ارائه مدل جهت شناسایی عوامل موثر بر ایجاد بیماری آسم با استفاده از داده کاوی

مرجان قاضی سعیدی[۱]، عباس شیخ طاهری[۲]، نسرین بهنیا فرد۳، فاطمه السادات آقایی میبدی[٤]، روح الله خارا[٥]، مجید کارگربیده[*۱]

۱) گروه مدیریت اطلاعات سلامت، دانشکده پیراپزشکی، دانشگاه علوم پزشکی تهران، تهران، ایران
۲) گروه مدیریت اطلاعات سلامت، دانشکده مدیریت و اطلاع رسانی، دانشگاه علوم پزشکی ایران، تهران، ایران
۳) گروه کودکان، دانشکده پزشکی، دانشگاه علوم پزشکی و خدمات بهداشتی درمانی شهید صدوقی یزد، یزد، ایران
٤) گروه داخلی،بیمارستان شهید صدوقی، دانشگاه علوم پزشکی و خدمات بهداشتی درمانی شهید صدوقی یزد، یزد، ایران
۵) گروه مدیریت اطلاعات سلامت، دانشکده مدیریت و اطلاع رسانی پزشکی، دانشگاه علوم پزشکی تبریز، تبریز، ایران

تاریخ دریافت: ۱۳۹۶/۳/۲۴ **تاریخ پذیرش: ۱۳۹۶/۱۲/۱۵**

چکیده

مقدمه: شناخت عوامل محیطی خطر ایجاد آسم، نقش مهمی در پیشگیری یا کاهش شدت آن ایفا می کند. امروزه می توان این کار را با استفاده از تکنیک های نوین انجام داد. داده کاوی یکی از این تکنیک ها است که کاربردهای فراوانی در زمینه های تشخیص، پیش بینی و درمان دارد، هدف این پژوهش شناسایی عوامل موثر بر ایجاد بیماری آسم و ارائه مدل پیش بینی با استفاده از الگوریتم های داده کاوی است.

مواد و روش ها: این پژوهش از نوع توصیفی با رویکرد کاربردی می باشد. پایگاه داده آن شامل ۲۲۰ رکورد می باشد. داده ها با استفاده از چک لیست و با روش مصاحبه از بیماران مراجعه کننده به یک درمانگاه در سال ۱۳۹۴ جمع آوری گردید. تجزیه و تحلیل داده ها و مدلسازی با استفاده از نرم افزار IBM SPSS Modeler نسخه ۱۴/۲ انجام شده است. در بخش مدل سازی از الگوریتم های درخت تصمیم CHAID و C5، الگوریتم شبکه عصبی و الگوریتم شبکه بیز استفاده شده است.

یافته های پژوهش: در این مطالعه ۱۲ متغیر به عنوان موثرترین فاکتورها تعیین گردید و صحت مدل ایجاد شده بر روی داده های مورد استفاده در الگوریتم CHAID ۷۲/۷۳ درصد، C5 ۶۹/۱ درصد، شبکه بیز ۷۰/۹ درصد و شبکه عصبی ۶۵/۴۵ درصد بوده است.

بحث و نتیجه گیری: یافته ها نشان داد مدل حاصل از الگوریتم درخت تصمیم CHAID از صحت عملکرد(۷۲/۷۳ درصد) بالاتری نسبت به مدل های دیگر برخوردار است. با توجه به متغیرهای پیش بینی کننده و قوانین ایجاد شده برای یک نمونه جدید با ویژگی های مشخص، می توان احتمال ابتلا فرد به بیماری آسم را پیش بینی نمود.

واژه های کلیدی: آسم، داده کاوی، مدل پیش بینی

* نویسنده مسئول: گروه مدیریت اطلاعات سلامت، دانشکده پیراپزشکی، دانشگاه علوم پزشکی تهران، تهران، ایران
Email:majkarbid@yahoo.com

[DOI: 10.29252/sjimu.27.1.203]

مقدمه

شاید بتوان اولین انگیزه کاوش داده را رشد روز افزون آن دانست. در واقع رشد داده به حدی است که تنها در صورت وجود ابزار مکـانیزه برای بررسی آن می توان امیدوار به استفاده از آن بود. زیرا در غیر این صورت هم زمان با تحلیل حجم کوچکی داده، حجم بسیار زیادی از همان داده در حال تولید شدن است که هرگز فرصتی برای کاوش آن وجود نخواد داشت(۱). داده کاوی در مراقبت بهداشتی شاخه بسیار مهمی در تشخیص و فهم عمیق تر داده های پزشکی می باشد. داده کاوی بهداشتی در صدد حل مسائل دنیای واقعی در تشخیص و درمـــان بیماری ها می باشد. یکی از مهم ترین کاربردهای داده کاوی در حیطه استخراج دانش، بررسی الگوهای موجود در داده ها می باشد که منجر به شناخت عوامل موثر در ابتلا به بیماری و عوامل کاهنده یا تشدید کننده آن می شود(۲).

آسم بیماری التهابی رایج مزمن مجاری هوایی با ویژگی هایی مثل علائم متعدد و عودکننده، انسداد برگشت پذیر جریان هوا و اسپاسم برونش می باشد. نشانه های رایج آن خس خس سینه، سرفه و تنگی نفس می باشد و به علت ترکیبی از عوامل ژنتیکی و محیطی ایجاد می شود(۳،۴). به عبارتی اصولاً آسم به دلیل قرار گرفتن افراد مستعد از نظر ژنتیکی در برابر عوامل خطر محیطی ایجاد می گردد(۵).

آسم یک بیماری ناهمگون از تقابل بین عوامل ژنتیک و محیط است. عوامل خطرساز متعددی برای آسم مطرح شده است که هنوز نقش دقیق بیشتر آن ها به خوبی مشخص نشده است که شامل استعداد ژنتیکی، آتوپی، افزایش پاسخ دهی مجاری هوائی، نژاد، جنس، آلرژن ها، سیگار، عفونت های تنفسی، چاقی و عفونت های ویروسی زودرس می باشند(۶). اگر چه ژنتیک مهم ترین نقش را در بروز آسم دارد، ولی این سرعت افزایش که در طی دو دهه اخیر دیده می شود به وسیله تغییرات ژنتیکی توجیه نمی شود(۷). تشخیص بیماری آسم به دلیل پیچیدگی الگوها و نشانه ها از مسائل چالش برانگیز می باشد(۲). شناخت عوامل خطر ایجاد آسم نقش مهمی در پیشگیری از ایجاد و یا کاهش شدت علائم آن ایفا می کند. گر چه اطلاعات ما در مورد عوامل ژنتیکی این بیماری به طور مرتب در حال افزایش است، تغییر عوامل شناخته شده محیطی هم چنان بهترین رویکرد مـــوجود به این مـشکل می باشد(۸).

در مطالعات بررسی شده عواملی از قبیل جنسیت پسر در بین کودکان ۷-۶ سال(۱۱-۳،۹)، تولد فرد از طریق زایمان سزارین(۱۲،۱۳)، شاخص توده بدنی کم مادر در زمان حاملگی(۱۳)، سن حاملگی کم مادر(۱۳)، فتوتراپی نوزادی و طول مدت آن(۷)، وزن بالا هنگام تولد(۱۴)، وزن پایین هنگام تولد(۳،۱۵)، عدم تغذیه با شیر مادر(۳،۱۱،۱۵)، سابقه آسم در فامیل درجه یک(۳،۶،۱۳)، سابقـــه آلـــرژی در خانواده(۹،۱۵،۱۶)، حســـاسیـــت تنفـــسی در یکی از برادران یا خواهران(۱۱،۱۷)، وجود بیماری رینیت در فرد(۳،۹،۱۸)، عفونت های مکرر تنفسی(۱۵)، عفونت با رینوویروس (۱۷)، چـــاقی(۱۹)، مـــصرف سیـــگار توسـط والدین(۳،۱۱،۱۵،۱۶،۲۰،۲۱)، نگـــهداری از حیوانات دست آموز(۵،۱۱)، آلاینده های هوا(عمدتاً منوکسید کربن و ازن)(۲۲)، آلرژن های داخل خانه(مثل مایت غبار خانگی، اسپور قارچ ها و حیوانات)(۲۳)، به عنوان عامل موثر بر بیماری آسم شناخته شده اند. هم چنین در این ارتباط مطالعه ای با استفاده از تحلیل هایی از قبیل اثرگذاری کلیدی، دسته بندی بیماران و تشخیص استثنائات صورت گرفت که نتایج آن حاکی از آن است که بین ابتلا به آسم و سرفه های شدید همبستگی بالایی وجود دارد و این عامل را می توان فراوان ترین نشانه آسم دانست. هم چنین نتایج حاصل از تحلیل اثرگذاری کلیدی نشان داد که آسم در اکثر افراد در پی واکنش های هیجانی تشدید می شود و خس خس و تشدید بیماری در پی فعالیت های بدنی نیز از دیگر نشانه های اصلی آسم می باشد(۲).

مطالعات بررسی شده در مورد عوامل موثر بر بیماری آسم اغلب تک بعدی بوده و هر یک تعداد محدودی از ریسک فاکتورها را مورد بررســـی قرار داده اند و بعضاً نتایج ضد و نقیضی به دست آورده اند. در ضمن این که تمـــام مطـــالعات از روش ها و نرم افزارهای آماری جهت تحـــلیل داده ها استفاده کرده اند و هیچ یک از مطالعات از ابزارهای جدید مانند

[DOI: 10.29252/sjimu.27.1.203]

داده کاوی که برای شناسایی الگوهای پنهان در حجم وسیعی از داده ها استفاده می شوند استفاده نکرده اند. استخراج دانش و قوانین از پایگاه های داده به عنوان پایه اساسی برای سیستم پشتیبان تصمیم و پزشکی مبتنی بر شواهد هستند. با توجه به حجم عظیم و رشد روز افزون داده ها در پایگاه داده، استخراج قوانین و دانش از این انبارهای داده مستلزم استفاده از عناصر هوشمندی چون الگوریتم های داده کاوی است و همچنین نتایج فرآیند داده کاوی می توانند به شکل قوانین بیان شوند و این قوانین می توانند با پارامترهای ضریب اطمینان وزن دهی شوند(۲۵). در این مطالعه با استفاده از تکنیک های داده کاوی در میان داده های زیادی که از بیماران مبتلا به آسم تولید می شود، مدلی را جهت شناسایی عوامل موثر بر بیماری آسم و پش بینی آن در افراد به دست آورده ایم.

مواد و روش ها

این پژوهش از نوع توصیفی با رویکرد کاربردی می باشد. جامعه پژوهش بیماران مراجعه کننده به درمانگاه های تخصصی و فوق تخصصی بقایی پور، خاتم الانبیا(ص) و امام علی(ع) دانشگاه علوم پزشکی شهید صدوقی یزد با بیماری تنفسی که یکی یا همه علایم سرفه، تنگی نفس و خس خس سینه(ویز) را داشته اند، بوده است. نمونه پژوهش، بیماران با بیماری تنفسی که یکی یا همه علایم سرفه، تنگی نفس و خس خس سینه(ویز) را داشته اند به صورت مقطعی، با روش سرشماری و به صورت آینده نگر، از تاریخ ۹۴/۸/۱ تا ۹۴/۱۰/۱ در نظر گرفته شدند. در این بازه زمانی تعداد ۲۲۰ بیمار با داشتن شرایط ورود به مطالعه مراجعه نموده و با توجه به استفاده از روش های داده کاوی پیش بینانه، بیماران بر اساس تشخیص نهایی پزشک به دو گروه تقسیم شدند. گروه اول ۱۴۰ بیمار مبتلا به بیماری آسم و گروه دوم ۸۰ بیمار مبتلا به بیماری تنفسی غیر آسم بوده اند.

ابزار جمع آوری داده ها، چک لیست طراحی شده بر اساس متغیرهای شناسایی شده از مستندات کتابخانه ای و پایگاه مقالات می باشد. این چک لیست در قالب چهار جنبه ریسک فاکتورهای دموگرافیک، وراثت، محیطی و سابقه بیماری طراحی شده بود. داده ها نیز با روش مصاحبه از همه بیماران تنفسی جمع آوری گردید. سپس تشخیص نهایی بیمار بر اساس تشخیص پزشک در چک لیست ثبت گردید. در مورد بیماران کودک، مصاحبه با مادر وی جهت اخذ اطلاعات انجام گرفت. معیار ورود به این مطالعه داشتن بیماری تنفسی و مراجعه به مراکز مورد نظر در مطالعه بوده و معیار خروج از مطالعه عدم رضایت بیمار و یا والدین بیمار برای مصاحبه و اخذ اطلاعات بود. تجزیه و تحلیل داده ها به وسیله نرم افزار داده کاوی IBM SPSS Modeler نسخه ۱۴/۲ انجام گردید. در نهایت مدل های پیش بینی طراحی شده، مورد ارزیابی قرار گرفته و مدل نهایی ارائه گردیده است. به طور کلی این پژوهش در چند مرحله تحت عنوان«تعیین متغیرها»، «آماده سازی داده ها»، «پیش پردازش»، «مدل سازی» و «ارزیابی» انجام شد.

تعیین متغیرها: در این مرحله به شناسایی و توصیف متغیرهای پژوهش پرداخته شد. در این پژوهش برای تعیین عوامل احتمالی و شناخته شده تاثیرگذار در ابتلا به آسم، ابتدا منابع و مستندات موجود در پایگاه داده های PubMed، Google scholar، SID مورد بررسی قرار گرفت که با استفاده از این پایگاه داده ها عوامل قابل بررسی در قالب چهار جنبه ریسک فاکتورهای دموگرافیک، وراثت، محیطی و سابقه بیماری شناسایی گردید. سپس عواملی با نظر و بر اساس تجربه بالینی پزشکان مشاور(یک نفر فوق تخصص ایمونولوژی، آسم و آلرژی عضو هیات علمی دانشگاه علوم پزشکی یزد، و یک نفر فوق تخصص ریه عضو هیات علمی دانشگاه علوم پزشکی یزد) به عوامل استخراج شده از پایگاه داده ها اضافه گردید. در نهایت چک لیستی بر اساس این متغیرها برای جمع آوری داده های مورد نیاز تهیه گردید.

آماده سازی داده ها: در این مرحله پس از جمع آوری داده ها از طریق چک لیست، به توصیف داده های جمع آوری شده با استفاده از نرم افزار SPSS نسخه ۱۶ و در قالب جداول توزیع فراوانی، پرداخته شد. در هنگام ورود داده به نرم افزار SPSS متغیر کیفی شغل بر اساس تماس با مواد محرک در محل کار و شغل های شناخته شده به عنوان عامل موثر بر بیماری

[DOI: 10.29252/sjimu.27.1.203]

آسم در مطالعات دیگر و برای این که بتوان این عامل را در مدل سازی لحاظ نمود، به سه گروه پرخطر، متوسط و کم خطر تقسیم شد و هر یک از مشاغل با نظر اساتید مشاور در یکی از گروه های فوق قرار گرفت.

پیش پردازش: عملیات پیش پردازش شامل فرآیندهایی از قبیل: تصحیح و یا حذف داده های بی مقدار، تعیین محدوده مجاز و تصحیح مقادیر غیرمجاز، انجام محاسبات مجدد برای برخی ویژگی ها و تبدیل آن ها به ویژگی های دیگر بر روی داده ها و به منظور بهبود داده ها جهت استفاده در ابزار تحلیل می باشد. پس از ورود داده به نرم افزار SPSS و آماده سازی داده ها، فایل داده های آماده سازی شده به نرم افزار IBM SPSS Modeler نسخه ۱۴/۲ وارد گردید تا بررسی های کمی و کیفی قبل از مدل سازی بر روی داده ها انجام گیرد. با توجه به استفاده از روش مصاحبه برای جمع آوری داده در سایر فیلدها مقدار ثبت نشده ای وجود نداشت. در پایان این مرحله متغیرهای مهم و تاثیرگذار برای ساخت مدل با استفاده از گره انتخاب ویژگی و مدل رگرسیون لجستیک گزینش گردید.

مدل سازی: در این پژوهش از نرم افزار IBM SPSS Modeler نسخه ۱۴/۲ جهت مدل سازی استفاده گردیده است. با توجه به هدف پژوهش، نوع نمونه و متغیرها از الگوریتم های پیش بینی کننده درخت تصمیم C5، درخت تصمیم CHAID، شبکه بیز و شبکه عصبی استفاده شده است. متغیرهایی از قبیل علایم بیماری(سرفه، تنگی نفس و خس خس) و سن بیمار که تاثیری در ایجاد بیماری نداشتند برای جلوگیری از سوگیری، از جریان مدل سازی حذف گردید.

ارزیابی: در این مرحله پس از مدل سازی به ارزیابی نتایج حاصل از مدل سازی پرداخته شده است. قبل از اجرای دسته بندی، داده ها به دو دسته آموزش(۸۰ درصد)، برای ساخت روش دسته بندی و آزمایش(۲۰ درصد) برای آزمایش دسته بندی تقسیم شدند. در این پژوهش از ماتریس در هم ریختگی به علت سادگی و استفاده رایج توسط پژوهشگران حوزه علوم پزشکی برای ارزیابی مدل های به دست آمده مورد استفاده قرار گرفت. ماتریس در هم ریختگی(اغتشاش) معیارهای مختلفی برای ارزیابی روش های دسته بندی دارد که می توان حساسیت، ویژگی، دقت و صحت را نام برد. معیارها در زیر تعریف شده اند:

حساسیت(۶۶،۷۳):

$$\text{حساسیت} = \frac{\text{تعداد داده های کلاس مثبت که درست دسته بندی شدند}}{\text{تعداد کل داده های کلاس مثبت}}$$

ویژگی(۶۶،۷۳):

$$\text{ویژگی} = \frac{\text{تعداد داده های کلاس منفی که درست دسته بندی شدند}}{\text{تعداد کل داده های کلاس منفی}}$$

دقت(۶۶،۷۳):

$$\text{دقت} = \frac{\text{تعداد داده های کلاس مثبت که درست دسته بندی شدند}}{\text{تعداد داده های کلاس مثبت که درست دسته بندی شدند+تعداد داده های کلاس منفی که به اشتباه مثبت دسته بندی شدند}}$$

صحت(۶۶،۷۳):

$$\text{صحت} = \frac{\text{تعداد داده های کلاس منفی}}{\text{تعداد کل داده ها}} + \text{ویژگی} + \frac{\text{تعداد داده های کلاس مثبت}}{\text{تعداد کل داده ها}} + \text{حساسیت}$$

[DOI: 10.29252/sjimu.27.1.203]

یافته های پژوهش

یافته های مربوط به مهم ترین متغیرهای پیش بینی کننده در هر مدل و گره انتخاب ویژگی، جهت مشاهده و مقایسه بهتر در جدول شماره ۱ آورده شده است.

جدول شماره ۱. مهم ترین متغیرهای پیش بینی کننده در مدل های مختلف

الگوریتم / متغیر	گره Feature Selection	درخت تصمیم		شبکه بیزین	شبکه عصبی
		CHAID	C5		
۱	سن شروع بیماری	سن شروع بیماری	سابقه فامیلی حساسیت	–	سن شروع بیماری
۲	مصرف سیگار در بیمار	سابقه فامیلی آلرژی	شغل	–	مصرف سیگار در بیمار
۳	محل سکونت	مصرف سیگار بیمار	سابقه آلرژی در فرد	–	سابقه فامیلی آلرژی
۴	سابقه فامیلی آلرژی	محل سکونت	شاخص توده بدنی	–	شاخص توده بدنی
۵	نوع زایمان	مواجهه غیر فعال با دود سیگار	سابقه مصرف استامینوفن	–	مواجه غیر فعال با دود سیگار
۶	مواجهه غیر فعال با دود سیگار	محل تولد	محل سکونت	–	محل سکونت
۷	سابقه آلرژی در بیمار	–	مواجهه غیر فعال با دود سیگار	–	شغل
۸	شغل	–	سن شروع بیماری	–	سابقه آلرژی بیمار
۹	محل تولد	–	–	–	نوع زایمان
۱۰	ریفلاکس مری–معدوی	–	–	–	سابقه مصرف استامینوفن
۱۱	شاخص توده بدنی	–	–	–	محل تولد
۱۲	سابقه مصرف استامینوفن	–	–	–	ریفلاکس مری–معدوی

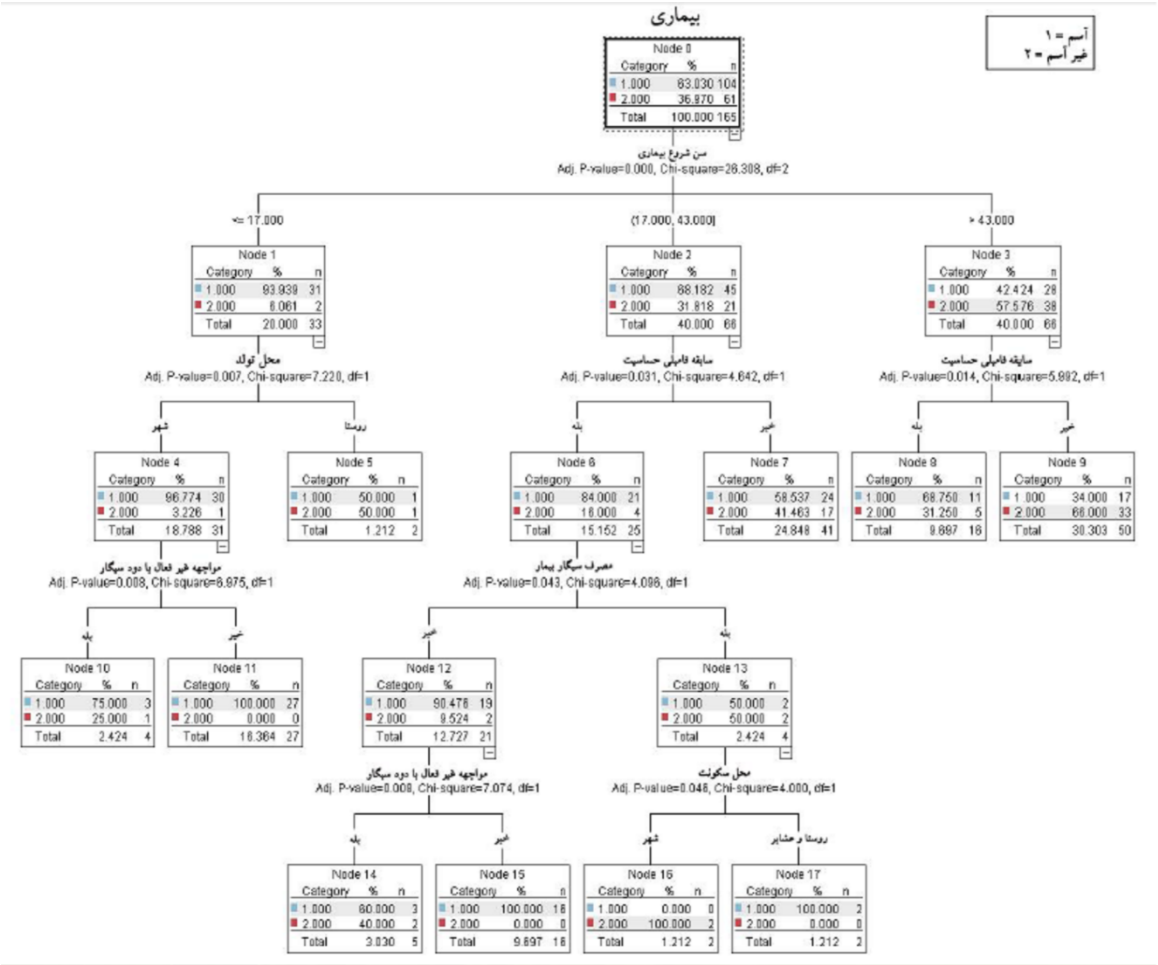
یافته های پژوهش صحت، دقت، حساسیت و ویژگی هر یک از الگوریتم های مورد بررسی را نشان می دهد جدول شماره ۲ که بیشترین صحت در بین مدل های ساخته شده برای داده های آزمایش به درخت CHAID و کمترین صحت به شبکه عصبی اختصاص دارد. بیشترین دقت در بین مدل های ساخته شده به شبکه بیز و کمترین دقت به شبکه عصبی اختصاص دارد. بیشترین حساسیت در بین مدل های ساخته شده به درخت CHAID و کمترین به شبکه عصبی اختصاص دارد. بیشترین ویژگی در بین مدل های ساخته شده به شبکه بیز و کمترین ویژگی به درخت CHAID اختصاص دارد. در نهایت الگوریتم CHAID به عنوان الگوریتم برتر در این مطالعه انتخاب شده است.

جدول شماره ۲. مقایسه مولفه های مربوط به ارزیابی همه مدل ها

گروه / مولفه		صحت	دقت	حساسیت	ویژگی
CHAID	آموزش	۷۳/۹۴	۷۷	۸۳/۶	۵۷/۳
	آزمایش	۷۲/۷۳	۷۳/۳	۹۱/۶	۳۶/۸
C5	آموزش	۸۹/۷	۹۰/۶۵	۹۳/۲۶	۸۳/۶
	آزمایش	۶۹/۱	۷۳/۱	۸۳/۳	۴۲/۱
شبکه بیز	آموزش	۷۸/۷۹	۸۲/۸۵	۸۳/۶۵	۷۰/۴۹
	آزمایش	۷۰/۹	۷۸/۳۷	۸۲/۸۵	۵۵/۵۵
شبکه عصبی	آموزش	۷۶/۹۷	۸۱/۱۳	۸۲/۶۹	۶۷/۲۱
	آزمایش	۶۵/۴۵	۷۰/۷۳	۸۰/۵۵	۳۶/۸۴

[DOI: 10.29252/sjimu.27.1.203]

الگوریتم CHAID. متغیرهای مهم برای پیش بینی بیماری در مدل CHAID سن شروع بیماری(۰/۴۱)، سابقه فامیلی آلرژی(۰/۲۶)، مـــصرف سیـــگار بیمار (۰/۱۴)، محل سکونت(۰/۰۸)، مواجهه غیر فعال با دود سیگار(۰/۰۶)، محل تولد(۰/۰۵) می باشد(شکل شماره ۱).



شکل شماره ۱. مدل ایجاد شده توسط الگوریتم CHAID

مدل های مختلف بر اساس چهار مولفه صحت، دقت، حساسیت و ویژگی ارزیابی و مقایسه شده اند. ارزیابی انجام شده مطابق با فرمول های ارائه شده در روش کار و با استفاده از داده های حاصل از ماتریس اغتشاش بر اساس داده های آزمایش، مطالبق جدول شماره ۳ است.

جدول شماره ۳. ماتریس در هم ریختگی مدل CHAID

			پیش بینی شده	
			آسم	غیر آسم
واقعی	آموزش	آسم	۸۷	۱۷
		غیر آسم	۲۶	۳۵
	آزمایش	آسم	۳۳	۳
		غیر آسم	۱۲	۷

[DOI: 10.29252/sjimu.27.1.203]

مطابق جدول شماره ۴ صحت عملکرد مدل ساخته شده به وسیله الگوریتم CHAID برای داده های آموزش ۷۳/۹۴ درصد و برای داده های آزمایش را ۷۲/۷۳ درصد است.

جدول شماره ٤. ارزیابی عملکرد مدل CHAID

مولفه / گروه	صحت	دقت	حساسیت	ویژگی
آموزش	۷۳/۹۴	۷۷	۸۳/۶	۵۷/۳
آزمایش	۷۲/۷۳	۷۳/۳	۹۱/۶	۳۶/۸

بحث و نتیجه گیری

یافته ها نشان می دهد مدل حاصل از الگوریتم درخت تصمیم CHAID با صحت ۷۲/۷۳ درصد، دقت ۷۳/۳ درصد، حساسیت ۹۱/۶ درصد و ویژگی ۳۶/۸ درصد به عنوان مدل بهینه در این مطالعه انتخاب گردیده است.

نکته قابل توجه در این یافته ها بالا بودن مولفه حساسیت و پایین بودن مولفه ویژگی در همه مدل ها به ویژه مدل منتخب CHAID می باشد که با توجه به تعریف این مولفه ها، بالا بودن حساسیت نشان می دهد بیماران گروه آسم به خوبی پیش بینی می شوند و پایین بودن ویژگی نشان می دهد بیماران تنفسی غیر آسم به درستی پیش بینی نمی شوند که دلیل آن را می توان متفاوت بودن و تنوع بیماری های قرار گرفته در گروه بیماران تنفسی غیر آسم دانست. هم چنین با توجه به فرمول محاسبه صحت، پایین بودن ویژگی باعث کاهش صحت عملکرد مدل نیز می شود.

صفدری و همکاران در مطالعه خود پس از مقایسه عملکرد درخت تصمیم گیری و شبکه عصبی در پیشگویی ابتلا به آنفارکتوس قلبی، مدل حاصل از الگوریتم درخت تصمیم C5 را با صحت عملکرد ۹۳/۴ درصد به عنوان مدل بهتر در این زمینه انتخاب نمود.

علیزاده و همکاران نیز در مطالعه خود صحت مدل ساخته شده توسط شبکه عصبی برای تشخیص بیماری آسم را ۱۰۰ درصد ارزیابی کرده اند. این در حالی است که در این مطالعه الگوریتم درخت تصمیم CHAID به عنوان الگوریتم بهینه شناخته شده است. هم چنین عامری و همکاران از دو الگوریتم درخت تصمیم C5 و شبکه عصبی برای بررسی احتمال بروز عوارض میکرو واسکولار، ماکرو واسکولار و یا هر دو نوع عارضه در بیماران دیابتی استفاده کرده اند. صحت مدل ایجاد شده توسط الگوریتم درخت تصمیم با ۸۹/۷۴ درصد در مقایسه با شبکه عصبی مصنوعی با ۵۱/۲۸ درصد بیشتر ارزیابی شده است. می توان گفت نتایج این مطالعه به مطالعه حاضر تقریبا هم راستا می باشد. مطالعه عامری و همکاران(۲۵) مناسب بودن داده ها، به کارگیری پیش پردازش مناسب و راه کار مناسب داده کاوی را عوامل موثر در کسب نتایج بهتر در ارتباط با داده های پزشکی نشان می دهد. به نظر می رسد به طور کلی صحت مدل های منتخب در رابطه با پیش بینی بیماری ها کمتر از مدل های منتخب در حوزه تشخیص باشد. به طور کلی می توان گفت مقایسه مولفه های ارزیابی مدل ها در حوزه های گوناگون به خصوص ابتلا به بیماری ها، به دلیل اختلاف در ماهیت و پیچیدگی موضوع، کار درستی نباشد.

References

1.Saniee Abadeh M, Mahmoudi S, Taherparvar M. Data mining application. 1th ed. Tehran Niaz Danesh Publication. 2012.

2.Samadsoltany T, Langarizadeh M. Data analysis and data mining in patients with asthma and report results. 2 Ed National Con Comput Sci Sanandaj2013; Iran.

3.Bilan N, Shiva S. [Risk factors in children 8-2 years old with asthma outpatient clinic of Tabriz University of medical sciences]. Med J Tabriz Uni Sci Health Serv 2007;29:47-50. (Persian)

4.Abedi S, Mahmoudi R, Sharifpour A, Azadeh H, Aliyali M, Abedian Kenari S, et

[DOI: 10.29252/sjimu.27.1.203]

al. [Risk factors associated with persistent airflow limitation in patients with severe Asthma]. J Mazandaran Uni Med Sci 2015;24:374-9. (Persian)
doi: 10.1378/chest.07-0713
5.Sharifi L, Pourpak Z, Bokaie S, Karimi A, Gharegozloo M, Movahhedi M, et al. Pet ownership and risk of asthma: a case-controlled study. Tehran Uni Med J 2008;66:338-42. (Persian)
6.Varqaee A. [Comparison of Asthma risk factors and absolute eosinophil count in asthmatic patients and control group]. Ardabil Uni Med Sci J2013; 3:21-8. (Persian)
7.Mosayebi Z, Heidarzadeh M, Movahedian AH, Abedi AR, Mousavi SGA, Eslamian MR. [The correlation between neonatal phototherapy and risk of childhood asthma in children referred to pediatric clinic of Kashan Shahid Beheshti Hospital in 2009]. Feyz J 2011;15:38-43. (Persian)
8.Faghihinia J, Asadyan A, Sohrabian N. The relationship between BMI and asthma in children. J Isfahan Med Sch 2009;27:468-77.
9.Bazazi H, Gharaghozlou M, Kasaee M, Parsikia A, Zahmatkesh H, Alikhani L, et al. [Wheezing and Asthma prevalence and associated factors in primary school students in the city of Gorgan in 2003]. Pejouhandeh J2006;4:259-64. (Persian)
10.Farrokhi S. [Prevalence of asthma and allergic diseases in people with school age 7-6 and 14-13 years based on ISAAC at Bushehr]. Bushehr Uni Med Sci Rep2014. (Persian)
11.Rajaeifard A, Moosavizadeh A, Pourmahmoudi A, Naeimi E, Hadinia A, Karimi A. [Evaluation of prevalence and related factors of pediatric asthma in children under six years old with logistic regression and probit]. Armaghane Danesh J 2010;16:272-81. (Persian)
12.Kolokotroni O, Middleton N, Gavatha M, Lamnisos D, Priftis KN, Yiallouros PK. Asthma and atopy in children born by caesarean section effect modification by family history of allergies-a population based cross sectional study. BMC Pediatr2012;12:179. doi: 10.1186/1471-2431-12-179
13.Metsala J, Kilkkinen A, Kaila M, Tapanainen H, Klaukka T, Gissler M, et al. Perinatal factors and the risk of asthma in childhood a population based register study in Finland. Am J Epidemiol 2008;168:170-8. doi: 10.1093/aje/kwn105
14.Heidarzadeharani M, Hajirezaei M, Ahmadi A. [Prevalence of abnormal birth weight among the asthmatic children 5-15 years referred to Kashan Asthma and allergy clinic during 2007-2008. Feyz 2013;17:300-4]. (Persian)
15.Inanluo S, Amin R. [Study of risk factors for asthma in children]. J Isfahan Med Sch 2002;20:3-8. (Persian)
16.Haghbin S, Ebrahimi S, Rezaei M, Pourmahmoodi A. [A study of some environmental factors in Asthma among children aged 6 months to 6 years in Yasuj]. Armaghan Danesh J 2003;8:33-8.
17.Jackson DJ, Gangnon RE, Evans MD, Roberg KA, Anderson EL, Pappas TE, et al. Wheezing rhinovirus illnesses in early life predict asthma development in high risk children. Am J Res Crit Care Med2008;178:667-72. doi: 10.1164/rccm.200802-309OC
18.Guerra S, Sherrill DL, Martinez FD, Barbee RA. Rhinitis as an independent risk factor for adult onset asthma. J Allerg Clin Immunol 2002;109:419-25.doi: 10.1067/mai.2002.121701
19.Taylor B, Mannino D, Brown C, Crocker D, Twumbaah N, Holguin F. Body mass index and asthma severity in the national Asthma survey. Thorax 2008;63:14-20.doi: 10.1136/thx.2007.082784
20.Karimi M, Mirzaee M. [Relationship exposure to the smoke and the prevalence of asthma and allergies in children in Yazd]. J Hormozgan Uni Med Sci2007;11:291-5. (Persian)
21.Sharifi L, Pourpak Z, Bokaie S, Karimi A, Movahedi M, Gharaghozlou M, et al. [Childhood asthma prevalence and parents daily cigarette smoking: a case-control study]. Tehran Uni Med J 2009;67:655-60. (Persian)
22.Khamutian R, Dargahi A, Soltanian M, Najafi F, Afshari A. [The relationship of air pollution and asthma patients admitted to hospitals in Kermanshah 2008-2009]. J Kermanshah Uni Med Sci 2015;18:568-92. (Persian)
doi: 10.22110/jkums.v18i10.1862

[DOI: 10.29252/sjimu.27.1.203]

[The relation of knowledge, attitude and self management behaviors in asthmatic patients with controlling Asthma]. Zahedan J Res Med Sci 2012;14:49-55. (Persian)
25.Ameri H, Alizadeh S, Barzegari A. Knowledge extraction of diabetics data by decision tree method. J Health Adm2013;16:58-72.

23.Alizadeh B, Safdari R, Zolnoori M, Bashiri A. Developing an intelligent system for diagnosis of Asthma based on artificial neural network. Acta Inform Med2015;23:220. doi: 10.5455/aim.2015.23.220-223
24.Mazloomi SS, Abbacimoghaddamniasar A, Saba MA, Morovati MA, Fallahzadeh H.

[DOI: 10.29252/sjimu.27.1.203]

A Proposed Model to Identify Factors Affecting Asthma using Data Mining

Ghazisaeedi M[1], Sheikhtaheri A[2], Behniafard N[3], Aghaiemaybodi F[4], Khara R[5], Kargarbideh M[1]*

(Received: June 14, 2017 Accepted: March 6, 2018)

Abstract

Introduction: The identification of asthma risk factors plays an important role in the prevention of the asthma as well as reducing the severity of symptoms. Nowadays, the identification process can be performed using modern techniques. Data mining is one of the techniques which has many applications in the fields of diagnosis, prediction, and treatment. This study aimed to identify the effective factors on asthma to provide a predictive model using data mining algorithms.

Materials & Methods: This descriptive study with a practical approach included 220 data bases. The data were collected using a checklist and interviews from the patients referred to clinical centers of Shahid Sadoughi Hospital in Yazd, Iran, during 2014. The data were analyzed in SPSS IBM Modeler software (Version 14.2). Moreover, the CHAID decision tree,C5 algorithm, neural network algorithm, and Bayesian network algorithm were utilized in the modeling.

Findings: In total, 12 variables were determined as the most influential factors in this study. The accuracy of the model on the data was estimated at 72.73%, 69.1%, 70.9%, and 65.45% in the CHAID algorithm, C5, Bayesian network, and the neural network, respectively.

Discussion & Conclusions: According to the results, the performance accuracy of the model obtained from CHAID decision tree algorithm (73/72%) was higher than that of the other models. Moreover, an individual's risk of asthma can be predicted with regard to the predictive factors and the established rules for a new sample with distinctive features.

Keywords: Asthma, Data Mining, Prediction model

1. Dept of Health Information Management, Faculty of Allied Medical Sciences, Tehran University of Medical Sciences, Tehran, Iran

2. Dept of Health Information Management, Faculty of Health Management and Information Sciences, Iran University of Medical Sciences,Tehran,Iran

3. Dept of Pediatrics, Faculty of Medicine, Shahid Sadoughi University of Medical Sciences,Yazd, Iran

4. Dept of Internal Medicine, Faculty of Medicine , Shahid Sadoughi General Hospital, Shahid Sadoughi University of Medical Sciences, Yazd, Iran

5. Dept of Health Information technology, Faculty of Management and Medical Informatics, Tabriz University of Medical Science, Tabriz, Iran

[DOI: 10.29252/sjimu.27.1.203]