگری و تشخیص بیماری افراد مظنون در دنیای واقعی استفاده کرد. همچنین به منظور ارزیابی سایر الگوریتم‌های درخت تصمیم در مقایسه با الگوریتم C4.5، معیارهای ارزیابی در الگوریتم‌های دیگر نیز محاسبه شد که در جدول ۴ آورده شده است.

جدول ۴: مقایسه معیارهای ارزیابی الگوریتم درخت تصمیم C4.5 با سایر الگوریتم‌های درخت تصمیم

معیار	C4.5 (مدل پیشنهادی)	CHAID	QUEST	C&RT
حساسیت	۹۵٪	۸۹٪	۵۹٪	۸۶٪
تشخیص	۵۰٪	۷۶٪	۶۷٪	۵۶٪
CWA	۹۰٪/۵	۸۷٪	۶۰٪	۸۳٪
مقدار پیش‌بینی مثبت	۳۴٪	۵۰٪/۵	۳۳٪	۳۵٪
مقدار پیش‌بینی منفی	۹۷٪	۹۶٪	۸۵٪	۹۳٪/۵

طبق بررسی‌های انجام‌شده، پژوهش حاضر، اولین پژوهشی است که از درخت تصمیم و به طور کلی الگوریتم‌های داده‌کاوی برای تشخیص بیماری تب کریمه‌کنگو استفاده کرده است. به همین منظور در جدول ۵، مطالعات گذشته‌ای که از الگوریتم درخت تصمیم در جهت تشخیص سایر بیماری‌ها استفاده کرده‌اند، آمده است.

جدول ۵: مقایسه نتایج پژوهش‌انجام‌شده با سایر پژوهش‌های انجام‌شده در حوزه داده‌کاوی پزشکی

نویسنده	روش انتخابی	نوع بیماری	حساسیت
Chan [۳۰] (۲۰۰۸)	درخت تصمیم	دیابت	۶۴٪/۲۵
Shouman [۳۱] (۲۰۱۱)	درخت تصمیم	بیماری قلبی	۷۷٪/۹
مدل پیشنهادی	درخت تصمیم	تب کریمه‌کنگو	۹۵٪

عملکرد درخت تصمیم بر روی پایگاه‌های داده مختلف و ویژگی‌های مختلف نتایج متفاوتی دارد. برای مثال در تحقیق Chan و همکاران از ویژگی‌هایی از جمله سن، BMI، جنسیت و کلسترول برای پیش‌بینی بیماری دیابت استفاده کرده‌اند. آن‌ها روش پیشنهادی خود را بر روی دو پایگاه داده آزمایش کردند که میانگین معیار حساسیت آن‌ها از اجرای الگوریتم بر روی این دو پایگاه داده در جدول ۵ ذکر شده است [۳۰]. در تحقیق دیگری که توسط Shouman و همکاران انجام شد، از ویژگی‌هایی از جمله درد قفسه سینه، فشارخون در زمان استراحت، کلسترول و قندخون ناشتا برای تشخیص بیماری قلبی استفاده شده است [۳۱]. همان‌طور که در جدول ۵ مشاهده می‌شود، حساسیت ارائه‌شده در این پژوهش نسبت به سایر پژوهش‌های انجام‌شده قابل قبول است و از آن‌جایی که این پژوهش بر روی داده‌های استاندارد انجام شده است، مدل پیشنهادی می‌تواند در سیستم‌های لازم برای تشخیص بیماری تب کریمه‌کنگو به کار گرفته شود و نتایج حاصل از تحقیق به عنوان مبنای ارزیابی برای پژوهش‌های آتی استفاده شود. با توجه به کاربردهای وسیعی که کامپیوتر در علومی همچون پزشکی و زیست‌پزشکی ایفا می‌کند، می‌توان گفت تشخیص بیماری‌ها یکی از حوزه‌های مهم تحقیقاتی به شمار می‌رود. از طرفی به واسطه رابطه تنگاتنگ این موضوع با داده‌های بالینی و شرایط بیمار، پیچیدگی و چالش‌های متعددی بر سر راه آن قرار دارد. این چالش‌ها، در کنار ویژگی‌های خاص و چالش‌برانگیز تشخیص بیماری به جذابیت‌های تحقیق در این زمینه می‌افزایند. مطالعات آینده می‌توانند بر ایجاد، بهبود و ارتقای الگوریتم‌ها و تکنیک‌های موجود در زمینه تشخیص بیماری‌ها تمرکز کنند. برای بهبود مدل‌سازی و همچنین بالا بردن احتمال پذیرش این سیستم تصمیم‌یار توسط متخصصان، موارد زیر پیشنهاد می‌شود:

۱–برای این‌که سیستم در عمل قابل استفاده باشد و بعد از مدتی کنار گذاشته نشود، یکی از مواردی که در طراحی سیستم تصمیم‌یار اهمیت زیادی دارد نحوه جمع‌آوری و دریافت ورودی‌ها است. به این خاطر، نیاز است که تا حد امکان سیستم به صورت خودکار مقادیر ورودی را از سیستم‌های موجود در مراکز درمانی دریافت کند و فقط پیش از انجام عملیات تشخیص، لیست مقادیر ورودی را به متخصص نمایش دهد تا در صورت نیاز و در صورت وجود تناقض در متغیرها اصلاحات لازم را انجام دهد.

۲–مسئله دیگر بروز شدن سیستم تصمیم‌یار بعد از جمع‌آوری داده‌های جدید در بازه‌های زمانی معین می‌باشد. این مسئله نیز می‌تواند نقش مهمی در پذیرش و استفاده کاربردی از سیستم ایفا کند.

۳– استفاده از دیگر روش‌های یادگیری ماشین، از جمله روش‌های مبتنی بر خوشه‌بندی و یا ماشین بردار پشتیبان و الگوریتم نایو بیز است.

هیچ‌گونه کمک مالی از هیچ نهاد و ارگانی دریافت نشده است و هیچ تعارض منافعی وجود ندارد.

References

1. World Health Organization(WHO). Crimean-Congo haenorrhagic fever (CCHF). [cited 2017 Aug 2]. Available http://www.who.int/csr/disease/crimean_congoHF/en.

2. 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.

3. Soni J, Ansari U, Sharma D, Soni S. Predictive data mining for medical diagnosis: an overview of heart disease prediction. International Journal of Computer Applications 2011; 17(8): 43-8.

4. Subbalakshmi G, Ramesh K, Rao MC. Decision support in heart disease prediction system using naive bayes: Indian Journal of Computer Science and Engineering 2011; 2(2): 170-6.

5. Al Jarullah AA. Decision tree discovery for the diagnosis of type II diabetes. International Conference on Innovations in Information Technology; 2011 Apr 25-27; Abu Dhabi, United Arab Emirates: IEEE; 2011. p. 303-7.

6. Fang X. Are You Becoming a Diabetic? A Data Mining Approach. Sixth International Conference on

Fuzzy Systems and Knowledge Discovery; 2009 Aug 14-16; Tianjin, China: IEEE; 2009. p. 18-22.
7. Fayyad UM, Piatetsky-Shapiro G, Smyth P, Uthurusamy R. Advances in Knowledge Discovery and Data Mining. Menlo Park: AAAI press; 1996.
8. Lavrač N. Selected techniques for data mining in medicine. Artif Intell Med 1999;16(1):3-23.
9. Khajehei M, Etemady F. Data mining and medical research studies. Second International Conference on Computational Intelligence, Modelling and Simulation; 2010 Sep 28-30; Tuban, Indonesia: IEEE; 2010. p. 119-22.
10. Jayalakshmi T, Santhakumaran A. A novel classification method for diagnosis of diabetes mellitus using artificial neural networks. International Conference on Data Storage and Data Engineering; 2010 Feb 9-10; Bangalore, India: IEEE; 2010. p. 159-63.
11. 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.
12. Witten IH, Frank E, Hall MA, Pal CJ. Data Mining: Practical Machine Learning Tools and Techniques. 3th ed San Francisco: Morgan Kaufmann; 2016.
13. Han J, Kamber M, Pei J. Data Mining: Concepts and Techniques. 3th ed. San Francisco: Morgan Kaufmann; 2011.
14. Tanner L, Schreiber M, Low JG, Ong A, Tolfvenstam T, Lai YL, et al. Decision tree algorithms predict the diagnosis and outcome of dengue fever in the early phase of illness. PLoS Negl Trop Dis 2008;2(3):e196.
15. Shaukat K, Masood N, Mehreen S, Azmeen U. dengue fever prediction: a data mining problem. J Data Mining Genomics Proteomics 2015; 6(3):181.
16. Saha S, Saha S. Combined committee machine for classifying dengue fever. International Conference on Microelectronics, Computing and Communications (MicroCom); 2016 Jan 23-25; Durgapur, India: IEEE; 2016. p. 1-6.
17. Saikia D, Dutta JC. Early diagnosis of dengue disease using fuzzy inference system: International Conference on Microelectronics, Computing and Communications (MicroCom); 2016 Jan 23-25; Durgapur, India: IEEE; 2016. p. 1-6.
18. Olanow CW, Watts RL, Koller WC. An algorithm (decision tree) for the management of Parkinson's disease (2001): treatment guidelines. Neurology 2001;56(11 Suppl 5):S1-S88.
19. Cai Z, Gu J, Chen HL. A new hybrid intelligent framework for predicting Parkinson's disease. IEEE Access 2017; 5: 17188-200.
20. Chen HL, Wang G, Ma C, Cai ZN, Liu WB, Wang SJ. An efficient hybrid kernel extreme learning machine approach for early diagnosis of Parkinson′ s disease. Neurocomputing 2016; 131-44.
21. Wang W, Richards G, Rea S. Hybrid data mining ensemble for predicting osteoporosis risk: 27th IEEE Annual International Conference of the Engineering in Medicine and Biology Society; 2005 Sep 1-4; Shanghai, China, IEEE; 2005. p. 886-9.
22. Gao Z, Hong W, Xu Y, Zhang T, Song Z, Liu J. Osteoporosis Diagnosis Based on the Multifractal Spectrum Features of Micro-CT Images and C4.5 Decision Tree. First International Conference on Pervasive Computing, Signal Processing and Applications; 2010 Sep 17-19; Harbin, China: IEEE; 2010. p. 1043-47.
23. Tsien CL, Fraser HS, Long WJ, Kennedy RL. Using classification tree and logistic regression methods to diagnose myocardial infarction. Stud Health Technol Inform1998;52 Pt 1:493-7.
24. Bellaachia A, Guven E. Predicting Breast Cancer Survivability Using Data Mining Techniques. Washington: Washington University; 2006.
25. Su CT, Yang CH, Hsu KH, Chiu WK. Data mining for the diagnosis of type II diabetes from three-dimensional body surface anthropometrical scanning data. Computers and Mathematics with Applications 2006; 51(6): 1075-92.
26. 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 Z-G, Zhang H, Sun C. Advances in Neural Networks – ISNN 2007: 4th International Symposium on Neural Networks, ISNN 2007, Nanjing, China; 2007 Jun 3-7; Berlin Heidelberg: Springer; 2007. p. 1274-9.
27. Rodriguez-Lujan I, Huerta R, Elkan C, Cruz CS. Quadratic Programming Feature Selection. Journal of Machine Learning Research 2010; 11: 1491-1516.
28. Naqvi C. A hybrid filter-wrapper approach for feature selection [dissertation]. Orebro: Orebro University; 2012.
29. Cohen G, Hilario M, Sax H, Hugonnet S, Geissbuhler A. Learning from imbalanced data in surveillance of nosocomial infection. Artif Intell Med 2006;37(1):7-18.
30. Chan CL, Liu YC, Luo SH. Investigation of diabetic microvascular complications using data mining techniques. Proceedings of the 2008 International Joint Conference on Neural Networks; 2008 Jun 1-8; Hong Kong, China: IEEE; 2008. p. 830-4.
31. Shouman M, Turner T, Stocker R. Using decision tree for diagnosing heart disease patients: Proceedings of the Ninth Australasian Data Mining Conference; 2011 Dec 1–2; Ballarat, Australia: Australian Computer Society; 2011. p. 23-30.

Original Article

Journal of Health and Biomedical Informatics
Medical Informatics Research Center
2017; 4(2): 108-121

Detection of Crimean-Congo Fever Using C4.5 Decision Tree

Esmaeeli Gohari Reza[1], Esmaeeli Gohari Elham[2]*, Shafiei Mehdi[3]

• **Received:** 23 Jul, 2017 • **Accepted:** 18 Sep, 2017

Introduction: The prevalence of Crimean-Congo fever, a common disease between human and animal, shows an increasing rate by coming summer season. Detection of this disease by the use of necessary tests, lasts at least about one week. There are several data mining and machine learning techniques to create predictive models for identifying at risk people. In this study, C4.5 decision tree method has been used due to its simplicity and efficiency.

Methods: In this applied descriptive study, data related to suspected cases of Crimean-Congo fever were used. These data have been collected from health centers of Iran in a four-year period since 2014 and contained 965 records with 29 features. First, by using the quadratic programming feature selection method, the variables which were effective on the model were selected and then, the C4.5 decision tree model was created through using input variables and determining the target variable. Data analysis was performed through Matlab software.

Results: According to the applied model, it was found that fever, bleeding, sudden onset of symptoms, increased liver enzyms, increased total Bilirubin, decreased Hemoglobin, Hematuria, Leukocytosis, Proteinuria and Leukopenia have the greatest impact in the diagnosis of this disease.

Conclusion: According to the obtained results, the sensitivity of the proposed model is 95% and its specificity is 50%. Therefore, this model showed acceptable efficiency in diagnosing this disease in comparison with other studies done in medical data mining field.

Keywords: Medical decision support system, Disease diagnosis, Crimean-Congo hemorrhagic fever, C4.5 Decision tree

• **Citation:** Esmaeeli Gohari R, Esmaeeli Gohari E, Shafiei M. Detection of Crimean-Congo Fever Using C4.5 Decision Tree. Journal of Health and Biomedical Informatics 2017; 4(2): 108-121.

1. M. Sc. Student in Compueter Engineering, Computer Engineering Dept., Bahmanyar Higher Education Institute of Kerman, Kerman, Iran.
2. M. Sc. in Compter Engineering, Computer Engineering Det., Technical and Engineering Campus, Yazd University, Yazd, Iran
3. General Practitioner-MPH, Deputy for Health, Kerman University of Medical Sciences, Kerman, Iran.
***Correspondence**: Faculty electrical and computer engineering, Yazd University, Yazd, Iran.
• **Tel:** 09133427435 • **Email:** g.elhamesmaeeli@gmail.com