

کاربرد داده کاوی در بیماران مبتلا به بیماری آسم

طاها صمدسلطانی^۱، دکتر مصطفی لنگری زاده^۲، مریم ذالنوری^۳

چکیده

زمینه و هدف: داده کاوی شاخه‌ی بسیار مهمی در فهم عمیق‌تر داده‌های پزشکی است که در صدد حل مسائل در تشخیص و درمان بیماری‌ها می‌باشد. یکی از مهمترین کاربردهای داده کاوی، بررسی الگوهای موجود در داده‌ها است. هدف از این پژوهش تعیین الگوهای موجود در داده‌های بیماران مبتلا به آسم است.

روش بررسی: جامعه‌ی پژوهش شامل بیماران دارای نشانه‌های تنفسی و نمونه پژوهش شامل داده‌های ثبت شده‌ی ۲۵۸ فرد مراجعه کننده با علائم تنفسی به بیمارستانهای امام خمینی(ره) و مسیح دانشوری شهر تهران بود. مجموعه داده‌ی مذکور با استفاده از فرم گردآوری داده جمع آوری گردید و در محیط نرم افزار Excel وارد شد و افزونه داده کاوی این نرم افزار مورد استفاده قرار گرفت و سپس تحلیل‌های اثرگذارهای کلیدی، دسته بندی بیماران و تشخیص استثنائات انجام شد. **یافته‌ها:** فراوان ترین نشانه بالینی مبتلایان به آسم در مجموعه‌ی حاضر، سرفه‌های شدید بود و این بیماری تحت تاثیر شدید هیجانات قرار داشت. داده‌ها، جهت انجام تحلیل‌های کلی‌تر، در پنج خوشه تجمیع شدند و وجه مشترک آنها ارائه شد و سوابقی که ویژگی‌های استثنایی داشتند، شناسایی شدند. سپس با تحلیل هزینه‌ها و تعیین مقدار آستانه برابر ۶۱۲، پرسش نامه‌ی تشخیص آسم به روش کارت امتیاز ارائه گردید.

نتیجه‌گیری: چارچوب ارائه شده برای تحلیل و داده کاوی، ابزاری مناسب جهت استخراج دانش از روی داده‌ها می‌باشد و می‌تواند هنگام استفاده از راهکارهای بالینی، شکاف موجود در تصمیم گیری را شناسایی و پر کند. **واژه‌های کلیدی:** داده کاوی، دسته بندی بیماران، بیماری آسم

* نویسنده مسئول :

دکتر مصطفی لنگری زاده؛
دانشکده مدیریت و اطلاع رسانی
پزشکی دانشگاه علوم پزشکی ایران
Email :
Langarizadeh.m@iums.ac

- دریافت مقاله : فروردین ۱۳۹۴ پذیرش مقاله : تیر ۱۳۹۴

مقدمه

انتخاب، اکتشاف و ساختن مدل‌ها با استفاده از انبوه داده‌های ذخیره شده برای کشف الگوهای از پیش موجود نیز یاد می‌شود(۲). از داده کاوی برای شناسایی روابط و الگوهای نو، صحیح، قابل فهم و به صورت بالقوه مفید در درون داده‌ها، با استفاده از ترکیب مجموعه داده‌ها، و استخراج الگوهای پیچیده برای انسان استفاده می‌شود(۳).

داده کاوی، گامی مهم در کشف و استخراج دانش می‌باشد؛ این اصطلاح به معنای کاوش مجموعه داده‌های بزرگ برای استخراج الگوهای ناشناخته بین

داده کاوی به عنوان فرآیندی برای یافتن الگوها و ارتباطات در پایگاه داده‌ها به همراه استفاده از اطلاعات برای ساختن مدل‌های پیش بینی تعریف شده است(۱) و همچنین از آن به عنوان فرآیندی برای

^۱ دانشجوی دکتری انفورماتیک پزشکی، گروه مدیریت اطلاعات سلامت، دانشکده پیراپزشکی، دانشگاه علوم پزشکی تهران، تهران، ایران
^۲ استادیار گروه مدیریت اطلاعات سلامت، دانشکده مدیریت و اطلاع رسانی پزشکی، دانشگاه علوم پزشکی ایران، تهران، ایران
^۳ دانشجوی دکتری انفورماتیک سلامت، گروه انفورماتیک سلامت، دانشکده انفورماتیک، دانشگاه ایالتی ایندیانا، ایندیانا، ایالات متحده آمریکا

ابزارها و روش‌های بسیاری برای داده کاوی و تحلیل آن‌ها وجود دارد که الگوریتم‌های مشابهی را پیاده سازی کرده‌اند. این پژوهش با استفاده از روش‌های تحلیل و داده کاوی، قصد دارد تا با تحلیل داده‌های گردآوری شده برای آسم، بخشی از دانش پنهان در آن را کشف نموده و گزارش کند.

از سوی دیگر، آسم یک بیماری مزمن التهابی رایج راه‌های هوایی است که بوسیله‌ی علائم تشدید شونده، انسداد برگشت پذیر راه‌های هوایی و اسپاسم نایژه‌ها مشخص می‌شود (۱). شیوع آسم از دهه ۱۹۷۰ به طرز چشمگیری رو به افزایش نهاد و در سال ۲۰۱۰، سیصد میلیون نفر در جهان به این بیماری مبتلا شدند و در سال ۲۰۰۹ باعث مرگ دویست و پنجاه هزار نفر در سراسر جهان شد (۲). این بیماری نرخ فزاینده‌ای در کشورهای در حال توسعه دارد (۳). تشخیص بیماری آسم به دلیل پیچیدگی الگوها و نشانه‌ها از مسائل چالش بر انگیز به شمار می‌رود. این پژوهش بر اساس داده‌های گردآوری شده از افراد مراجعه کننده به بیمارستان‌های امام خمینی (ره) و مسیح دانشوری تهران که در پژوهش ذوالنوری و همکاران گردآوری شده بود، انجام پذیرفت (۱۱). هدف این مطالعه تعیین روابط پنهان بین متغیرهای موثر در افراد مبتلا به آسم و استخراج الگوی آن جهت پیش بینی میزان ابتلا به آسم می‌باشد.

روش بررسی

در این مطالعه، جامعه‌ی پژوهش، بیماران دارای علائم تنفسی در شهر تهران بودند و نمونه پژوهش را داده‌های ثبت شده‌ی ۲۵۴ فرد مراجعه کننده با علائم تنفسی به بیمارستان‌های امام خمینی (ره) و مسیح دانشوری در شهر تهران تشکیل می‌داد. این دو بیمارستان بزرگترین مراکز ارائه‌ی خدمات درمانی در

داده‌ها است (۴). از آنجایی که کشف روابط بین داده‌ها در روش‌های سنتی آماری بسیار مشکل است (۵)، لذا کاربرد داده کاوی به سرعت در بخش‌های وسیعی از قبیل سازمان‌های ارائه خدمات بهداشتی، پیش بینی مالی و پیش بینی هوا گسترش یافته است (۶).

داده کاوی در مراقبت سلامت، شاخه‌ی بسیار مهمی در تشخیص و فهم عمیق‌تر داده‌های پزشکی می‌باشد. داده کاوی بهداشتی درمانی، در صدد حل مسائل دنیای واقعی در تشخیص و درمان بیماری‌هاست (۷). همچنین در مراقبت بهداشتی درمانی، زمینه‌ی تحقیقاتی مهمی برای پیش بینی بیماری‌ها و درکی عمیق‌تر از داده‌های بهداشتی درمانی محسوب می‌شود. داده کاوی بهداشتی درمانی قصد دارد تا مسائل دنیای واقعی را در تشخیص و درمان بیماری‌ها حل کند (۸). پژوهشگران از این روش برای تشخیص بیماری‌های مختلف استفاده می‌کنند و برای این کار از روشها و الگوریتم‌های مختلفی بهره می‌برند که میزان صحت و دقت آنها با یکدیگر متفاوت است (۹).

کاربردهای داده کاوی می‌توانند در تمام بخشهای صنعت مراقبت بهداشتی درمانی به صورت جدی مثر ثمر باشند. برای مثال، سازمان‌های ارائه دهنده خدمات بیمه بهداشتی درمانی می‌توانند کلاهبرداری و سوء استفاده‌ها را شناسایی نمایند؛ سازمان‌های بهداشتی درمانی تصمیمات مربوط به مدیریت ارتباط با مراجعه کنندگان را اتخاذ نمایند؛ پزشکان، درمانهای موثر و بهترین شیوه‌های درمانی را شناسایی نمایند، و بیماران، خدمات مراقبتی بهتر و رضایت بخش‌تر را دریافت نمایند (۳). شناخت روشهای پیشگیرانه و درمانی در بیماری‌های مزمن از قبیل آسم از طریق داده کاوی میسر می‌شود و این روش نقش بارزی در شناخت الگوها دارد (۱۰).

دلیل ماهیت پیش بینی در مقاله و لزوم بررسی گذشته نگر، و دسترسی به مجموعه داده‌ها. در ادامه این مجموعه داده در محیط نرم افزار صفحه گسترده مایکروسافت اکسل Excel وارد شد و افزونه‌ی داده کاوی مورد استفاده قرار گرفت. محیط‌های موجود در پایگاه داده مورد استفاده در جدول ۱ نشان داده شده است.

حوزه‌ی بیماری‌های دستگاه تنفسی می‌باشند که دارای بخش‌های تخصصی ریه هستند و حجم مراجعه کننده بزرگی دارند و در شهر تهران قرار دارند. مجموعه داده‌ی مذکور در پژوهش ذوالنوری و همکاران با استفاده از فرم گردآوری داده، جمع آوری گردید (۱۱). دلایل انتخاب این مجموعه داده عبارتند از: تکمیل بودن داده‌های آن و وجود نقص در داده‌ها یا فقدان گردآوری صحیح آن در سال‌های بعدی، گذر زمان به

جدول ۱: صفات مجموعه داده بیماری آسم

ویژگیها	بیماری آسم (تعداد بیماران)
۱: سابقه بیماری در یکی از والدین (par1)	۱۶۹ (آسمی)
۲: سابقه بیماری در هر دو والدین (par2)	۸۵ (غیرآسمی)
۳: سابقه بیماری در خواهران و برادران (relative)	
۴: رینیت آلرژی (AR)	
۵: شاخص توده بدنی (BMI)	
۶: سابقه ریفلاکس دستگاه گوارشی (GER)	
۷: حساسیت به آلرژن‌ها (R_AI)	
۸: واکنش در عصبانیت (R_Ir)	
۹: واکنش به هیجانات (R_Em)	
۱۰: تنگی قفسه سینه (CT)	
۱۱: تشدید علائم در هنگام فعالیت بدنی (Exe.S)	
۱۲: تنگی نفس (Dysp)	
۱۳: سابقه اگزما و تب (Ecz)	
۱۴: بستری قبل از سه سالگی به دلیل عفونت ویروسی (Hosp)	
۱۵: آلرژی به غذا قبل از سه سالگی (Food)	
۱۶: سرماخوردگی (Cold)	
۱۷: در معرض آلودگی هوا (AP)	
۱۸: در معرض دود سیگار (Smok.p)	
۱۹: استعمال سیگار توسط مادر در دوران حاملگی (Preg)	
۲۰: در معرض حیوانات خانگی قرار گرفتن در کودکی (Animal)	
۲۱: زندگی در گرما و شرایط رطوبتی (Damp)	
۲۲: سرفه (Cough)	
۲۳: خلط (Phlegm)	
۲۴: خس خس (Wheeze)	

پاکسازی در پژوهش‌های پیشین، بر روی این مجموعه داده، توسط پژوهشگران انجام شده بود و در این پژوهش، تحلیل‌های اثرگذاران کلیدی (Key Influencers)، تشخیص دسته‌ها (Categories Detect)، و تشخیص استثنائات (Highlight Expectations) انجام شد. در پایان نیز یک ماشین پیشگو (Prediction Calculator) بصورت پرسش نامه، جهت تشخیص آسم بر اساس داده‌ها ارائه شد. ابزار مورد استفاده جهت تحلیل و داده کاوی، افزونه‌ی Excel و ابزار تحلیلی SQL Server بود. پس از نصب افزونه در محیط Excel، دو برگه‌ی جدید با عناوین تحلیل و داده کاوی افزوده شدند که گزینه‌های موجود در برگه تحلیل، منطبق با عناوین تحلیل‌های انجام شده در این پژوهش می‌باشد.

داده‌های خام گردآوری شده با پیش پردازش‌هایی از قبیل گسسته سازی و تبدیل عددی و همچنین حذف داده‌های ناقص و حذف ویژگی‌های غیر موثر توسط روش‌های انتخاب ویژگی به دسته بندی مطابق جدول ۱ منجر شد. این مجموعه داده پس از طی مراحل مذکور شامل ۲۵۰ سابقه و متشکل از ۲۴ صفت ورودی و ۱ صفت خروجی گسسته بود. هر یک از ویژگی‌های ذکر شده در جدول ۱، دامنه مقادیر کمی دارند. برای مثال، سرفه می‌تواند بسته به شدت، مقادیر عددی بین ۰ تا ۵ را اخذ نماید. کلاس‌های خروجی نیز شامل ۲ کلاس، ۰ یا ۱ بود که نشان دهنده‌ی تشخیص آسم یا غیر آسم می‌باشد. مراحل داده کاوی جهت کشف الگوها و بررسی روابط، بر اساس مدل پیشنهادی زیر انجام پذیرفت که در آن شناخت مسئله، جمع آوری داده‌ها، و تبدیل و

یافته‌ها

جدول ۲: نتایج تحلیل اثرگذاران کلیدی روی مجموعه داده

ستون	مقدار	وضعیت ابتلا به آسم	تأثیر نسبی
Cough	=>۴.۳۳۰۵۲۵۰۳۳۶	۱	۱۰۰
R_Em	۱	۱	۷۴
Wheeze	۲	۱	۷۳
Exe_S	۱	۱	۷۳
Damp	۰	۱	۵۹
Hosp	۰	۱	۴۳
Dysp	۱	۱	۳۳
R_Ir	۱	۱	۳۱
R_AL	۱	۱	۲۷
Phlegm	۰/۷۵	۱	۱۸
Cough	۰/۹۲۲۲۳۹۸۸۵۵	۰	۸۵
Dysp	۰	۰	۷۶
R_Em	۰	۰	۷۴
Exe_S	۰	۰	۷۳

Cough	۲/۴۲۱۶۱۷۸۵۵۲-۰/۹۲۲۳۹۸۸۵۵	۰	۶۲
Damp	۱	۰	۵۹
Wheeze	۰	۰	۴۳
Hosp	۱	۰	۴۳
AR	۰	۰	۳۵
R_Ir	۰	۰	۳۱
R_AL	۰	۰	۲۷
Phlegm	۱	۰	۲۲
par2	۰	۰	۱۶
AR	۱	۰	۱۳

مقدار ستون دوم را در رخداد ستون سوم به ازای هر سطر، نشان می‌دهد.

با انجام تحلیل دسته‌های مسئله پژوهش، افراد بر اساس شاخص‌های مشتق شده از ویژگی‌ها به گروه‌هایی دسته بندی شدند و مشخصات این افراد درون این دسته‌ها یا خوشه‌ها مشابه هم بودند. برای هر دسته تنها دو عامل با بیشترین تاثیر نسبی محاسبه شدند و نام دسته هم نام با موثرترین عامل انتخاب شد.

جدول ۲، نتایج تحلیل اثرگذاران کلیدی را نشان می‌دهد. در ستون اول، ویژگی‌های موجود در مجموعه داده‌ها ارائه شده است که نشان دهنده نشانه‌ها، سوابق و عوامل خطرزا در افراد مراجعه کننده به بیمارستان می‌باشد. ستون دوم، مقادیر مربوط به ستون مورد نظر را نشان می‌دهد. ستون سوم از صفر یا یک تشکیل شده است که نشان دهنده وضعیت تشخیص فرد است؛ به طوری که عدد یک، نشان دهنده‌ی ابتلا به آسم و عدد صفر، نشان دهنده‌ی عدم ابتلاست. ستون چهارم نیز تاثیر نسبی ستون اول با

جدول ۳: نتایج تحلیل تشمیص دسته‌ها برای مجموعه داده‌ی آسم

نام دسته	تعداد رکوردها	عوامل اصلی	مقدار عامل	تأثیر نسبی عامل
سابقه‌ی بستری پس از سه سالگی به دلیل عفونت ویروسی	۷۱	Hosp	۰	۱۰۰
		R_Ir	۱	۹۸
وضعیت ابتلا به آسم	۵۷	Asthma	۰	۱۰۰
		Damp	۱	۳۹
وضعیت سرفه	۶۰	Cough	۴.۳۳۰	۱۰۰
		Wheeze	۲	۷۳
سابقه‌ی بیماری در هر دو والدین	۴۶	Par2	۲	۱۰۰
		AR	۳	۱۰۰
آلرژی به غذا بعد از سه سالگی	۲۰	Food	۱	۱۰۰
		Cold	۱	۸۰

موجود در پایگاه داده با داده‌های غیر معمول یا احتمالاً خطا دار مشخص شدند. این پژوهش با تعیین آستانه‌ای برابر ۰/۷۵، تنها ۲۵٪ داده‌های استثنائی را شناسایی نمود. هرچه میزان آستانه کمتر باشد، تعداد بیشتری از سوابق به عنوان استثناء شناسایی می‌شوند.

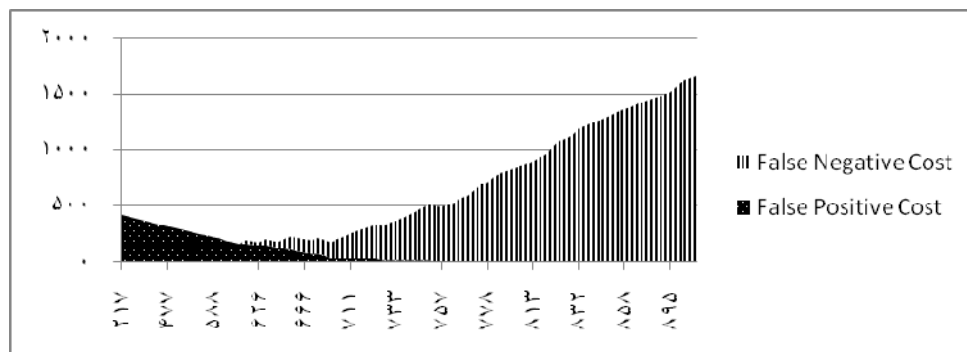
جدول ۳ دسته‌های تشخیص داده شده را به همراه تعداد سوابق خوشه بندی شده درون آن و دو عامل اصلی برای دسته بندی به همراه تاثیر نسبی آنها نشان می‌دهد. با انجام تحلیل تشخیص استثنائات، سوابق

جدول ۴: نتایج تملیل تشخیص استثنائات

عامل	فراوانی
Cough	۵
Hosp	۳
Total	۸

مقدار تشخیص (False Positive) که تنها هزینه‌ی بررسی بیشتر را شامل می‌شود قرار داده شد (FN=۱۰)، در شکل ۱ این مقادیر و نحوه‌ی محاسبه‌ی مقدار آستانه برای تفکیک بیماران آسمی از غیرآسمی نشان داده شده است که بر اساس روش‌های آماری و به صورت بهینه صورت می‌گیرد. مقدار آستانه که محل تلاقی هزینه‌ها می‌باشد برابر ۶۱۲ محاسبه شد. از این رو، امتیازات بالای آن، آسمی و پایین‌تر، غیر آسمی محسوب می‌شوند.

جدول ۴ نتایج تشخیص استثنائات را با مقدار آستانه ۷۵ درصد نشان می‌دهد. ستون اول عامل موثر در پرت شدن سوابق و ستون دوم فراوانی این عامل را در استثناء شدن سوابق نشان می‌دهد. برای این مجموعه داده با انجام تحلیل‌های ماشین حساب پیشگو توسط ابزارهای داده کاوی، معیاری جهت تشخیص مبتلایان به آسم طراحی و پیشنهاد شد. با توجه به هزینه‌ی عدم تشخیص صحیح (False Negative) در تشخیص بیماری‌ها و خطرات ناشی از آن برای بیماران، مقدار این متغیر دو برابر



شکل ۱: هزینه‌های (مقدار تشخیص غلط غیربیمار و تشخیص غلط بیمار و مقدار آستانه بر اساس نمونه گیری تکراری

جدول ۵: پرسش نامی پیشنهادی برای تشخیص بیماری آسم بر اساس نشانه‌ها

Cough	< ۰.۹۲۲۳۹۸۸۵۵ Very Low	۰	<input type="checkbox"/>	GER	۰ no	۳۰	<input type="checkbox"/>
	۰.۹۲۲۳۹۸۸۵۵ - ۲.۴۲۱۶۱۷۸۵۵۲ Low	۶	<input type="checkbox"/>		۱ yes	۰	<input type="checkbox"/>
	۲.۴۲۱۶۱۷۸۵۵۲ - ۳.۳۳۵۱۷۲۹۷۶ Medium	۱۸۱	<input type="checkbox"/>	Par2	۰ Asthma	۰	<input type="checkbox"/>
	۳.۳۳۵۱۷۲۹۷۶ - ۴.۳۳۰۵۲۰۳۳۶ High	۱۸۵	<input type="checkbox"/>		۱ Allergy	۹	<input type="checkbox"/>
	>= ۴.۳۳۰۵۲۰۳۳۶ Very High	۱۸۴	<input type="checkbox"/>		۲ fever	۱۱	<input type="checkbox"/>
Phlegm	۰ no phlegm	۱۱۷	<input type="checkbox"/>	Par1	۰ no	۰	<input type="checkbox"/>
	۰.۵ low	۱۲۱	<input type="checkbox"/>		۱ yes	۲	<input type="checkbox"/>
	۰.۷۵ medium	۱۲۴	<input type="checkbox"/>	Rel	۰ no	۱۰	<input type="checkbox"/>
	۱ High	۰	<input type="checkbox"/>		۱ yes	۰	<input type="checkbox"/>
Wheeze	۰ no wheeze	۰	<input type="checkbox"/>	R_AL	۰ no	۰	<input type="checkbox"/>
	۰.۲۵ very low	۱۹	<input type="checkbox"/>		۱ yes	۷	<input type="checkbox"/>
	۱ low	۱۵	<input type="checkbox"/>	R_Ir	۰ no	۰	<input type="checkbox"/>
	۱.۵ medium	۱۳	<input type="checkbox"/>		۱ yes	۵	<input type="checkbox"/>
	۲ high	۲۰	<input type="checkbox"/>	R_Em	۰ no	۰	<input type="checkbox"/>
	۳ very high	۱۰	<input type="checkbox"/>		۱ yes	۴۷	<input type="checkbox"/>
Dysp	۰ no dysp	۶	<input type="checkbox"/>	Ecz	۰ no	۲۱	<input type="checkbox"/>
	۰.۵ low	۵	<input type="checkbox"/>		۱ yes	۰	<input type="checkbox"/>
	۱ medium	۱۷	<input type="checkbox"/>	Hosp	۰ no	۵۰	<input type="checkbox"/>
	۲ high	۰	<input type="checkbox"/>		۱ yes	۰	<input type="checkbox"/>
۳ very high	۱۹	<input type="checkbox"/>					
CT	۰ no	۰	<input type="checkbox"/>	Food	۰ no	۰	<input type="checkbox"/>
	۱ yes	۱۱	<input type="checkbox"/>		۱ yes	۳۴	<input type="checkbox"/>
Exe_S	۰ no	۰	<input type="checkbox"/>	Cold	۰ ۰ cold in year	۴۲	<input type="checkbox"/>
	۱ yes	۷۶	<input type="checkbox"/>		۱ cold in year	۰	<input type="checkbox"/>
AR	۰ no AR	۸۲	<input type="checkbox"/>		۲ or more cold in year	۵۳	<input type="checkbox"/>
	۱ low	۰	<input type="checkbox"/>	AP	۰ no	۰	<input type="checkbox"/>
	۲ medium	۸۱	<input type="checkbox"/>		۱ yes	۱	<input type="checkbox"/>
	۳ high	۷۲	<input type="checkbox"/>	P_smoke	۰ exposure to smoking at present	۶۴	<input type="checkbox"/>
			۱ exposure to smoking at infancy		۶۲	<input type="checkbox"/>	
			۲ self-smoking		۰	<input type="checkbox"/>	
BMI	۰ below ۳۰	۵۰	<input type="checkbox"/>				
	۱ above ۳۰	۰	<input type="checkbox"/>				
	Missing	۴۷	<input type="checkbox"/>				
Preg	۰ no	۴۶	<input type="checkbox"/>	Damp	۰ no	۰	<input type="checkbox"/>
	۱ yes	۰	<input type="checkbox"/>		۱ yes	۵	<input type="checkbox"/>
Animal	۰ no	۴۹	<input type="checkbox"/>				
	۱ yes	۰	<input type="checkbox"/>				
	Missing	۴۷	<input type="checkbox"/>				

دانش مهمی پیرامون نشانه‌ها و عوامل خطرزای بیماری آسم ارائه می‌دهد.

نتایج تحلیل تشخیص دسته‌ها منجر به گروه بندی بیماران جهت انجام پژوهش‌های بعدی شده است. بزرگترین وجه اشتراک سوابق حاضر در مجموعه داده، نبود سابقه‌ی بستری پس از سه سالگی به دلیل عفونت ویروسی می‌باشد که بزرگترین خوشه را تشکیل می‌دهد. البته این عامل به تنهایی در خوشه بندی لحاظ نشده و به همراه سایر عوامل محاسبه گردیده است. از جمله، سایر خوشه‌ها را می‌توان افراد غیر آسمی یا افراد با علائم سرفه‌ی شدید و یا سابقه‌ی هر دوی والدین نام برد که بر اساس مطالعات قبلی نیز مورد تایید قرار می‌گیرند (۱۵ و ۱۴).

در تحلیل تشخیص استثنائات، دو مقدار در مجموعه داده‌ها بسیار به ندرت یافت می‌شود و کمترین نمونه آماری را تشکیل می‌دهند. اولین عامل، سرفه‌های با مقدار ۰/۵ است که به لحاظ داده‌ای، معنای خاصی ندارد؛ اما دومین عامل، سابقه‌ی بیماری یا آلرژی یا تب هر دو والدین است که در آن متغیر Par2 برابر مقدار ۲ قرار داده شده است که نشان دهنده‌ی سابقه‌ی بروز تب در تعداد موارد نادری از بیماران است و داده‌ای واقعی است و خطا محسوب نمی‌شود. بنابراین، می‌توان ادعا نمود که داده‌ها، با کیفیت و کم خطا محسوب می‌شوند. اینگونه بررسی، در مطالعات قبلی یافت نشد و در نوع خود جدید بوده لذا نتایج قابل قیاس در دسترس نمی‌باشد و قضاوت و بحث در این مورد امکان پذیر نیست.

ساخت پرسش نامه برای تشخیص بیماران مبتلا به آسم از افراد غیر آسمی روالی می‌باشد که در مطالعات متعددی انجام شده است (۱۹-۱۶). در این پژوهش، پرسش نامه‌ی مذکور که به صورت ماشینی حسابی است باید مورد ارزیابی نیز قرار گیرد. اما روی

در جدول ۵، پرسش نامه‌ای بر اساس روش کارت امتیاز پیشنهاد شده است که امکان پیش بینی و تشخیص را بدون نیاز به نرم افزار فراهم می‌نماید. در این جدول بر اساس ویژگی‌ها و مقادیر آنها که به صورت عددی و نشان دهنده‌ی شدت یا تناوب و فراوانی نشانه‌ها و علائم بیماری می‌باشند، به انتخاب گزینه‌ی مناسب برای وضعیت بیمار، پرداخته می‌شود و سپس مقادیر تاثیر نسبی هر گزینه جمع زده شده و اگر مقدار نهایی از مقدار آستانه که معادل ۶۱۲ می‌باشد، بیشتر باشد فرد دارای نشانه‌های آسم بوده و به احتمال بسیار زیاد مبتلا به آسم می‌باشد و بالعکس. مقادیر عددی ویژگی‌ها بر اساس مقیاس‌های تعریف شده در مجموعه داده تعریف شده است. برای مثال، سرفه که مقداری ما بین صفر و پنج در مجموعه داده دارد بر اساس شدت به ۵ دسته‌ی مشابه نگاشت شده است.

بحث

بر اساس یافته‌های پژوهش بر روی مجموعه داده‌ی بیماری آسم، می‌توان دریافت که بین ابتلا به آسم و سرفه‌های بسیار شدید همبستگی بسیار بالا وجود دارد و فراوان ترین نشانه‌ی آسم شناخته شده است. با تحلیل سوابق غیر آسمی، بر این نکته تاکید می‌شود که سرفه‌های شدید اصلی‌ترین شاخصه‌ی مبتلایان به آسم است. نتیجه‌ی داده کاوی با مطالعات قبلی پزشکی پیرامون آسم مطابقت دارد و تاکید کننده بر این یافته می‌باشد (۱۲). از نتایج این تحلیل جهت استدلال مبتنی بر موارد (Case Based Reasoning) می‌توان استفاده نمود یا برای تدوین راهکارهای بالینی بهره گرفت تا شکاف دانشی موجود در راهکارهای بالینی جهت اتخاذ تصمیمات پزشکی شناسایی و تکمیل شود (۱۳). نتایج تحلیل اثر گذاران کلیدی،

مجموعه داده‌ها وجود دارد. روش پیشنهاد شده در این پژوهش و نتایج پژوهش، علاوه بر معرفی روش‌های تحلیل و داده کاوی، چارچوبی برای بررسی همبستگی و روابط بین اطلاعات هویتی، علائم، عوامل خطرزا و نشانه‌های بیماری‌ها را در مبتلایان به آن بیماری‌ها بررسی می‌نماید، و علاوه بر این ایجاد پرسش نامه‌ای مبتنی بر محاسبات جهت تشخیص این بیماری‌ها بر اساس هزینه‌های تشخیصی خروجی کاربرد دیگری می‌باشد که می‌تواند متعاقباً جهت شناسایی و پر نمودن شکاف دانشی موجود در راهکارهای بالینی به کار رود. این روش بیشتر در حوزه‌های مالی و فروش به کار می‌رود، اما در حوزه سلامت، و به خصوص بررسی مبتلایان به بیماری‌ها نیز کاربرد دارد.

مجموعه داده‌های موجود به طور قطع بهترین عملکرد را برای تفکیک بیماران از افراد سالم دارد و معیار مناسبی محسوب می‌شود و می‌توان آن را پس از ارزیابی و اعتبار سنجی به صورت استاندارد به عنوان ابزار بالینی کمک تشخیصی به کار برد. در مطالعات انجام شده‌ی پیشین، پرسش نامه‌هایی مبتنی بر روش کارت امتیاز طراحی و ارزیابی شده‌اند که غالباً مبتنی بر سوالاتی کمتر و کلی‌تر هستند و از این رو دارای دامنه‌ی تشخیص محدود می‌باشند (۲۰).

نتیجه گیری

داده کاوی و تحلیل داده‌ها در نظام سلامت، همانند سایر حوزه‌ها، منجر به استخراج دانش مفید در تشخیص، درمان و پیشگیری از بیماری‌ها می‌شود. تحلیل‌های مختلفی برای بررسی روابط و ویژگی‌های

منابع

1. Kincade K. Data mining: Digging for healthcare gold. Insurance & Technology 1998; 23(2): 2-7.
2. Fayyad U, Piatetsky-Shapiro G & Smyth P. Knowledge discovery and data mining towards a unifying framework. Available at: <https://www.aaai.org/Papers/KDD/1996/KDD96-014.pdf>. 1996.
3. Koh HC & Tan G. Data mining applications in healthcare. Journal of Healthcare Information Management 2005; 19(2): 64-72.
4. Han J, Kamber M & Pei J. Data mining: Concepts and techniques. USA: Morgan Kaufmann Publishers Inc; 2005: 5-6.
5. Lee IN, Liao SC & Embrechts M. Data mining techniques applied to medical information. Medical Informatics and the Internet in Medicine 2000; 25(2): 81-102.
6. Obenshain MK. Application of data mining techniques to healthcare data. Infection Control and Hospital Epidemiology 2004; 25(8): 690-5.
7. Samad Soltani Heris T, Lagarizadeh M, Mahmoodvand Z & Zolnoori M. Intelligent diagnosis of asthma using machine learning algorithms. International Research Journal of Applied and Basic Sciences 2013; 4(12): 4041-6.

8. Liao SC & Lee IN. Appropriate medical data categorization for data mining classification techniques. *Medical Informatics and the Internet in Medicine* 2002; 27(1): 59-67.
9. Shouman M, Turner T & Stocker R. Applying k-nearest neighbour in diagnosing heart disease patients. *International Journal of Information and Education Technology* 2012; 2(3): 220-3.
10. Bereznicki BJ, Peterson GM, Jackson SL, Walters EH, Fitzmaurice KD & Gee PR. Data-mining of medication records to improve asthma management. *Medical Journal of Australia* 2008; 189(1): 21-5.
11. Zolnoori M, Zarandi MH, Moin M & Teimorian SH. Designing fuzzy expert system for diagnosing and evaluating childhood asthma, Iran: *Medical Engineering Conference*, 2010.
12. Morice AH, Fontana GA, Sovijarvi AR, Pistolesi M, Chung KF, Widdicombe J, et al . The diagnosis and management of chronic cough. *European Respiratory Journal* 2004; 24(3): 481-92.
13. Toussi M, Lamy JB, Le Toumelin P & Venot A. Using data mining techniques to explore physicians' therapeutic decisions when clinical guidelines do not provide recommendations: Methods and example for type 2 diabetes. *BMC Medical Informatics & Decision Making* 2009; 9(1): 28.
14. Louviere JJ & Woodworth G. Design and analysis of simulated consumer choice or allocation experiments: An approach based on aggregate data. *Journal of Marketing Research* 1983; 20(4): 350-67.
15. Motschnig Pitrik R. Analyzing the notions of attribute, aggregate, part and member in data/knowledge modeling. *Journal of Systems and Software* 1996; 33(2): 113-22.
16. Brisman J, Toren K & Jarvholm B. Asthma and asthma-like symptoms in adults assessed by questionnaires. A literature review. *Chest* 1993; 104(2): 600-8.
17. Jenkins MA, Clarke JR, Carlin JB, Robertson CF, Hopper JL, Dalton MF, et al. Validation of questionnaire and bronchial hyperresponsiveness against respiratory physician assessment in the diagnosis of asthma. *International Journal of Epidemiology* 1996; 25(3): 609-16.
18. Mata Fernandez C, Fernandez-Benitez M, Perez Miranda M & Guillen Grima F. Validation of the Spanish version of the phase III ISAAC questionnaire on asthma. *Journal of Investigational Allergology and Clinical Immunology* 2005; 15(3): 201-10.
19. Abramson MJ, Hensley MJ, Saunders NA & Wlodarczyk JH. Evaluation of a new asthma questionnaire. *Journal of Asthma* 1991; 28(2): 129-39.
20. Shin B, Cole SL, Park SJ, Ledford DK & Lockey RF. A new symptom-based questionnaire for predicting the presence of asthma. *Journal of Investigational Allergology & Clinical Immunology* 2010; 20(1): 27-34.

Data Mining And Analysis: Reporting Results For Patients With Asthma

Samad Soltani Taha¹(MSc.) – Langarizadeh Mostafa²(Ph.D)
Zolnoori Maryam³(MSc.)

1 Ph.D Student in Medical Informatics, Health Information Management Department, School of Allied Medicine, Tehran University of Medical Sciences, Tehran, Iran

2 Assistant Professor, Health Information Management Department, School of Health Management and Information Sciences, Iran University of Medical Sciences, Tehran, Iran

3 Ph.D Student in Health Informatics, Bio Health Informatics Department, School of Informatics and Computing, Indiana University, Indianapolis, USA

Abstract

Received : Mar 2015
Accepted : Jun 2015

Background and Aim: Data mining is a very important branch in deeper understanding of medical data, which attempts to solve problems in the diagnosis and treatment of diseases. One of the most important data mining applications is to examine the existing data patterns. The present study aims to examine the existing data patterns of patients with asthma.

Materials and Methods: This study was performed on 258 patients with respiratory symptoms, who referred to Imam Khomeini and Masih Daneshvari Hospitals in 2009. All records were entered into Excel software, and data mining add-ins were used. Analyses such as key influencers, cluster analysis of patients, and detecting exceptions have been done.

Results: The most common clinical sign of asthma among subjects was severe coughing, which was highly affected by thrills. The data were aggregated into 5 clusters for more general analyses. Their common denominator was then identified and the records with exceptional features were determined. Then, following cost analysis and setting the threshold value at 612, a questionnaire was developed based on data features for diagnosis of asthma.

Conclusion: The developed framework for data mining and analysis is an appropriate tool for knowledge extraction based on the data and their relationships. Meanwhile, it can identify and fill the existing gap in medical decision-making when using clinical guidelines.

Key words: Data Mining, Patient Scale, Asthma

* Corresponding Author:
Langarizadeh M;
E-mail:
Langarizadeh.m@iums.ac.ir