مجله دانشکده پیراپزشکی دانشگاه علوم پزشکی تهران (پیاورد سلامت)
دوره ۱۳ شماره ۱ فروردین و اردیبهشت ۱۳۹۸، ۹۰-۸۱

مقاله پژوهشی

مقایسه‌ی الگوریتم‌های داده کاوی برای تشخیص بیماری‌های کبد

محمدرضا شهرکی[۱]، محبوبه مسگر[۲]

دریافت مقاله : شهریور ۱۳۹۷
پذیرش مقاله : دی ۱۳۹۷

* نویسنده مسئول :
محبوبه مسگر؛
دانشکده مهندسی شهید نیکبخت دانشگاه سیستان و بلوچستان
Email :
mhb@pjs.usb.ac.ir

چکیده

زمینه و هدف: کبد به‌عنوان یکی از بزرگ‌ترین اندام‌های داخلی بدن، وظیفه‌ی انجام اعمال حیاتی مختلفی از جمله تصفیه و پالایش خون، تنظیم هورمون‌های بدن، ذخیره‌ی گلوکز و ... را در بدن به عهده دارد. بنابراین اختلال در کارکرد آن مشکلات گاه جبران‌ناپذیری به دنبال خواهد داشت. لذا پیش‌بینی به‌موقع این بیماری به درمان‌های اولیه و مؤثر آن کمک می‌کند. با توجه به اهمیت بیماری‌های کبد و افزایش تعداد مبتلایان، مطالعه‌ی حاضر با هدف پیش‌بینی بیماری‌های کبد با استفاده از الگوریتم‌های داده‌کاوی صورت گرفت.

روش بررسی: این پژوهش از نوع توصیفی بوده و با استفاده از ۷۲۱ داده‌ی جمع‌آوری‌شده از بیماران کبدی شهر زاهدان انجام‌شده است. در این بررسی پس از پیش‌پردازش داده‌ها، تکنیک‌های داده کاوی از قبیل ماشین بردار پشتیبان (Support Vector Machines)، CHAID، Exhaustive CHAID و C5.0 تقویت‌شده در نرم‌افزار IBM SPSS Modeler 18 بررسی، مقایسه و تحلیل‌شده است.

یافته‌ها: یافته‌ها نشان داد که الگوریتم C5.0 تقویت‌شده با دقت ۹٤/۰۹ درصد، الگوریتم Exhaustive CHAID با دقت ۸۸/۷۱ درصد، ماشین بردار پشتیبان با دقت ۸۷/۰۹ درصد و الگوریتم CHAID با دقت ۸۵/٤۷ درصد بیماری های کبد را پیش‌بینی کردند. بنابراین بهترین الگوریتم از لحاظ دقت عملکرد، الگوریتم C5.0 تقویت‌شده شناخته شد.

نتیجه‌گیری: با توجه به دقت الگوریتم C5.0 تقویت‌شده و قوانین حاصل از آن، برای یک نمونه‌ی جدید با ویژگی‌های مشخص، می‌توان احتمال ابتلای فرد به بیماری‌های کبد را با دقت قابل قبولی پیش‌بینی کرد.

واژه‌های کلیدی: بیماری‌های کبد، ماشین بردار پشتیبان (SVM)، الگوریتم CHAID، درخت تصمیم C5.0، داده‌کاوی

۱ استادیار گروه مهندسی صنایع، دانشکده مهندسی شهید نیکبخت، دانشگاه سیستان و بلوچستان، زاهدان، ایران

۲ دانشجوی کارشناسی ارشد مهندسی صنایع، دانشکده مهندسی شهید نیکبخت، دانشگاه سیستان و بلوچستان، زاهدان، ایران

مقدمه

با تغییر سبک زندگی افراد، شیوع بیماری‌ها به‌طور پیوسته در حال گسترش است. یکی از بیماری‌هایی که امروزه افراد بسیاری با آن دست‌وپنجه نرم می‌کنند، بیماری‌های کبد(مانند: کبدچرب، سیروز و سرطان کبد) می‌باشد. بیماری‌های کبد عامل حدوداً دو میلیون مرگ و میر در سراسر جهان به حساب می‌آید که میزان مرگ و میر از نظر نوع بیماری‌های کبد، بر اساس منطقه‌ی جغرافیایی، نژاد، جنسیت و قومیت متفاوت است(۱). کبد یکی از بزرگ‌ترین اندام‌های بدن انسان است که در سمت راست، بالای شکم و زیر دیافراگم قرار دارد(۳و۲). کبد خون را تصفیه کرده و با شناسایی مواد سمی آن‌ها را دفع می‌کند(٤). مشکلات کبدی باعث به وجود آمدن بیش از صد نوع بیماری در بدن انسان می‌شود؛ بنابراین می‌توان گفت هرگونه بیماری کبد و عوارض ناشی از آن تمامی بدن را تحت تأثیر قرار خواهد داد(۵). با توجه به اینکه تعداد افراد مبتلا به بیماری‌های کبد در حال افزایش است و شیوع بیماری‌های کبد از ۱۸ تا ۵۵ درصد در مطالعات ذکرشده است (۷-۵)، بنابراین تشخیص به‌موقع این بیماری‌ها می‌تواند در پیشگیری از عوارض، کنترل و درمان بسیار مؤثر باشد. لذا استفاده از روش‌های مختلف، جهت پیش‌بینی و تشخیص زودهنگام بیماری‌های کبد یکی از اولویت‌های اصلی پژوهش‌های علوم پزشکی می‌باشد.

امروزه با گسترش کاربردهای فناوری اطلاعات در حوزه پزشکی، جمع‌آوری و ذخیره‌سازی حجم زیادی از داده‌ها تسهیل شده و امکان به‌کارگیری روش‌های داده کاوی در این زمینه برای پژوهشگران فراهم‌شده است. روش‌های داده کاوی از الگوریتم‌های یادگیری برای ایجاد مدل‌های پیش‌بینی استفاده می‌کنند که با توجه به برخی عوامل نظیر ویژگی داده‌ها و حیطه‌ی کاربردی مدل‌ها، برخی الگوریتم‌ها نسبت به سایر آن‌ها در اولویت قرار می‌گیرند(۸). استفاده از ابزارهای هوشمند داده کاوی می‌تواند به عنوان تکنیکی برای شناسایی و تشخیص بیماری‌ها، دسته‌بندی بیماران در مدیریت بیماری، پیدا کردن الگوهایی برای تشخیص سریع‌تر بیماری و جلوگیری از بروز عوارض در آن‌ها کمک بسیار بزرگی باشد(۹).

الگوریتم‌های داده کاوی به دو دسته‌ی کلی نظارتی و غیرنظارتی و یا پیش‌بینی و توصیفی تقسیم می‌شوند. در الگوریتم‌های توصیفی، هدف، استخراج الگو از داده‌هاست که نیاز به تحلیل نتایج دارد. اما در الگوریتم‌های پیش‌بینی، هدف پیش‌بینی یک ویژگی خاص بر مبنای ویژگی‌های دیگر است(۱۰). یکی از عملکردهای پیش‌بینی، دسته‌بندی است. دسته‌بندی فرایند یافتن مدلی است که با تشخیص دسته‌ها یا مفاهیم داده می‌تواند دسته‌ی ناشناخته‌ی اشیای دیگر را پیش‌بینی کند(۱۱). مطالعات زیادی در زمینه‌ی پیش‌بینی در حوزه‌های مختلف پزشکی انجام شده است. یکی از این حوزه‌ها، پیش‌بینی بیماری‌های کبد می‌باشد. Abdar و همکاران، از برخی الگوریتم‌های داده کاوی برای تشخیص زودهنگام بیماری‌های کبد استفاده نمودند و نشان دادند که سن فرد، آلکالن فسفاتاز، بیلی روبین مستقیم، آسپارتات آمینوترانسفراز و نسبت آلبومین به گلوبولین مهم‌ترین فاکتورهای شناسایی‌شده توسط الگوریتم C5.0 جهت پیش‌بینی این بیماری است. بنابراین آنان استفاده از الگوریتم C5.0 بهبودیافته، با دقت ۹۳/۷۵ درصد را جهت پیش‌بینی بیماری‌های کبد توصیه کردند(۲). Ramana و همکاران، مطالعه‌ای تطبیقی بین مجموعه داده‌ی ILPD (Indian Liver Patient Dataset) و BUPA برگرفته از مخزن داده UCI (University of California Irvine) انجام دادند که در تمام الگوریتم‌های انتخابی، مجموعه داده‌ی ILPD دارای عملکرد بهتری نسبت به مجموعه داده‌ی BUPA بود، و در بین دسته‌بندهای انتخابی، KNN (K-nearest neighbor) بهترین نتیجه را بر روی مجموعه داده‌ی ILPD با ترکیب تمام ویژگی‌ها نشان داده است(۱۱). Yeh و همکاران در مقاله ای با عنوان پیش بینی بیماری کبدچرب با الگوریتم های یادگیری ماشین، از چهار مدل طبقه‌بندی(بیزی ساده، شبکه عصبی مصنوعی، رگرسیون لجستیک و جنگل تصادفی) برای پیش بینی بیماری استفاده کردند. نتایج آنان نشان داد که جنگل تصادفی دارای عملکرد بهتری در پیش‌بینی بیماری است(۱۲). هدف این مطالعه مقایسه الگوریتم‌های داده کاوی در پیش‌بینی بیماری‌های کبد و شناسایی الگوریتم کاراتر می‌باشد.

روش بررسی

این مطالعه توصیفی گذشته‌نگر است و داده‌های مورداستفاده، مربوط به اطلاعات بیماران یک آزمایشگاه در شهر زاهدان می‌باشد که داده‌ها به تفکیک هر سال کار آزمایشگاه از سال ۱۳۹٤ تا ۱۳۹٦ در دسترس ما قرارگرفته است. در مجموع داده‌های ۸۷٤ بیمار با ۱۰ ویژگی از جمله سن، جنسیت، بیلی‌روبین تام (TB)، بیلی‌روبین مستقیم

(DB)، آلکالن فسفاتاز (Alkphos)، آلانین آمینوترانسفراز (SGPT)، آسپارتات آمینوترانسفراز (SGOT)، پروتئین تام (TP)، آلبومین (ALB) و نسبت آلبومین به گلوبولین (A/G) از پایگاه داده‌ی مذکور استخراج شدند. اولین مرحله در ایجاد هر مدلی براساس تکنیک‌های داده‌کاوی، پیش‌پردازش (Preprocessing) می‌باشد. بنابراین در این مطالعه به‌منظور بهبود کیفیت داده‌ها، پیش‌پردازش در سه مرحله مقادیر گمشده، مقیاس بندی داده‌ها و دسته‌بندی داده‌ها انجام شد. ابتدا مقادیر داده‌ها از لحاظ وجود مقادیر گمشده (Missing) بررسی گردید. وجود این مقادیر تأثیر بسزایی در عملکرد الگوریتم‌های متعدد یادگیری ماشین دارد(۱۳). بسیاری از تکنیک‌های پردازش داده توسط Han و Kamber (۱٤) ارایه‌شده است. در این مطالعه مواردی را که ارزش صفر برای ۱۰ ویژگی پژوهش داشتند، حذف شدند. زیرا مرسوم‌ترین روش در برخورد با داده‌های گمشده، حذف آن‌هاست(۱۵). به‌علاوه Witten و Frank (۱٦) نیز ثابت کردند که حذف عاقلانه، یک روش کارآمد به‌جای جایگزین کردن ارزش‌ها با تکنیک‌هایی مانند میانگین، انتساب تصادفی، انتساب رگرسیون و مدل‌های بیزی است. بنابراین، تعدادی از رکوردهای داده به دلیل نقص در اطلاعات و یا غلط بودن اطلاعات حذف گردید. با این کار تعداد ۱۱۷ نفر حذف شدند. هم‌چنین رکوردهای تکراری متعلق به یک بیمار حذف‌شده، که تعداد ۳٦ نفر دارای چنین وضعیتی بودند. در نهایت تعداد داده‌های مورد بررسی به ۷۲۱ نفر رسید.

در مرحله‌ی بعد با توجه به اینکه متغیر هدف پژوهش یک متغیر طبقه‌ای است و هم‌چنین مقادیر ورودی در مقیاس‌های متفاوتی هستند و این مسئله، تأثیر منفی در روند همگرایی، افزایش زمان و تعداد دفعات تکرار اعمال آموزش خواهد گذاشت، به‌منظور دستیابی به نتیجه‌ی دقیق‌تر، مقیاس بندی یا نرمال‌سازی داده‌ها ضروری است. نرمال‌سازی داده‌ها باعث یکدست شدن و هماهنگ شدن داده‌ها در دامنه مورد بررسی می‌شوند و هم‌چنین باعث می‌شود که داده‌های با مقیاس بزرگ نتیجه را به سمت خویش منحرف نکنند. بنابراین در این مطالعه قبل از آموزش الگوریتم‌ها طبق رابطه ۱ از نرمال‌سازی Min-Max که یک تبدیل خطی بر روی داده‌های اصلی انجام می‌دهد، استفاده شده و تمامی مشاهدات عددی در بازه [۰،۱] نرمال‌سازی شدند.

$$Z_i = \frac{x_i - \min(x)}{\max(x) - \min(x)} \qquad (۱)$$

که در آن منظور از min(x) کمترین مقدار یک متغیر، max(x) بیشترین مقدار یک متغیر، x_i درایه مربوط به ماتریس داده‌ها و Z_i مقدار نرمال شده برای هر مشاهده می‌باشد. در مرحله‌ی آخر به‌منظور استفاده از داده‌ها در الگوریتم‌های داده‌کاوی، کل نمونه به دو دسته‌ی آموزش و آزمایش تقسیم‌بندی شد. بر این اساس از نمونه‌گیری تصادفی و با تخصیص ۷۰ درصد داده‌ها برای آموزش و ۳۰ درصد داده‌ها برای تست سیستم‌های طبقه‌بندی استفاده‌شده و تعداد دفعات تکرار آموزش–تست، ۱۰ مرتبه تعیین شد.

مجموعه کلاس‌ها به دو گروه مبتلا به بیماری‌های کبد و سالم تقسیم شدند که کلاس اول شامل ۲۰۳ نفر سالم و کلاس دوم شامل ۵۱۸ نفر بیمار می‌باشد. بنابراین متغیر هدف در این مطالعه، ابتلا یا عدم ابتلا به بیماری‌های کبدی است که در مورد هر کدام از افراد مورد بررسی یکی از این دو حالت ثبت گردیده است. مقدار یک برای این متغیر، نشان‌دهنده‌ی ابتلا به بیماری کبد و صفر، نشان‌دهنده‌ی عدم ابتلا به این بیماری است. در نهایت بعد از جمع‌آوری و ثبت داده‌ها، از نرم‌افزار IBM SPSS Modeler 18 برای تجزیه‌وتحلیل الگوریتم‌های ماشین بردار پشتیبان، CHAID، Exhaustive CHAID و C5.0 تقویت‌شده استفاده گردید. در این بخش از نمونه‌گیری تصادفی و با تخصیص ۷۰ درصد داده‌ها برای آموزش و ۳۰ درصد داده‌ها برای تست سیستم‌های طبقه‌بندی استفاده شده و اعتبار الگوریتم‌ها به روش اعتبارسنجی k-fold با k=۱۰ ارزیابی گردید.

● ماشین بردار پشتیبان (SVM)

ماشین بردار پشتیبان یکی از تکنیک‌های دسته‌بندی است که برای طیف متنوعی از مجموعه داده‌ها قابل‌استفاده است. در واقع فلسفه‌ی ماشین بردار پشتیبان این است که برای تفکیک داده‌هایی با ساختار پیچیده و غیرخطی، داده‌ها توسط توابع ریاضی کرنل (Kernel)، در فضای جدیدی نگاشت شوند(۱۷). بنابراین این روش برای انجام عمل پردازش از توابع کرنل استفاده می‌کند. کرنل نوع خط مرز تصمیم را نشان می‌دهد و چهار تابع خطی، چندجمله‌ای، حلقوی و تابع پایه‌ای شعاعی را می‌پذیرد و با تغییر توابع، نتایج متفاوتی حاصل می‌گردد (۱۸). لذا الگوریتم ماشین بردار پشتیبان در قالب مسئله‌ی بهینه‌سازی با تابع هدف(رابطه ۲) و محدودیت‌های(رابطه ۳) بیان می‌شود.

$$maximize_{\alpha}\left[\sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i,j=1}^{n}\alpha_i\alpha_j y_i y_j K(X_i, X_j)\right] \quad (2)$$

$$\text{Subject to} \quad \sum_{i=1}^{n}\alpha_i y_j = 0 \qquad 0 \le \forall\alpha_i \ll L \quad (3)$$

که X بردار یادگیری، y برچسب مرتبط با هر بردار یادگیری، α بردار پارامترهای دسته‌بندی کننده فراصفحه، K تابع کرنل(تابع کرنل که باعث می‌شود مدل بتواند فضاهای غیرخطی پیچیده‌تر را هم پوشش دهد) و L پارامتر جریمه برای کنترل تعداد دسته‌بندی‌های اشتباه می‌باشد. اگر L بی‌نهایت باشد، الگوریتم به دسته‌بندی‌های اشتباه جریمه‌ی بی‌نهایت می‌دهد و از وقوع خطا در دسته‌بندی جلوگیری می‌کند. بزرگ بودن L دقت دسته‌بندی را بالا می‌برد و به‌تبع آن زمان محاسبات نیز افزایش می‌یابد درحالی‌که مقادیر کمتر از L انعطاف‌پذیری بیشتری روی دسته‌بندی ایجاد می‌کند.

ازآنجایی‌که راه ساده‌ای برای دانستن اینکه کدام تابع با مجموعه داده، بهترین عملکرد را دارد، وجود ندارد؛ در این مطالعه توابع مختلف به‌نوبت انتخاب و نتایج با یکدیگر مقایسه شده است. در روش‌های دسته‌بندی یک فیلد به‌عنوان فیلد خروجی در نظر گرفته می‌شود که در این تحقیق بیمار یا سالم بودن افراد به‌عنوان خروجی و مشخصات آزمایشگاهی و دموگرافیک به‌عنوان ورودی در نظر گرفته‌شده است. برای ارزیابی ماشین بردار پشتیبان داده‌ها به زیرمجموعه‌های آموزش و تست تقسیم گردید. سپس داده‌های بیماران کبد با چهار تابع کرنل الگوریتم، مورد آزمون قرارگرفتند. پارامترهای دیگر ماشین بردار پشتیبان با تابع کرنل چندجمله‌ای شامل Regularization Parameter، Regression Precision (epsilon)، RBF gamma، Bias و Degree است که در این مطالعه به ترتیب دارای مقادیر ۱۸، ۰/۰۵، ۱/۵، ۱/۳ و ۵ هستند. در نهایت مشخص شد که هیچ‌یک از توابع کرنل نمی‌تواند نتایج را به‌خوبی تابع چندجمله‌ای برای این مجموعه داده به‌دست آورد.

● الگوریتم C5.0

الگوریتم C5.0 برای ساخت درخت تصمیم (Decision Tree) یا مجموعه قوانین (Rule Set) استفاده می‌شود. این الگوریتم با تجزیه‌ی داده بر اساس ویژگی‌ها، حداکثر بهره از اطلاعات را می‌برد. هر زیر بخش به بخش‌های کوچک‌تری تقسیم‌شده و مجدداً زیر بخش‌ها نیز به بخش‌های کوچک‌تر تقسیم می‌شوند و این تقسیم‌ها تا جایی ادامه می‌یابد که تقسیم‌بندی جدیدی انجام نشود. نهایتاً پایین‌ترین بخش‌بندی‌ها مجدداً امتحان شده و چنانچه کمک قابل‌توجهی برای ارزش مدل نداشته باشند، حذف یا هرس می‌گردند. الگوریتم C5.0، نسبت به سایر مدل‌ها درک آسانی دارد خصوصاً موقعی که قواعد به‌دست آمده از مدل، تفسیر مشخصی دارند. C5.0 همچنین با استفاده از روش boosting قابلیت افزایش دقت در دسته‌بندی را دارد(۱۴). در این مطالعه، الگوریتم C5.0 بر روی اطلاعات ۷۲۱ بیمار کبدی اعمال گردید. جهت آموزش مدل، ۱۰ متغیر(سن، جنسیت، بیلی‌روبین تام، بیلی‌روبین مستقیم، آلکالن فسفاتاز، آلانین آمینوترانسفراز، آسپارتات آمینوترانسفراز، پروتئین تام و نسبت آلبومین به گلوبولین) به‌عنوان ورودی و یک متغیر(بیمار یا سالم بودن فرد) به‌عنوان خروجی در نظر گرفته شد.

● الگوریتم CHAID

این الگوریتم با مدل‌سازی روابط بین یک متغیر وابسته و تعداد زیادی متغیر مستقل به مطالعه‌ی رابطه‌ی میان آن‌ها می‌پردازد. آزمون آماری مورداستفاده بستگی به سطح اندازه‌گیری متغیر هدف دارد. اگر متغیر هدف پیوسته باشد، آزمون F و اگر دسته‌ای باشد، آنالیز استقلال کای دو (Chi-Square) به‌کاربرده می‌شود(۱۴). در این مطالعه با توجه به اینکه متغیر هدف، از نوع دسته‌ای است، از آزمون استقلال کای دو استفاده‌شده است. بنابراین ابتدا CHAID اهمیت هر یک از متغیرهای پیش‌بینی‌کننده را برای پیش‌بینی متغیر هدف پژوهش با توجه به آزمون استقلال کای دو تعیین می‌کند. این الگوریتم، متغیرهای تأثیرگذار در پیش‌بینی متغیر هدف را براساس P-Value شناسایی می‌کند(کوچک‌ترین مقدار p). برای کاهش تورش در مدل‌سازی استفاده از روش اعتبارسنجی K-Fold برای این تفکیک توصیه‌شده است(۹). در این مطالعه، روش اعتبارسنجی K-Fold، مقدار k برابر پنج درنظرگرفته شد، یعنی رکوردها به صورت تصادفی به پنج گروه نسبتاً مساوی تقسیم گردید و پنج بار عملیات آموزش (Train) برای چهار گروه و عملیات آزمون (Test) برای گروه پنجم تکرار گردید. در نهایت الگوریتم درخت تصمیم روی زیرمجموعه آموزش، اعمال و مدل پیش‌بینی ایجاد شد.

● الگوریتم Exhaustive CHAID

Exhaustive CHAID یک نسخه‌ی اصلاح‌شده از CHAID است که برای برطرف کردن برخی ضعف‌های روش CHAID توسعه‌یافته است. در واقع Exhaustive CHAID تمرکز و دقت بیشتری برای پیدا کردن کلیه گسست‌های ممکن دارد. در این روش پس از پیدا کردن کلیه گسست‌ها، ادغام دسته‌ها تا زمانی‌که تنها دو دسته برای هر پیش‌بینی کننده باقی بمانند، ادامه می‌یابد. پس‌ازآن به گام انتخاب گسست متغیر رفته و از بین پیش‌بینی کننده‌ها، آن موردی را که بیش‌ترین گسست مشهود را دارد، انتخاب می‌کند. Exhaustive

CHAID با CHAID در آزمون‌های آماری مورداستفاده و نحوه‌ی برخورد با مقادیر مفقودی یکسان است، اما روش این الگوریتم در ترکیب دسته‌های متغیرها، دقیق‌تر از CHAID است و برای داده‌های زیاد و هم‌چنین داده‌های دارای متغیرهای پیش‌بینی پیوسته، به زمان محاسبه‌ی بیشتری نیاز دارد. بنابراین اگر زمان کافی وجود داشته باشد، Exhaustive CHAID به‌طور عمومی مطمئن‌تر و دقیق‌تر است. این روش اغلب، بخش‌های مفید بیشتری را می‌یابد، هرچند که ممکن است تفاوتی بین نتایج این دو نباشد(۱۹).

- **ارزیابی الگوریتم‌ها**

الگوریتم‌ها دارای دو مرحله‌ی آموزش و تست هستند. در مرحله‌ی آموزش مدلی ایجاد گردیده که با استفاده از این مدل، پیش‌بینی وضعیت سایر نمونه‌ها امکان‌پذیر می‌شود و عملکرد آن در مرحله‌ی دوم توسط مجموعه تست سنجیده می‌شود. در مرحله‌ی تست، هدف تخمین کارایی الگوریتم از جنبه‌های مختلف است(۹).

در روش دسته‌بندی، مجموعه‌ای از داده‌ها به نام داده‌های آموزشی وجود دارند که از پیش طبقه‌بندی شده و دارای برچسب‌های مشخصی هستند. هدف یافتن روش، تابع و یا قوانینی بر اساس ویژگی‌های داده‌های آموزشی جهت طبقه‌بندی داده‌هایی است که در آینده وارد سیستم می‌شوند. در روش دسته‌بندی یکی از ویژگی‌های داده‌ها همان برچسب دسته یا متغیر هدف است و سایر ویژگی‌ها، متغیرهای پیش‌بینی‌کننده نام دارند و هدف، پیش‌بینی مقدار متغیر هدف بر اساس متغیرهای پیش‌بینی‌کننده می‌باشد. در این مقاله با استفاده از داده‌های بخش آموزش(۷۰ درصد) مدل تولید گردیده و داده‌های بخش تست(۳۰ درصد) مدل تولیدشده را آزمون و برچسب مربوط به رکوردهای مذکور را تعیین می‌نمایند. برای آموزش، متغیر طبقه‌ای بیمار یا سالم بودن فرد به عنوان خروجی و مقادیر به‌دست‌آمده از آزمایش‌های بیماران به عنوان ورودی در نظر گرفته‌شده است.

با پیاده‌سازی الگوریتم‌های ماشین بردار پشتیبان CHAID، Exhaustive CHAID و C5.0 تقویت‌شده یک ماتریس درهم ریختگی (Confusion matrix) مطابق جدول ۱ ارایه می‌گردد. این ماتریس یکی از ابزارهای مهم طبقه‌بندی به شمار رفته که چگونگی عملکرد الگوریتم را با توجه به مجموعه داده‌های ورودی بررسی می‌کند. اگر تعداد دسته‌های موجود در طبقه‌بندی m باشد، ماتریس درهم ریختگی یک ماتریس m×m به دست خواهد آمد. در این ماتریس اگر i شماره سطر و j شماره ستون باشد، عنصر Cij تعداد مشاهداتی از دسته i است که توسط الگوریتم دسته‌بندی j تشخیص داده‌شده است(۱۴).

جدول ۱: ماتریس درهم ریختگی

ماتریس درهم ریختگی		رکوردهای پیش‌بینی‌شده	
		دسته−	دسته+
رکوردهای واقعی	دسته+	TN	FP
	دسته−	FN	TP

TP، تعداد نمونه‌هایی که به‌درستی مثبت تشخیص داده می‌شوند؛ TN، تعداد نمونه‌هایی که به‌درستی منفی تشخیص داده می‌شوند؛ FP، تعداد نمونه‌هایی که به اشتباه مثبت تشخیص داده می‌شوند و FN، تعداد نمونه‌هایی که به اشتباه منفی تشخیص داده می‌شوند را نشان می‌دهد. به‌طورکلی در این ماتریس عناصر قطر اصلی، نشان‌دهنده‌ی تعداد مواردی است که به درستی دسته‌بندی‌شده و عناصر قطر فرعی، مواردی می‌باشد که به درستی دسته‌بندی نشده‌اند(۲۰).

به‌منظور ارزیابی عملکرد الگوریتم‌های C5.0 تقویت‌شده، ماشین بردار پشتیبان، CHAID و Exhaustive CHAID شاخص‌های«صحت»، «حساسیت»، «دقت» و «تشخیص» محاسبه می‌گردد. در میان شاخص‌های مطرح‌شده، مهم‌ترین معیار برای تعیین کارایی یک الگوریتم طبقه‌بندی، «دقت» می‌باشد. منظور از دقت، تعیین میزان توانایی یک تست در برآورد تشخیص صحیح و اشتباه از سایر موارد است. در واقع دقت، به میزان نزدیکی مقدار اندازه‌گیری شده به مقدار واقعی عبارت از تعداد نمونه‌هایی که به‌درستی تشخیص داده می‌شوند، اشاره دارد(۲۰). رابطه‌ی ٤ و ٥ به ترتیب بیانگر دقت و صحت در یک الگوریتم پیش‌بینی می‌باشد.

$$Precision= TP/TP+FP \quad (٤)$$

$$Accuracy= TP+TN/TP+TN+FP+FN \quad (٥)$$

دو معیار حساسیت و تشخیص از دیگر معیارهای ارزیابی عملکرد در آمار شاخص‌هایی برای ارزیابی نتیجه یک دسته‌بندی دودویی(دوحالته) هستند. زمانی‌که بتوان داده‌ها را به دو گروه مثبت و منفی تقسیم کرد، دقت نتایج یک آزمایش که اطلاعات را به این دو دسته تقسیم می‌کند با استفاده از شاخص‌های حساسیت و تشخیص قابل‌اندازه‌گیری و توصیف است. حساسیت به معنی نسبتی از موارد مثبت است که آزمایش آن‌ها را به درستی به عنوان مثبت علامت‌گذاری می‌کند

و یا به عبارتی احتمال پیش‌بینی درست ابتلا به بیماری توسط الگوریتم می‌باشد. تشخیص به معنی نسبتی از موارد منفی است که آزمایش آن‌ها را به درستی به عنوان منفی علامت‌گذاری می‌کند و یا احتمال پیش‌بینی درست عدم ابتلا به بیماری توسط الگوریتم می‌باشد(۲۱). محاسبه‌ی حساسیت و تشخیص به ترتیب براساس روابط ٦ و٧ صورت می‌گیرد.

Sensitivity= TP/(TP+FN) (٦)

Specifity= TN/(TN+FP) (٧)

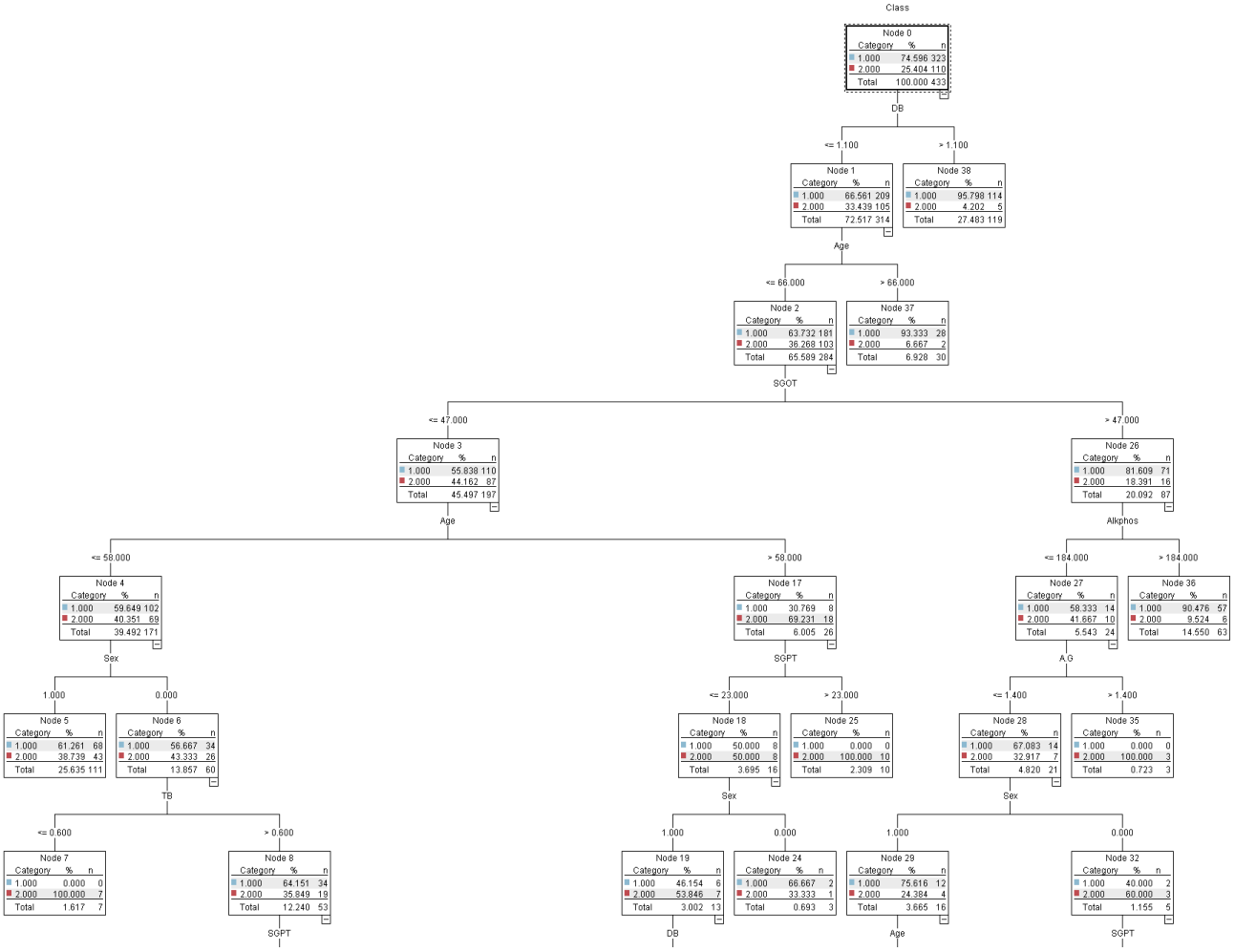
یافته‌ها

یافته‌های حاصل از مطالعه نشان داد که دقت طبقه‌بندی با استفاده از درخت تصمیم C5.0 تقویت‌شده نسبت به سایر الگوریتم‌ها بالاتر است. نتایج حاصل از بررسی در جدول خلاصه‌شده است.

جدول ۲: عملکرد پیش‌بینی الگوریتم‌ها

معیار عملکردی / الگوریتم	دقت	حساسیت	ویژگی	صحت
SVM	٪۸۷/۰۹	٪٦۹/۲۳	٪٦۳/٦۳	٪٦۸
CHAID	٪۸۵/٤۷	٪٦۷/۱۸	٪٦۸/۱۸	٪٦۷/۳۳
Exhaustive CHAID	٪۸۸/۷۱	٪۷۳/۸٤	٪۷۰/۲۱	٪۷۱/۵۹
Boosting C5.0	٪۹٤/۰۹	٪۸۱/۲۳	٪۸٤/٦۳	٪۹٤

همان‌طورکه در جدول ۲ مشاهده می‌شود درخت تصمیم C5.0 تقویت‌شده با دقت ۹٤/۰۹ درصد و صحت ۹٤ درصد بهترین عملکرد را در میان سایر الگوریتم‌ها دارد. لذا با توجه دقت بالای این الگوریتم در پیش‌بینی بیماری‌های کبد استفاده از آن به صورت عملی توصیه می‌گردد. شکل ۱ قسمتی از درخت حاصل از الگوریتم C5.0 را نشان می‌دهد.



شکل ۱: قسمتی از درخت تصمیم C5.0

عمق درخت تولیدشده توسط الگوریتم C5.0 تقویت‌شده (boosting)، ۱۰ است که در مقایسه با عمق درخت حاصل از الگوریتم CHAID بسیار بیشتر است که نشان می‌دهد C5.0 اهمیت عوامل بیشتری را در پیش‌بینی بیماران کبد در نظر گرفته و توانایی نمایش جزئیات بیشتر را دارد. درخت تصمیم C5.0 همچنین تأثیرگذارترین متغیرها را به همراه میزان اهمیت هر متغیر شناسایی می‌کند. در این مطالعه آسپارتات آمینوترانسفراز، جنسیت، آلکالن فسفاتاز، سن، بیلی‌روبین مستقیم، آلبومین، بیلی‌روبین تام، نسبت آلبومین به گلوبولین و پروتئین تام به ترتیب مهم‌ترین متغیرها در پیش‌بینی بیماری کبد شناسایی شدند. همچنین در این مطالعه، C5.0 تقویت‌شده، ۳۶ قانون برای کلاس ۱ و ٤١ قانون برای کلاس ۲ استخراج کرده است، مشاهده می‌شود که الگوریتم تقریباً تعداد قوانین یکسانی برای دو کلاس به‌دست آورده اما در شناسایی افراد بیمار موفق‌تر است. در واقع هر چه تعداد قوانین تولیدشده بیشتر باشد بهتر است؛ زیرا نشان می‌دهد که مدل پیش‌بینی کننده، جزئیات بیشتری را مدنظر قرار داده است. به‌طور نمونه در قوانین تولیدشده آمده که اگر سن فردی کمتر از ٦٦ سال، بیلی‌روبین مستقیم کمتر از ۱/۱، آسپارتات آمینوترانسفراز بیشتر از ٤٧، آلکالن فسفاتاز کمتر از ١٨٤ و نسبت آلبومین به گلوبولین کمتر از ۱/٤ داشته باشد برچسب دسته ۲ گرفته و نشان می‌دهد فرد بیمار است. لازم به ذکر است که قوانین تولیدشده جزئیات بیشتری در مورد هر دو کلاس نشان می‌دهند و این قوانین برای پزشکان و همکارانشان در بیمارستان ساده و قابل‌درک است. هم‌چنین با استفاده از قوانین ایجادشده، برای یک نمونه‌ی جدید با ویژگی‌های مشخص، می‌توان ابتلای فرد به بیماری کبد را پیش‌بینی کرد. با شناخت مهم‌ترین عوامل بروز بیماری های کبد و پیش‌بینی به‌موقع می‌توان امیدوار بود از بروز عارضه تا حدی اجتناب کرد و یا آن را به تعویق انداخت.

بحث

در این مطالعه، از داده‌کاوی برای پیش‌بینی، تشخیص و به دست آوردن مهم‌ترین متغیرهای بروز بیماری‌های کبد استفاده‌شده است. این بیماری با توجه به شیوع و سهمی که در مرگ‌ومیر انسان‌ها دارد از اهمیت بالایی برخوردار است. ویژگی متمایز مطالعه‌ی حاضر دقت بالاتر پیش‌بینی بیماری‌های کبد می‌باشد. تعداد کل متغیرهای مورد استفاده ۱۰ مورد بود. یافته‌ها نشان داد درخت تصمیم C5.0 (تقویت‌شده) در شاخص‌های حساسیت، صحت، ویژگی و دقت بیشترین مقدار را دارد و نسبت به سایر الگوریتم‌ها تعداد بیشتری از نمونه‌ها را در جای درست خود دسته‌بندی کرده است. درخت تصمیم C5.0 (تقویت‌شده) منجر به تولید ۳۹ برگ گردید. ازآنجایی‌که ویژگی درخت تصمیم C5.0 (تقویت‌شده)، دقت بالاتر آن نسبت به پژوهش‌های پیشین است، بنابراین قدرت بیش‌تری در پیش‌بینی بیماران کبد دارد. در سال‌های اخیر پژوهش‌های زیادی بر روی انواع داده‌های بیماران کبدی جهت پیش‌بینی و تشخیص این بیماری در حوزه‌ی داده‌کاوی انجام‌شده است. Abdar و همکاران در سال ۲۰۱۷ از الگوریتم‌های داده‌کاوی جهت پیش‌بینی بیماری کبد استفاده کردند. مجموعه داده‌ی آن‌ها شامل ٤١٦ پرونده بیماری کبد و ١٦٧ پرونده از افراد سالم بود که توسط الگوریتم‌های C5.0 و CHAID مورد تجزیه‌وتحلیل قرار گرفت. مقایسه کارایی این دو الگوریتم مشخص کرد که الگوریتم C5.0 تقویت‌شده عملکرد بهتری نسبت به CHAID دارد، که مطابق با یافته‌های تحقیق حاضر می‌باشد. Abdar و همکاران هم‌چنین متغیرهای بیلی‌روبین مستقیم، آلبومین، آسپارتات آمینوترانسفراز، بیلی‌روبین تام و نسبت آلبومین به گلوبولین را مؤثرترین پارامترها در تشخیص بیماری کبد معرفی کردند(۲)، که پارامترهای مشترکی با پارامترهای شناسایی‌شده توسط درخت تصمیم تحقیق حاضر دارد. Pahareeya و همکاران سیستم تشخیص خودکار بیماری‌های کبد را با رویکرد ترکیبی دو الگوریتم ژنتیک و شبکه عصبی طراحی کردند؛ که بهترین راه‌حل را از اطلاعات مشخصه‌های محلی در مسایل پیچیده پیدا می‌کند. این سیستم با دو مجموعه داده‌ی ILPD و BUPA ارزیابی‌شده که بالاترین دقت پیش‌بینی با دسته‌بندی با نظارت به دست آمد(۲۲). این پژوهش نیز مطابق تحقیق Pahareeya از دسته‌بندی با نظارت استفاده کرده است. Babu و همکاران از فن‌های هوش محاسباتی ازجمله رگرسیون خطی چندگانه، ماشین بردار پشتیبان، شبکه عصبی پیشخور چندلایه، جنگل تصادفی و الگوریتم ژنتیک برای دسته‌بندی بیماران کبدی استفاده نمودند. بررسی‌ها نشان داد که جنگل تصادفی با (٪۲۰۰) Oversampling عملکرد بهتری نسبت به دیگر تکنیک‌ها دارد(۲۳). Ramana و همکاران از الگوریتم‌های بیز ساده، C4.5، KNN، شبکه عصبی پس انتشار و ماشین بردار پشتیبان برای دسته‌بندی بیماران کبد استفاده کردند. الگوریتم‌های KNN، پس انتشار

و ماشین بردار پشتیبان براساس چهار معیار دقت، صحت، حساسیت و مشخصه نتایج بهتری با تمام ترکیبات ویژگی‌ها نشان دادند(۱۱).

بررسی مطالعات داخلی و خارجی نشان می‌دهد که تاکنون از داده‌کاوی برای پیش‌بینی و تشخیص بیماری‌های کبد استفاده گردیده است. نتایج حاصل‌شده، کاربردی بودن داده‌کاوی در حوزه‌ی بیماری‌های کبدی را تأیید می‌نماید. بسیاری از متغیرهای مهم در مطالعات ذکرشده مرتبط با بیماری‌های کبد، با عناصر اطلاعاتی مطالعه‌ی حاضر همخوانی دارند که این مسئله اهمیت این متغیرها در زمینه‌ی پیش‌بینی بیماری‌های کبد را نشان می‌دهد. به‌علاوه پس از شناسایی متغیرهای مهم ذکرشده در مطالعات و استخراج آن‌ها، این متغیرها به تأیید پزشکان و متخصصان این حوزه نیز رسید. از نظر نوع تکنیک مورد استفاده و با توجه به به‌کارگیری بیش‌تر تکنیک‌های دسته‌بندی و درخت تصمیم در مطالعات پیشین و نیز تفسیرپذیری و قابل‌فهم بودن نتایج حاصل از آن‌ها، این تکنیک‌ها به‌عنوان روش پیش‌بینی در این مطالعه انتخاب شد. نقطه‌ی قابل‌بهبود در این مطالعه، پیش‌بینی دقیق‌تر بیماری جهت افزایش اطمینان تصمیم و کاهش مرگ‌ومیر این بیماران و ایجاد سیستم‌های تصمیم‌یار جهت کمک در تشخیص بیماری‌های کبد می‌باشد.

نتیجه‌گیری

در حال حاضر با توجه به اهمیت بیماری‌های کبد، شناسایی روشی برای پیش‌بینی و تشخیص به‌موقع این بیماری‌ها ضروری است. روش‌های طبقه‌بندی مختلفی می‌توانند بیماری‌های کبد را پیش‌بینی کنند که در این پژوهش و با استفاده از داده‌های بیماران کبدی شهر زاهدان، درخت تصمیم C5.0 تقویت‌شده توانست بهترین عملکرد را در طبقه‌بندی ارایه دهد. با توجه به نتایج به‌دست‌آمده از بررسی علایم تأثیرگذار در بروز بیماری‌های کبد، در این پژوهش می‌توان پرخطرترین علایم این بیماری‌ها را در شهر زاهدان شناسایی کرد که شامل ویژگی‌هایی از قبیل آسپارتات آمینوترانسفراز، جنسیت، آلکالن فسفاتاز، سن، بیلی روبین مستقیم، آلبومین، بیلی‌روبین تام، نسبت آلبومین به گلوبولین و پروتئین تام می‌باشند. به‌علاوه متغیرهای شناسایی‌شده به تأیید پزشکان این حوزه رسیده و آنان نیز به اهمیت این متغیرها در پیش‌بینی بیماری‌های کبد اذعان داشتند. با توجه به دقت درخت تصمیم C5.0 (تقویت‌شده) روی داده‌ها، درخت پیشنهادی دقیق، معتبر و قابل استناد است. به‌علاوه مطالعه‌ی حاضر، از روش‌های پیش‌بینی برای استخراج قوانین مرتبط با سیستم‌های خبره استفاده نموده است که می‌توانند در پیش‌بینی بیماری‌های کبد در مراکز درمانی مفید واقع شوند. برای بررسی بیشتر در این زمینه می‌توان در مطالعات بعدی از داده‌های مراکز درمانی دیگر و الگوریتم‌های دیگر نیز استفاده کرده و نتایج را با هم مقایسه نمود. هم‌چنین توصیه می‌شود سیستم تصمیم‌یار بالینی از طریق نتایج این بررسی طراحی‌شده تا پزشکان بتوانند با استفاده از محیطی مناسب به بررسی و پیش‌بینی بیماران کبدی بپردازند و در نتیجه از بروز صدمات جبران‌ناپذیر در افراد پیشگیری نمایند. در نهایت پیشنهاد می‌گردد که از روش‌های فازی با هدف بهره‌برداری از دانش انسانی جهت ایجاد مدل پیش‌بینی استفاده گردد.

تشکر و قدردانی

این مقاله بخشی از پایان‌نامه کارشناسی ارشد، با عنوان«ارایه الگوریتمی جهت پیش‌بینی بیماری کبد با ترکیب دو روش ویکور و تاپسیس» و کد ثبت به شماره ۲۵۱۹۸۹۳ در دانشکده مهندسی صنایع دانشگاه سیستان و بلوچستان می‌باشد. لذا از تمام استادان گروه صنایع که ما را یاری کردند، صمیمانه تشکر می‌شود.

منابع

1. Asrani SK, Devarbhavi H, Eaton J & Kanath PS. Burdent of Liver diseases in the world. Journal of Hepatology 2019; 70(1): 151-71.

2. Abdar M, Zomorodi-Moghadam M, Das R & Ting IH. Performance analysis of classification algorithms on early detection of Liver disease. Expert Systems with Applications 2017; 67(1): 239-51.

3. Acharya UR, Faust O, Molinari F, Sree SV, Junnarkar SP & Sudarshan V. Ultrasound-based tissue characterization and classification of fatty Liver disease: A screening and diagnostic paradigm. Knowledge- Based Systems 2015; 75(1): 66-77.

4. Chitturi S, Farrell GC & George J. Non-alcoholic steatohepatitis in the Asia-Pacific region: Future shock? Journal of Gastroenterology and Hepatology 2004; 19(4): 368-74.

5. Pourshams A, Malekzadeh R, Monawari A, Akbari MR, Mohamadkhani A, Yarahmadi S, et al. Prevalence and etiology of persistently elevated alanine aminotransterase levels in healthy Iranian blood donors. Journal of Gastroenterology and Hepatology 2005; 20(2): 229-33.

6. Abedian E, Ayoobi A, Yusefi Poor M & Ghafari HR. Prognosis of hepatic dysfunction using a multi-layer perceptron artificial neural network, University of Torbat-e-Heydariyeh: 4th National Conference on Information Technology, Computer & Telecommunication, 2017.

7. Ameri H, Alizadeh S & Barzgari A. Knowledge extraction of Diabetics' data by decision tree method. Journal of Health Administraction 2013; 16(53): 58-72[Article in Persian].

8. Abolmasum F, Alizadeh S & Asghari M. Utilizing data mining techniques for investigating factors influencing the failure of intrauterine insemination infertility treatment. Journal of Health Administraction 2014; 16(54): 46-55[Article in Persian].

9. Dormohammadi S, Alizadeh S, Asghari M & Shami M. Proposing a prediction model for diagnosing Causes of infertility by data mining algorithms. Journal of Health Administraction 2014; 17(57): 46-57[Article in Persian].

10. Azizi AA, Zarei J, Nabovati E, Vakili-Arki H, Abbasi E & Razavi AR. Determining of the factors affecting mortality in burn patients using a decision tree data mining algorithm. Journal of Health Administraction 2014; 16(54): 34-45[Article in Persian].

11. Ramana BV, Babu MSP & Venkateswarlu NB. A critical study of selected classification algorithms for Liver disease diagnosis. International Journal of Database Management Systems (IJDMS) 2011; 3(2): 101-14.

12. Yeh WC, Hsu WD, Islam MM, Nguyen PA, Poly TN, et al. Prediction of fatty Liver disease using machine learning algorithms. Computer Methods and Programs in Biomedicine 2019; 170(1): 23-9.

13. Karim H, Etminani K, Tara SM & Mardani M. Identifying factors associated with length of hospital stay using decision tree. Journal of Health Administraction 2015; 18(61): 57-68[Article in Persian].

14. Han J & Kamber M. Data mining: Concepts and techniques. 2nd Edition. San Francisco: Morgan Kaufmann Publishers; 2006: 78-82.

15. Shahrabi J & Zare A. Data mining with celementine. Tehran: Academic Jahad, Amirkabir University of Technology; 2013: 101-4[Book in Persian].

16. Witten IH & Frank E. Data mining: Practical machine learning tools and techniques. 2nd edition. San Francisco: Morgan Kaufmann Publishers; 2005: 35-7.

17. Wang Z & Xue X. Multi-class support vector machine. New York: Springer International Publishing; 2014: 23-48.

18. Mahmoodi M. Designing a Heart disease prediction system using support Vector machine. Journal of Health and Biomedical Informatics Medical Informatics Research Center, 2017; 4(1): 1-10.

19. Felea MG, Felea V & Gavrilescu CM. Using CHAID algorithm in low-risk metabolic syndrome patients, Chisinau: 3rd International Conference on Nanotechnologies and Biomedical Engineering, 2016.

20. Deng N, Tian Y & Zhang C. Support Vector machines: Optimization based theory, algorithms and extensions. United States: CRC press; 2012: 25-30.

21. Mazaheri S, Ashoori M & Bechari Z. A model to predict Heart disease treatment using data mining. Journal of Payavard Salamat 2017; 11(3): 287-96[Article in Persian].

22. Pahareeya J, Vohra R, Makhijani J & Patsariya S. Liver patient classification using intelligence techniques. International Journal of Advanced Research in Computer Science and Software Engineering 2014; 4(2): 295-9.

23. Babu MSP, Venkata Ramana B & Sarath Kumar BR. New automatic diagnosis of liver status using bayesian classification, Malaysia: International Conference on Intelligent Network and Computing (ICINC), 2010.

Evaluation of Data Mining Algorithms for Detection of Liver Diseases

Mohammad Reza Shahraki[1] (Ph.D.) – Mahboubeh Mesgar[2] (B.S.)

1 Assistant Professor, Department of Industrial Engineering, Faculty of Engineering Shahid Nikbakht, Sistan and Balochestan University, Zahedan, Iran

2 Master of Sciences Student in Industrial Engineering, Faculty of Engineering Shahid Nikbakht, Sistan and Baluchestan University, Zahedan, Iran

Abstract

Received: Aug 2018

Accepted: Dec 2018

Background and Aim: The liver, as one of the largest internal organs in the body, is responsible for many vital functions including purifying and purifying blood, regulating the body's hormones, preserving glucose, and the body. Therefore, disruptions in the functioning of these problems will sometimes be irreparable. Early prediction of these diseases will help their early and effective treatment. Regarding the importance of liver diseases and increasing number of patients, the present study, using data mining algorithms, aimed to predict liver disease.

Materials and Methods: This descriptive study was performed using 721 data from liver patients from Zahedan. In this study, after preprocessing data, data mining techniques such as SVM: Support Vector Machine, CHAID, Exhaustive CHAID and boosting C5.0, data were analyzed using IBM SPSS Modeler 18 data mining software.

Result: According to the findings, the liver diseases can be predicted by the enhanced C5.0 algorithm with precision of 94/09, exhaustive CHAID algorithm with precision of 88/71, SVM with the precision of 87/09, and CHAID algorithm with the precision of 85/47. The enhanced C5.0 algorithm showed the best performance among other algorithms.

Conclusion: According to the rules created by boosting C5.0 algorithm, for a new sample, one can predict the likelihood of a person for developing liver disease with high precision.

Keywords: Liver Disease, Support Vector Machine (SVM), CHAID Algorithms, C5.0 Decision Tree, Data Mining

*** Corresponding Author:**
Mesgar M
Email:
mhb@pjs.usb.ac.ir