فصلنامه رویکردهای پژوهشی نوین در
مدیریت و حسابداری
سال چهارم، شماره ۳۳، تابستان ۱۳۹۹- جلد سوم، ص ۶۵-۵۳

Journal of New Research Approaches
in Management and Accounting
ISSN: 2588-4573
http://www.majournal.ir

کاربرد داده کاوی در پیش بینی وضعیت بیماران نادر با استفاده از درخت تصمیم گیری؛ مطالعه موردی سامانه بنیاد بیماری های نادر ایران (سبنا)

محسن سجودی [۱]
فریبا ابراهیم بابایی [۲]

تاریخ دریافت: ۱۳۹۹/۰۴/۲۵ تاریخ چاپ: ۱۳۹۹/۰۴/۲۸

چکیده

پایگاه داده ها در حوزه ی سلامت حاوی میزان وسیعی از داده های بالینی است که کشف ارتباطات و الگوها در آن می تواند به دانش جدید پزشکی بیانجامد. امروزه با توجه به پیدایش نظام‌های اطلاعات یکپارچه و رشد فن آوری اطلاعات، این مهم بیش از پیش نمایان شده است. این مطالعه با هدف شناسایی مزایای بالقوه‌ای که داده‌کاوی می‌تواند به بخش بهداشت و درمان، با استفاده از داده‌های بنیاد بیماری های نادر ایران به عنوان مطالعه موردی ارائه دهد، انجام پذیرفته است. معمول‌ترین روش داده‌کاوی که درخت تصمیم‌گیری می‌باشد برای تولید مدل پیش‌بینی با مصورسازی درخت برای انجام تحلیل پیش بینی بیماری نادر مورد استفاده قرارگرفته است. تمام مراحل فرایند داده‌کاوی با ابزاری به نام وکا (WEKA) انجام شده است. علاوه بر این، از وکا برای ارزیابی عملکرد پیش‌بینی از طریق اندازه‌گیری دقت، ویژگی و تحلیل حساسیت استفاده شده است. از جمله نتایج پژوهش حاضر، برخی از عواملی است که مراکز حمایتی بیمار می‌توانند در هنگام پیش‌بینی هزینه‌های درمان بیمار مورد توجه قرار دهند. شاخص جنس یا سن، به شدت تحت تاثیر مدت زمان بستری بیمار قرار می‌گیرد. بیمار سالخورده با بیماری نادر باید در مراکز درمانی تحت مراقبت و مدت زمان بستری طولانی‌تری نسبت به افراد جوان‌تر قرار گیرد. در نتیجه، درخت تصمیم‌گیری یک روش مفید و آموزنده برای انجام داده‌کاوی پیش‌بینی شده است.

واژگان کلیدی

داده‌کاوی، مدل پیش‌بینی، بیماری نادر، درخت تصمیم‌گیری، تجزیه و تحلیل حساسیت، وکا

۱. دانشجوی دکتری مدیریت تحقیق در عملیات، دانشگاه فردوسی مشهد، ایران (mohsened@gmail.com)
۲. دانشجوی دکتری روانشناسی، دانشکده ادبیات و علوم انسانی، دانشگاه لرستان، خرم آباد، ایران (faribaebrahimbabaie@yahoo.com)

۱. مقدمه

سازمان‌های مراقبت بهداشتی-درمانی به دلیل پیچیدگی و تکرار فعالیت‌ها، سرشار از داده‌های بالینی هستند اما نیازمند تکنیک‌ها و ابزارهایی از قبیل داده‌کاوی می‌باشند که این داده‌ها را به اطلاعات مفید تبدیل نمایند (گیوسیدی[1]،۲۰۰۳). داده‌کاوی یا استخراج دانش از پایگاه داده‌ها[2] فرایند مهم شناسایی الگوهای معتبر، جدید و قابل فهم در میان انبوهی از داده‌ها است. مفهوم داده‌کاوی شامل الگوریتم‌ها و متدهایی است که باعث استخراج اطلاعات از داده‌ها می‌شود (کلیرلی[3]،۲۰۰۶). داده‌ها در عصر امروزی یعنی عصر اطلاعات، عمده‌ترین دارایی برای سازمان‌های سلامت بوده است (پیاتسکی[4]،۲۰۰۰) و موفقیت سازمان‌های سلامت در گروی جمع‌آوری، ذخیره و تحلیل آن‌ها است (بالی[5]، ۲۰۰۵). با این وجود، جمع‌آوری و ذخیره‌ی میزان زیادی از داده‌ها می‌تواند یک نوع اتلاف محسوب شود؛ مگر این که داده‌ها به شکل سودمند استفاده شده و تبدیل به یک منبع مالی برای سازمان گردد. برای تبدیل این ارزش بالقوه به اطلاعات استراتژیک، بسیاری از سازمان‌ها به داده‌کاوی روی آورده‌اند (یوسفی و همکاران، ۲۰۱۶)؛ چرا که به واسطه‌ی داده‌کاوی امکان کشف روابط، روندها و الگوهای مخفی بین داده‌ها و دستیابی به دانش نوین در زمینه‌ی چالش‌های آشکار و نهان سازمان میسر خواهد شد. هدف این مطالعه شناسایی مزایای بالقوه‌ای است که داده‌کاوی می‌تواند به بخش سلامت در ایران ارائه نماید. این مطالعه براساس اطلاعات بیماران نادر متعلق به بنیاد بیماری‌های نادر ایران برای پیش‌بینی عوامل موثر بر بیماری‌های نادر با استفاده از درخت تصمیم‌گیری به عنوان تکنیک داده‌کاوی بهره گرفته است.

۲. مبانی نظری و ادبیات پژوهش

داده کاوی در اواخر دهه ی 1980 پدیدار گشت، در دهه ی ۱۹۹۰ گام های بلندی در این شاخه از علم برداشته شد و انتظار می‌رود در قرن حاضر نیز به رشد و توسعه‌ی خود ادامه دهد و پیش بینی ها حاکی از آن است که در دهه های آتی با توسعه ای انقلابی مواجه شود. مؤسسه ی فن آوری ماساچوست داده کاوی[6] را یکی از ده فن‌آوری برتری می‌داند که نقش چشم گیری در تحول جهان خواهد داشت (راجرز[7] و جوینر، ۲۰۱۱).

در ایران سیستم‌های اطلاعاتی، به پزشکان در انجام وظایف خود در رابطه با تصمیم‌گیری (سیستم پشتیبان تصمیم‌گیری) کمک می‌کند. بخش قابل توجهی از تحقیقات حوزه انفورماتیک پزشکی به صورت مستقیم و غیرمستقیم به این مقوله اختصاص دارد (ادل[8] و همکاران، ۲۰۱۰). مزایای فناوری اطلاعات نه تنها در تصمیم‌گیری بلکه همچنین در پیاده‌سازی بهداشت الکترونیکی، مراقبت‌های پرستاری و غیره نیز دیده شود که می‌تواند خدمات بهداشت عمومی ایران را بهبود بخشد؛ پرایز[9] و همکاران (۲۰۱۳) اظهار داشتند که کارکنان مدیریت اطلاعات مراقبت‌های بهداشتی برای کنترل هزینه‌های مراقبت بهداشتی و مدیریت استفاده از خدمات، از تلاش‌های مختلف مانند پیاده‌سازی پرونده‌های مدیریتی، بررسی بهره-وری و مدیریت بیماری‌ها استفاده می‌کنند. با این حال، به نظر نمی‌رسد تمام این برنامه‌ها درکنترل هزینه‌ها درست عمل نمایند. آنها روش‌های مختلفی را برای شناسایی بیماران مبتلا به بیماری نادر (به دلیل ریسک بالا در پذیرش) و کنترل هزینه

---

[1] Giudici
[2] KDD (knowledge discovery in database)
[3] cleary
[4] Piatetsky-Shapiro
[5] Bali
[6] Massachusetts institute of technology
[7] Rogers
[8] Adel
[9] Price

ها به روش حرفه‌ای پزشکی و مدل‌های پیش‌بینی طراحی‌شده، پیشنهاد می‌کنند که از داده‌کاوی به عنوان یکی از انواع این روش‌ها یاد می‌کنند (پرایز و همکاران، ۲۰۱۳)؛ بنابراین، بنیاد بیماری‌های نادر ایران (رادو-آی-آر) باید از مدیریت داده خوبی بهره‌مند باشد. با این حال، بنیاد با برخی از چالش‌ها در فرایند مدیریت داده‌ها، به ویژه در بخش سلامت، جایی که حجم زیادی اطلاعات باید سازمان یافته و ذخیره شود، روبرو است. بنیاد بیماری های نادر ایران (رادو-آی-آر)، مرکز حمایتی، تشخیصی، درمانی و خدمات اجتماعی است و تنها حامی نهاد بیماران نادر در ایران می‌باشد. این نهاد کلیه خدمات را برای انواع بیماری‌های نادر مدیریت می‌کند. یک نمونه از آن، مدیریت بیماری‌های نادر که از مرگبارترین بیماری‌ها در سراسر جهان است می‌باشد (جولی[10] و همکاران، ۲۰۱۱). بنیاد بیماری های نادر ایران سالهاست که از سیستم های اطلاعاتی برای ارائه انواع خدمات به بیماران و پزشکان بهره می‌برد (سایت رادیور[11]، ۲۰۱۴).

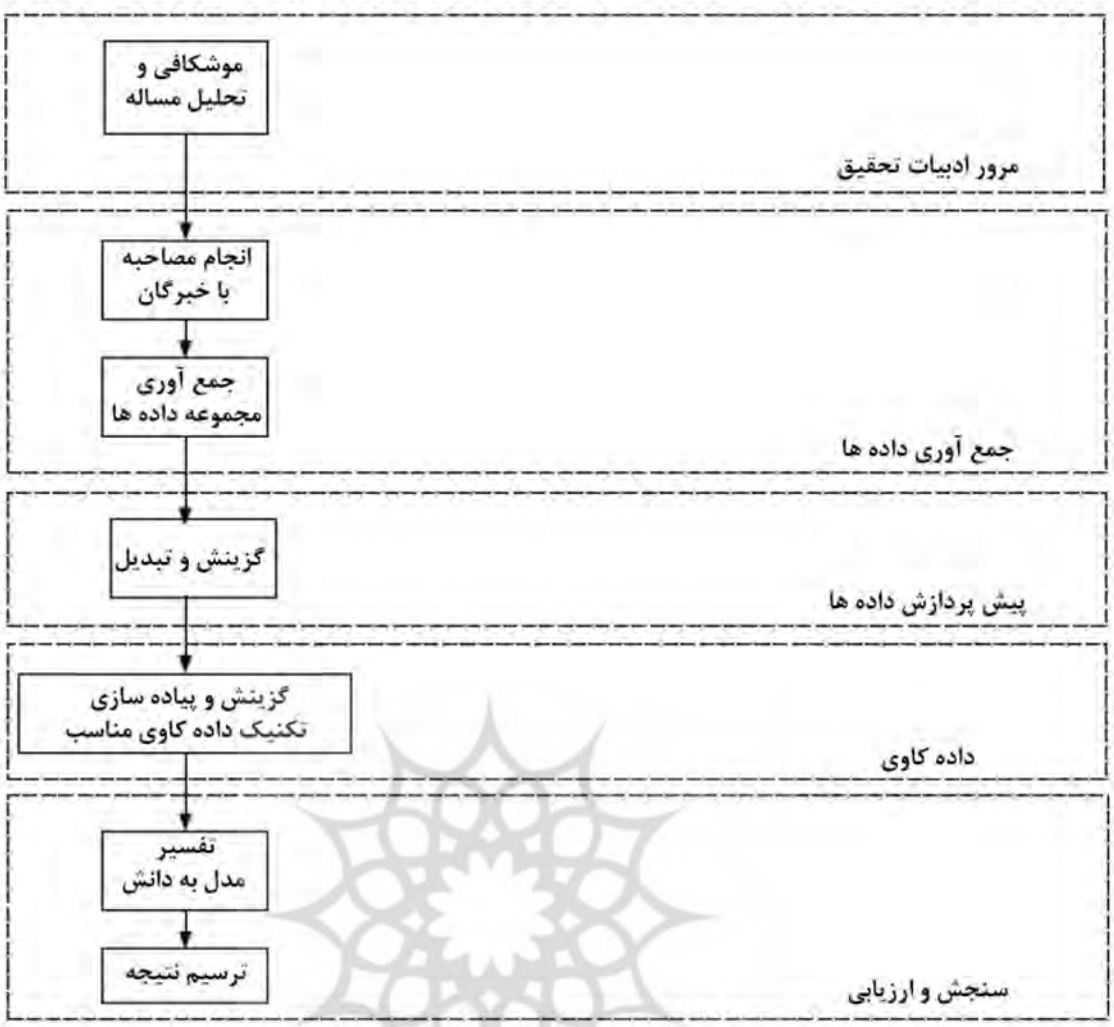
بنابراین، داده‌کاوی می‌تواند با ارائه تکنیک خود، یکی از راهکارها برای استخراج اطلاعات از مقادیر زیادی از داده‌ها باشد و قادر است کیفیت تصمیم‌گیری مدیریت داده‌ها را بهبود بخشد (اسلامی و همکاران، ۲۰۱۱). هدف اصلی داده‌کاوی پیش‌بینی است. درخت تصمیم‌گیری، به عنوان یکی از تکنیک‌های داده‌کاوی ثابت کرده که در میان دیگر تکنیک‌ها، از جمله شبکه‌های عصبی مصنوعی و مدل رگرسیون، دقیق‌ترین پیش‌بینی را ارائه می‌دهد (دلن[12] وهمکاران، ۲۰۰۵). داده کاوی پیشگویانه شایع‌ترین نوع داده‌کاوی و دارای بیشترین آمار برنامه‌های کاربردی تجاری است (کیودسی و همکاران، ۲۰۱۷).

۳. روش‌شناسی پژوهش

روش تحقیق این پژوهش داده‌کاوی می‌باشد. داده‌کاوی یک روش و تکنولوژی است که از سال ۱۹۹۴ معرفی شده است (ترایبولا[13]، ۱۹۹۷) روش و تکنولوژی داده‌کاوی حجم زیادی از داده‌ها را به اطلاعات معنی‌دار برای پشتیبانی کارآمد از تصمیم‌گیری تبدیل می‌کند (کوه[14] و همکاران، ۲۰۱۱). در این تحقیق، از الگوریتم طبقه‌بندی داده‌ها (الگوریتم C4.5 برای برای پیش‌بینی عوامل موثر بر بیماری نادر بر اساس چندین شاخص از قبیل سن، جنسیت، مدت زمان اقامت و بیماری استفاده شده است.

وکا (WEKA) فرآیند داده‌کاوی را با استفاده از درخت تصمیم‌گیری که برای کمک به تصمیم‌گیری سازمان براساس عوامل طبقه بندی شده است، مورد استفاده قرار می‌دهد (ریچا[15]، ۲۰۱۳). داده‌های جمع‌آوری شده از طریق برخی جلسات مصاحبه با آنالیزورهای اطلاعات و مدیران فناوری اطلاعات بنیاد بیماری‌های نادر ایران از طریق تعاملات چهره به چهره، تلفن یا ایمیل به دست آمده است. هدف اصلی این جلسات بحث و بررسی، دستیابی به بینش در مورد نوع داده‌ها برای مرور برخی از احتمالات در خصوص کارآیی داده‌کاوی است که می تواند برای رادو-آی-آر سودمند باشد. شکل ۱ طرح تحقیقاتی را نشان می دهد.

[10] Julie
[11] Radoir.org
[12] Delen
[13] Trybula
[14] Koh
[15] Richa



شکل ۱: طرح تحقیق

٤. یافته‌های پژوهش

۱) جمع‌آوری و درک اطلاعات

داده‌های جمع‌آوری شده به‌صورت وضعیت کلینیک درمانی بنیاد بیماری‌های نادر ایران از دوره ۱ فروردین ۱۳۹۸ تا ۱ خرداد ۱۳۹۸، به عنوان داده‌های خام برای این تحقیق مورد استفاده قرار گرفته است. مجموع ۲۳۵۲ گزارش اطلاعات مربوط به بیماران، شامل ۶۰۵ بیمار بستری و ۱۷۴۷ اطلاعات سرپایی است که می‌تواند مورد بررسی و تحلیل قرار گیرند.

از ۲۳۵۲ گزارش اطلاعات، ۹۷۶ بیماری نادر و ۱۳۷۶ عدد آن بیماری‌های غیر نادر است. این داده‌ها به عنوان ویژگی‌های ورودی برای پردازش و اجرای الگوریتم درخت تصمیم‌گیری C4.5 در طبقه‌بندی و پیش‌بینی مدل استفاده می‌شود. خروجی مدل پیش‌بینی "بیمار نادر" یا "غیر نادر" است.

در این تحقیق، داده‌های خام با چهارچوبی مجزا از علامت (CSV) ذخیره می‌شوند. سه فایل پایگاه داده بیماران بستری از مراکز وابسته به بنیاد، مرکز تصویربرداری پزشکی پارسیان، آزمایشگاه بیماران نادر، کلینیک فوق تخصصی بیماران نادر و یک فایل پایگاه داده بیماران سرپایی در یک جدول در مایکروسافت اکسل ادغام شده و به فرمت CSV تبدیل شده تا با ابزار داده‌کاوی وکا (WEKA) سازگار و به راحتی پردازش گردد.

پس از اتمام ادغام داده‌ها، پاک‌سازی داده‌ها انجام شد، داده‌های نامتناسب، داده‌های فاقد ارزش مقدار (فاقد پرونده) و داده‌های بیش از حد (داده‌های حاوی بیش از یک پرونده با همان مقادیر) حذف شده‌اند زیرا وجود آنها می‌تواند کیفیت یا دقت نتایج داده‌کاوی را کاهش دهد.

پس از ادغام و پاک‌سازی داده‌ها، یک جدول چهار ستونی که شامل (سن، جنسیت، مدت زمان اقامت و بیماری) است تعریف می شود. جزئیات این ویژگی‌ها در جدول ۱ دیده می‌شود.

جدول ۱ – توصیف ویژگی ها در مجموعه داده ها

ویژگی ها	توصیف	نوع	مقادیر ممکن
Age	سن بیمار	عددی	(۱ و ۲ و ۳ و...) سال
Gender	جنسیت بیمار	اسمی	مرد، زن
Length Of Stay	مدت زمان اقامت بیمار در مراکز درمانی	عددی	۰ روز، ۱ روز، ۲ روز و ...
Disease	آیا بیماری نادر است یا خیر؟	اسمی	نادر، غیر نادر

۲) تجزیه و تحلیل تجربی

۱) پیش پردازش داده‌ها

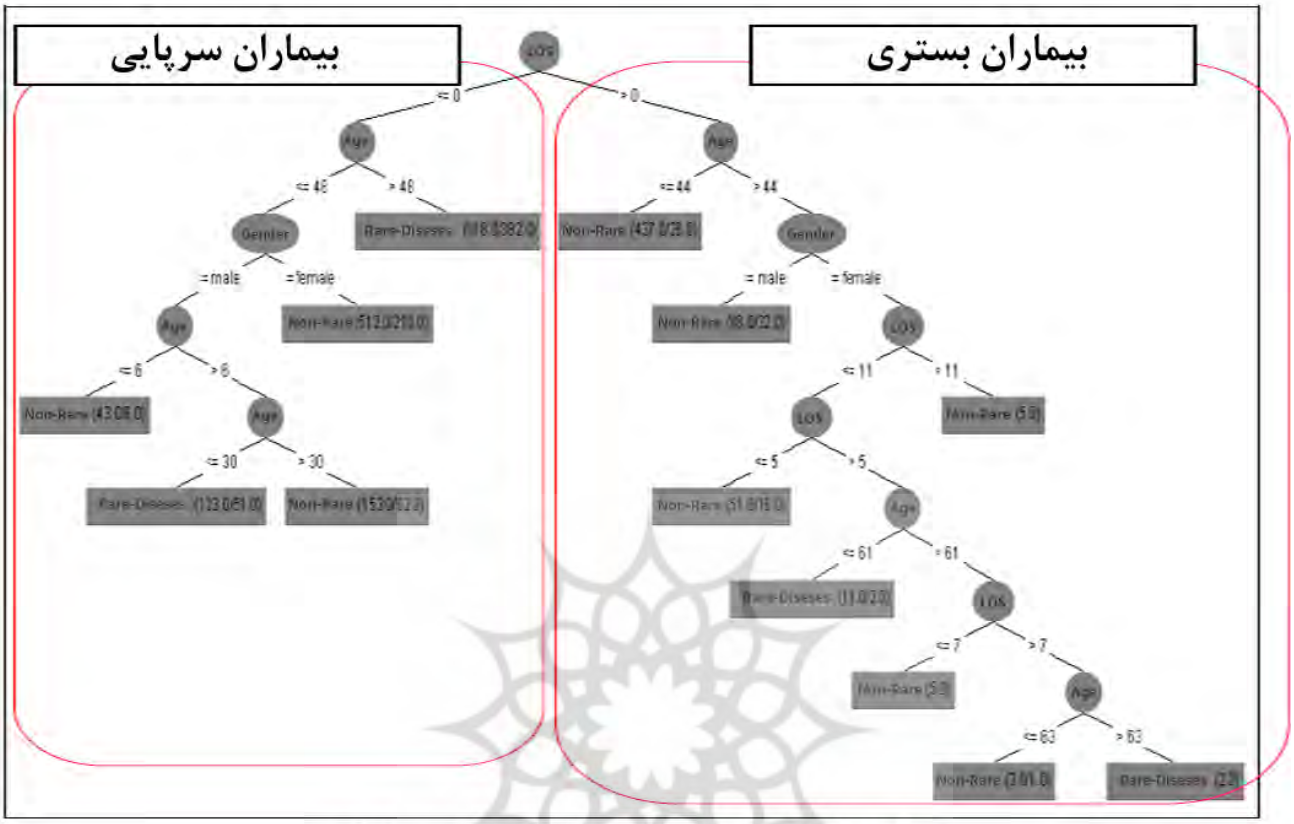
تجزیه و تحلیل تجربی بر روی داده‌های بیماران با استفاده از نرم‌افزار وکا برای به دست آوردن تصویری از درخت تصمیم-گیری و سنجش دقت طبقه‌بندی، ماتریس درهم ریختگی و نرخ TP (شخص بیمار، به درستی بیمار تشخیص داده شود) و نرخ TN (شخص سالم، به درستی سالم تشخیص داده شود) انجام شد. پس از بارگذاری فایل به پنجره اکسپلورر وکا، هیستوگرام تولید می شود.

این گراف نشان می‌دهد که تعداد بیماران زن بیشتر از مردان با تعداد ۱۳۴۸ نفر است. بیماران مردی که از بیماری‌های نادر رنج می‌برند تقریبا همانند کسانی می‌باشند که از بیماری‌های مزمن رنج نمی‌برند، درحالی‌که تعداد بیماران زن مبتلا به بیماری نادر کمتر از افرادی است که از آنها رنج نمی‌برند. همچنین یافته‌ها نشان می‌دهد که حداقل میزان اقامت بیمار در مراکز خدماتی ۱ روز و حداکثر ۳۷ روز است و تعداد بیماران مبتلا به بیماری نادر کمتر از کسانی است که دارای بیماری غیر نادر با تعداد ۹۷۶ بیمار نادر در مقایسه با سال ۱۳۷۶ بیمار غیر نادر می‌باشند.

پس از اتمام فرایند آماده سازی داده‌ها، مدل‌های طبقه‌بندی ساخته شدند. روش درخت تصمیم‌گیری ارائه شده توسط وکا که طبقه بندی الگوریتم C4.5 (J4.8) می‌باشد در مجموعه داده‌ها برای تولید درخت طبقه‌بندی انجام پذیرفته است. داده-های طبقه‌بندی شده توسط مجموعه داده‌های آموزشی، ارزیابی می‌شود.

درخت طبقه‌بندی تولید شده در شکل ۲ نشان داده شده است. برخی از گره‌ها در شکل ۲ با عنوان مدت زمان اقامت (LOS)، به عنوان مهم‌ترین عامل پیش‌بینی بیماری می‌باشد. گره مدت زمان اقامت (LOS) ریشه درخت تصمیم‌گیری و طبقه‌بندی می‌شود، زیرا از میان دیگر ویژگی‌ها بیشترین اطلاعات را به خود اختصاص داده است. از این رو، قوانین طبقه-

بندی شده از درخت تصمیم گیری در دو قسمت توضیح داده می‌شود، یکی بیماران سرپایی (برای بیمارانی که در یک روز اقامت دارند) و دیگری تحت عنوان بیماران بستری (برای بیماران که بیش از یک روز اقامت دارند) توضیح داده شده است.



شکل ۲. درخت تصمیم گیری که در آن مدت زمان اقامت (LOS) به عنوان مهمترین عامل است

۲) خلاصه ارزیابی

عملکرد الگوریتم با استفاده از مصورسازی درخت تصمیم‌گیری، تجزیه و تحلیل بیماران سرپایی و بستری، دقت طبقه بندی، ماتریس درهم ریختگی و نرخ TP و TN مورد بررسی قرار می‌گیرد.

از نتایج فوق، می‌توان این جمع‌بندی را نمود که طول مدت اقامت (LOS) مهم‌ترین عاملی است که بر وضعیت بیماری ها تأثیر می‌گذارد و بنیاد را قادر می‌سازد تا طول مدت اقامت بیمار را با توجه به عوامل مختلف مانند سن، جنس و نوع بیماری شناسایی نماید. جدول ۲ عوامل موثر بر بیماری نادر را نشان می‌دهد که می‌توان از آن نتیجه‌گیری کرد که اکثر بیماران مبتلا به بیماری‌های نادر سرپایی، دارای سن بیش از ۴۸ سال می باشند.

جدول ۲ – عوامل مؤثر بر بیماری های نادر

نوع	گروه سنی (سال)	جنسیت	مدت زمان اقامت (روز)	تعداد
بیمار سرپایی	۳۰-۷	مرد	۰	۷۲
بیمار سرپایی	بیش از ۴۸	هردو	۰	۵۳۴
بیمار بستری	۶۱-۴۵	زن	۱۱-۶	۱۲
بیمار بستری	بیش از ۶۳	زن	۱۱-۸	۲

میزان دقت طبقه‌بندی برای پیش‌بینی عملکرد، نسبتا بالاتر از میانگین و در حدود ۶۶٫۳۷٪ است. ۱۵۶۱ مورد از ۲۳۵۲ به درستی طبقه بندی شدند که در آن درصد بیماران غیرنادر (۶۸/۴ درصد) بیشتر از بیماران نادر (۶۳/۵ درصد) است. نتیجه تحقیق در مقایسه با نتایج تحقیقات قبلی تفاوت چندانی را در پی نداشت در حالی که نتایج دقت حاصل از انجام آن تحقیقات در حدود ۶۷٪ بوده است؛ به طور مثال هوانگ، بیماری نادر را از طریق آنومالی و بر اساس علائم حیاتی بیماران مانند ECG پیش بینی نموده است (لی و همکاران، ۲۰۰۸).

مقدار آماری کاپا (Kappa) برابر 0.3154 است که ارزیابی میزان سازش میان مقادیر پیش‌بینی‌شده توسط مدل می باشد. مقادیر خطای زیر (حاشیه خطای مطلوب، خطای جذر میانگین مربعات، خطای مطلق نسبی و خطای جذر میانگین نسبی مربعات) خطای پیش بینی را تخمین می‌زند. علاوه بر این، محدوده ROC برای هر دو دسته درخت تصمیم‌گیری برابر ۰٫۷۱۳ است که نشان می‌دهد اعتبار طبقه‌بندی بالا است. از آنجا که ROC نزدیکتر به ۱ است، توان تفکیک‌کننده طبقه‌بندی بیشتر است.

وکا خطاهای حاصل از طبقه‌بندی را مشخص می‌نماید. خطای طبقه‌بندی، تعداد عوامل مورد نیاز برای تجزیه و تحلیل را که باید به طورگسترده مورد توجه قرار گیرد نشان می‌دهد. به عنوان مثال، بیماران جوانتر مبتلا به بیماری‌های غیر نادر نیز وجود داشتند که به احتمال زیاد و به‌طور اشتباه به‌عنوان بیمار نادر طبقه‌بندی شده‌اند. شاید این مساله به دلیل فقدان اطلاعات بیماران نادر به ویژه برای بیماران جوانتر با توجه به نسبت بین بیماران جوان (کمتر از ۲۵ سال) مبتلا به بیماری های نادر و غیر نادر ۳۸۷:۱۴۰ رخ داده باشد. در نتیجه، الگوریتم برای پیش‌بینی بیماری‌های نادر برای بیماران جوانتر اطلاعات کافی را در اختیار ندارد.

داده‌کاوی به طور معمول یک فرآیند تکراری است که در آن چندین مرحله باید چندین بار تکرار شود (دینی، ۲۰۱۶). در این مورد مطالعاتی، با توجه به دقت که زیاد بالا نیست، پیش پردازش داده‌ها به منظور افزایش دقت مجدداً انجام می‌شود، مانند حذف خطا به صورت یک به یک (بر اساس خطای طبقه‌بندی)، حذف داده‌های متراکم، تعیین داده‌های خارج از محدوده و به حداقل رساندن داده‌های عددی با طبقه‌بندی سن به کودکان نوپا، خردسالی، اوایل نوجوانی، اواخر نوجوانی، اوایل جوانی، اواخر جوانی، اوایل سالمندی، اواخر سالمندی و کهنسالان (بر اساس طبقه بندی گروه سنی اداره بهداشت جمهوری اسلامی ایران)؛ اما حتی پس از انجام پردازش داده‌ها برای دومین بار، هنوز هم پارامتر دقت آنچنان که باید، افزایش نیافت.

جدول ۳ ماتریس درهم ریختگی را برای هر طبقه‌بندی‌کننده جهت تفسیر و درک نتایج نمایش می‌دهد.

جدول ۳ – ماتریس درهم ریختگی

		کلاس های واقعی	
		الف	ب
کلاس های پیش بینی شده	الف: بیمار غیر نادر	$\frac{TP}{941}$	$\frac{FP}{435}$
	ب: بیمار نادر	$\frac{FN}{356}$	$\frac{TN}{620}$
J4.8	جمع کل	۱۲۹۷	۱۰۵۵
	دقت	٪۷۲٫۵۵	٪۵۸٫۷۶

یک ماتریس درهم‌ریختگی، اطلاعاتی را به صورت عددی ارائه می‌دهد که برای توصیف عملکرد مدل طبقه بندی بسیار مناسب است و باعث می‌شود مقایسه عملکرد مدل‌های مختلف آسان‌تر شود. علاوه بر این، ماتریس درهم ریختگی نه تنها نشان می‌دهد که مدل پیش‌بینی خوبی دارد، بلکه جزئیاتی را در طول فرآیند استخراج اطلاعات که ممکن است اشتباه باشد نمایان می‌سازد (دیوید بو، ۲۰۱۲).

ماتریس درهم ریختگی در جدول ۳ نشان داده شده است که در آن ستون‌ها طبقه‌های واقعی و سطرها طبقه‌های پیش‌بینی-شده را نشان می‌دهد. در جدول ۳ اعداد ۹۴۱ و ۶۲۰ شاخص تعداد مواردی است که مقادیر واقعی و پیش‌بینی‌شده مشابه هستند. ۹۴۱ شاخص بیمارانی است که با بیماری غیرنادر پیش‌بینی می‌شوند و واقعا بیماران غیرنادر هستند، در حالی که ۶۲۰ بیمارانی هستند که با بیماری‌های نادر پیش‌بینی می‌شوند و واقعا بیماران نادر هستند؛ به عبارت دیگر، قطر ماتریس همه پیش‌بینی‌های درست را نشان می‌دهد.

از طرفی تعداد ۴۳۵ شاخص مواردی است که نتیجه واقعی بیماری، غیر نادر بوده اما به عنوان یک بیماری نادر پیش‌بینی‌شده و تعداد ۳۵۶ شاخص مواردی که در آن نتیجه یک بیماری نادر بوده اما به عنوان بیماری غیرنادر پیش‌بینی‌شده است. لذا می‌توان چنین تجزیه و تحلیل کرد که درصد دقت در رده‌بندی بیماری غیر نادر (۷۲٫۵۵٪) نسبت به بیماری نادر (۵۸٫۷۶٪) بالاتر است.

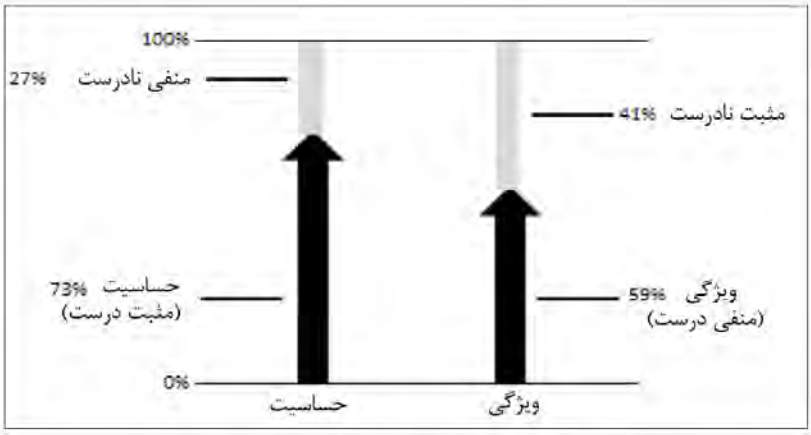
ضریب اطمینان آزمون پیش‌بینی را می‌توان با محاسبه حساسیت و ویژگی اندازه‌گیری نمود (پاتیل، ۲۰۰۹). در جدول ۳، مقدار TP (مثبت درست) ۹۴۱ و مقدار FP (مثبت نادرست)، ۴۳۵ تعیین شده است. علاوه بر این، ۳۵۶ مورد از FN (منفی نادرست) و ۶۲۰ مورد از TN (منفی نادرست) یافت شده است. سلول‌هایی که با TP (مثبت درست) برچسب‌گذاری شده اند، تعداد موارد واقعی است که توسط الگوریتم طبقه‌بندی به‌طور دقیق پیش‌بینی شده است، در حالی که سلول‌های دیگر، FP، FN و TN به روش‌های مشابه تفسیر می شوند. عملکرد الگوریتم را نه تنها با محاسبه دقت اجرا، بلکه با حساسیت و ویژگی نیز می توان مورد ارزیابی قرار داد.

- حساسیت به توانایی آزمون پیش بینی در شناسایی درست بیماران مبتلا به بیماری غیر نادر را مشخص می‌کند. حساسیت برابر با نرخ مثبت درست است:

حساسیت= TP / TP + FN => 941/(941+356) =0.73

- ویژگی به توانایی آزمون پیش‌بینی در شناسایی بیماران مبتلا به بیماری نادر اشاره می‌کند. ویژگی برابر با نرخ منفی درست است:

ویژگی= TN / TN + FP => 620/(620+435) =0.59



شکل ۳: نسبت حساسیت به ویژگی

شکل ۳ حساسیت بالا اما ویژگی کم (مثبت نادرست بیشتر) را نشان می‌دهد که بیان‌کننده تعداد مواردی است که نتیجه واقعی آن غیر نادر بوده اما به عنوان بیماری نادر پیش‌بینی شده است. این امر ممکن است ناشی از داده های نامتقارن باشد که در آن تعدادی از داده های غیر نادر بسیار بالاتر از داده های نادر است. مسئله ناپایداری داده ها گسترش یافته و به طور جهانی رخ داده است که بر نتیجه فرایند داده‌کاوی تأثیر می‌گذارد (چاولا، ۲۰۰۵). به عنوان مثال، در این تحقیق تعداد بیماران سرپایی بسیار بزرگتر از داده های بیماران بستری با نسبت ۱۷۴۷: ۶۰۵ می باشد و تعداد بیماران مبتلا به بیماری های نادر بسیار کمتر از بیماری های غیر نادر با نسبت ۹۷۶: ۱۳۷۶ است. با توجه به نتیجه بالا، نرخ TP به حساسیت و از آنجا که نرخ TP به حساسیت برابر ۰,۷۳ است لذا پیش بینی تمامی موارد توسط مدل درست است.

۵. بحث و نتیجه‌گیری

هدف اولیه این تحقیق، پیش‌بینی عوامل جمعیت شناختی بر بیماری نادر است که بر اساس نادر یا غیر نادر بودن طبقه‌بندی می‌شود؛ اما پس از انجام فرایند داده‌کاوی، طول مدت اقامت (LOS)، به عنوان دانش حاصل از مدل پیش‌بینی، به عنوان مهمترین عامل که یک بیماری نادر را پیش‌بینی می‌کند، ظاهر می‌شود؛ به عبارت دیگر، طول مدت اقامت (LOS) قوی‌ترین شاخص پیش‌بینی بیماری‌های نادر است که این بنیاد را قادر می‌سازد تا طول مدت اقامت (LOS) بیماران را با توجه به وضعیت بیماری، جنسیت و سن را پیش‌بینی نماید؛ بنابراین، ثابت می‌شود که داده‌کاوی بیش از یک موضوع آماری می‌باشد و به عنوان یک ابزار تجزیه و تحلیل پیش بینی مطرح می‌باشد.

علاوه بر این، الگوریتم C4.5 یک مدل طبقه‌بندی را در قالب مصورسازی (درخت تصمیم گیری) با موفقیت ایجاد می‌کند که می‌تواند به راحتی برای پیش‌بینی بیماری نادر مورد مطالعه و تفسیر قرار گیرد. دقت پیش‌بینی آن ۶۶,۳۷ درصد بوده است. با این حال، اندازه حساسیت تا ۰,۷۳ به دست آمد که این بدان معناست که ضریب اطمینان آزمون پیش‌بینی، خوب است و معتبر بودن داده‌ها ۱۰۰ درصد ثابت شده‌اند؛ بنابراین، علیرغم دقت پایین مدل مشتق شده، نتیجه برخی از عواملی را که بنیاد بیماری‌های نادر می‌تواند در هنگام پیش بینی هزینه‌های درمان بیمار مورد توجه قرار دهد، نشان می‌دهد. به عنوان مثال، مدل نشان می‌دهد که شاخص جنس یا سن، به شدت تحت تاثیر مدت زمان بستری بیمار قرار می‌گیرد. بیمار سالخورده با بیماری نادر باید در مراکز درمانی تحت مراقبت و مدت زمان بستری طولانی‌تری نسبت به افراد جوان‌تر قرار گیرد. در نتیجه، درخت تصمیم‌گیری یک روش مفید و آموزنده برای انجام داده‌کاوی پیش‌بینی شده است.

مزیت داده کاوی برای مراکز درمانی بیماران نادر را می توان اینگونه بیان کرد؛ دانشی که می‌تواند از الگوهای ایجاد شده پس از داده‌کاوی شناسایی شود، احتمالا می‌تواند به عنوان یک سیاست جدید برای بنیاد بیماری‌های نادر ایران (رادو-آی-آر) مورد توجه قرار گیرد. مصورسازی از طریق درخت تصمیم‌گیری که خواندن و تفسیر را آسان می نماید نیز به افراد شاغل در بنیاد در ترسیم اطلاعات، کمک شایانی را می‌نمایند. برخی پیشنهاداتی که می‌تواند در مراکز درمانی بیماران نادر به کارگرفته شود به شرح زیر می‌باشد؛ اینکه با این روش ممکن است بنیاد بیماری‌های نادر ایران قادر به شناسایی مدت زمان بستری شدن بیماران مبتلا به بیماری نادر شود که این موضوع می‌تواند مخاطب را در اختصاص میزان بودجه بیشتر برای خدمات بهداشتی و داروها به بیماران نادر بستری شده در مقابل بیماران سرپایی مبتلا به بیماری‌های نادر، جهت برآورد نمودن میزان پوشش بیمه‌ای که باید برای یک بیمار بستری شده در نظر گرفته شود و نیز برآورد نمودن میزان پرداختی هزینه حق بیمه پذیرش‌شدگان در هر ماه، کمک و راهنمایی نماید. همچنین مشاوره‌های بهداشتی می‌بایست صورت گیرند، مخصوصا برای زنانی که مستعد ابتلا به بیماری‌های نادر در مقایسه با همتایان مرد خود می‌باشند، دانش، توانایی، آگاهی از بیماری‌های نادر و درک پذیرش‌شدگان می‌بایست افزایش یابند و حداقل یک بار در سال برای پذیرش شدگان بالای ۴۰ سال، غربالگری پزشکی انجام شود. هدف این مطالعه شناسایی عوامل مخاطره‌آمیز برای بیماری‌های نادر است.

قبل از انجام داده‌کاوی، اطمینان حاصل نمودن از اینکه نسبت نمونه محاسبه شده متعادل باشد حائز اهمیت است. نیازی نیست این نسبت یکی باشد اما برای سیستم محاسبه‌گر مناسب است تا چگونگی طبقه‌بندی را برای هر دسته ایجاد شده فرابگیرد. داده های مورد استفاده در این مطالعه عمدتا توسط یک طبقه غالب از نمونه ها انتخاب شده اند. در نتیجه، داده های نامتعادل منجر به کمبود دقت و طبقه بندی نادرست شده است و در نتیجه تعداد بیماری های غیرنادر بیش از بیماری نادر شده است.

بنیاد می تواند از روش های مختلفی برای افزایش دقت استفاده کند. به عنوان مثال، یکی از روش ها جنگل تصادفی یا ماشین بردار پشتیبانی (SVM) است که یکی از گزینه های موجود برای افزایش دقت می باشد (ادکی، ۲۰۱۲). با این حال، روش‌های دیگر ممکن است دقت بیشتری داشته باشند، اما تصویری مانند درخت تصمیم گیری را ارائه نمی دهد. به ویژه برای مراکزی که افراد با داده کاوی آشنایی ندارند، استخراج اطلاعات از درخت تصمیم مصور شده، ساده تر است.

بر اساس این مطالعه، مشخص شد که داده های دریافتی از مراکز درمانی بنیاد، به لحاظ کیفی مورد انتظار نیست. در نتیجه، یک سری تکنیک های پاکسازی داده اتخاذ گردید. این موضوع نشان داد که عملکرد روش های داده کاوی بستگی به کیفیت داده ها دارد که بنیاد باید این مساله را در نظر بگیرد. داده های مورد استفاده در این تحقیق حاوی بسیاری از مقادیر گم شده، داده های نامناسب یا اطلاعات ناقص است که بر کیفیت داده هایی که باید قبل از انجام فرایند داده کاوی پاک شوند (پاک سازی داده ها) تاثیر می گذارد؛ بنابراین، برای اطمینان از اینکه داده های نگهداری شده پاک شده اند، لازم است برخی از مکانیزم ها مورد استفاده قرار گیرند. لذا داده کاوی می تواند به راحتی داده ها را بدون نیاز به مداخله نوع خاصی از توجیه یا اعمال نظر شخصی که می تواند مانع ادامه فعالیت شود، درصورتی که بنیاد پیش بینی نماید که داده های پاکسازی شده را از ابتدا در اختیار داشته باشد، پردازش نماید.

داشتن یک مدیریت داده خوب، یکی از روش های حفظ کیفیت داده است. کیفیت داده ها می تواند زمانی که رادو-آی-آر میل به تامین اعتبار برای روند تجزیه و تحلیل جهت درک شکاف میان سطح بلوغ مدیریت داده ها و بهبودی الگوها و

روش های مدیریت داده ها دارد، بکارگفته شود. سطح بلوغ مدیریت داده ها می تواند برای کمک به ارزیابی وضعیت فعلی مدیریت داده ها بنیاد استفاده شود. بر اساس یافته های این تحقیق، تاکنون بنیاد بیماری های نادر ایران هیچ فرایندی را در مراکز درمانی خود برای دانستن اینکه آیا داده ها پاک سازی شده اند یا خیر، در اختیار ندارد؛ بنابراین، بر اساس پنج سطح بلوغ کیفیت داده شده توسط روش لری انگلیسی (ادلمن، ۲۰۰۵)، کیفیت داده ها، بلوغ را در سطح دوم نشان نمی دهد که در آن بنیاد نه تنها نیاز به انجام نمایه سازی داده و پاکسازی داده ها دارد بلکه باید پشتیبانی گسترده ای از بنیاد برای بهبود کیفی داده ها صورت پذیرد

٦. منابع و مآخذ

1. Adell, A., Ahmadi, P., & Sebt, M. (2010). Desining Model for Choosing human resources with data mining approach. Journal of Iranian Technology, 2(4), 5.
2. Bali, R. K. (Ed.). (2005). Clinical knowledge management: opportunities and challenges. IGI Global.
3. Bowes, D., Hall, T., & Gray, D. (2012, September). Comparing the performance of fault prediction models which report multiple performance measures: recomputing the confusion matrix. In Proceedings of the 8th international conference on predictive models in software engineering (pp. 109-118).
4. Chawla, N. V. (2009). Data mining for imbalanced datasets: An overview. In Data mining and knowledge discovery handbook (pp. 875-886). Springer, Boston, MA.
5. Cleary, R. J. (2006). Applied data mining: statistical methods for business and industry.
6. Delen, D., Walker, G., & Kadam, A. (2005). Predicting breast cancer survivability: a comparison of three data mining methods. Artificial intelligence in medicine, 34(2), 113-127.
7. Edeki, C., & Pandya, S. (2012). Comparative study of data mining and statistical learning techniques for prediction of cancer survivability. Mediterranean journal of Social Sciences, 3(14), 49-49.
8. Giudici, P. (2005). Applied data mining: statistical methods for business and industry. John Wiley & Sons.
9. Helmerhorst, H. J., Schultz, M. J., van der Voort, P. H., Bosman, R. J., Juffermans, N. P., de Wilde, R. B., ... & de Keizer, N. F. (2016). Effectiveness and clinical outcomes of a two-step implementation of conservative oxygenation targets in critically ill patients: a before and after trial. Critical care medicine, 44(3), 554-563.
10. Julie, D. M., & Kannan, B. (2012). Statistical Machine Learning Techniques for the Prediction of Learning Disabilities in School-Age Children (Doctoral dissertation, Cochin University of Science and Technology).
11. Koh, H. C., & Tan, G. (2011). Data mining applications in healthcare. Journal of healthcare information management, 19(2), 65.
12. Li, J., Huang, K. Y., Jin, J., & Shi, J. (2008). A survey on statistical methods for health care fraud detection. Health care management science, 11(3), 275-287.
13. Moss, L. T., Abai, M., & Adelman, S. (2005). How to improve data quality. Data Strategy; Addison-Wesley Professional: Boston, MA, USA.
14. Patil, B. M., Toshniwal, D., & Joshi, R. C. (2009, March). Predicting burn patient survivability using decision tree in weka environment. In 2009 IEEE International Advance Computing Conference (pp. 1353-1356). IEEE.
15. Piatetsky-Shapiro, G. (2000). Knowledge discovery in databases: 10 years after. Acm Sigkdd Explorations Newsletter, 1(2), 59-61.
16. Price, S., Dobbs, D., Oliveira, J., Beidas, S., Burkhart, T., & Sharp, J. (2013). Clinical & Business Intelligence: Data Management–A Foundation for Analytics.
17. Qudsi, D. H. (2016). Predictive Analytics Data Mining in Imbalanced Medical Dataset. Jurnal Komputer Terapan, 2(2), 195-204.
18. Qudsi, D. H., Kartiwi, M., & Saleh, N. B. (2017). Predictive data mining of chronic diseases using decision tree: a case study of health insurance company in Indonesia. International Journal of Applied Engineering Research, 12(7), 1334-1339.

19. Rogers, G., & Joyner, E. (1997, March). Mining your data for health care quality improvement. In SAS User Group International Conference (pp. 641-647).
20. Sharma, R., Ghosh, A., & Joshi, P. K. (2013). Decision tree approach for classification of remotely sensed satellite data using open source support. Journal of Earth System Science, 122(5), 1237-1247.
21. Trybula, W. J. (1997). Data mining and knowledge discovery. Annual review of information science and technology (ARIST), 32, 197-229.
22. Yousofi, M. H., Esmaeili, M., & Sharifian, M. S. (2016). A study on image mining; its importance and challenges. American Journal of Software Engineering and Applications, 5(3-1), 5-9.

Decision Tree-Based Use of Data Mining Techniques to Determine & Predict the Rare Patients Status
The Case Study: Rare Patients Registry System (SABNA)

Mohsen Sojoudi [1]
Fariba Ebrahimbabaie [*2]

Date of Receipt: 2020/07/15 Date of Issue: 2020/07/18

Abstract

The health database contains a broad range of clinical data through which the discovered communication patterns and algorithms are led to new medical data achievements. Nowadays due to the existing integrated data as well as the informatics technology growth, it has turned into be a significant emergence. This study aims to identify the potential advantages with which data mining techniques can be applied for health and treatment trends using RADOIR's patients' data as a case study. The most usual method is WEKA based data pre-processing, data mining and classification with decision tree to create prediction modalities via picturing a tree in order to analyze a rare disease. Result & Also we have used WEKA to evaluate the prediction processing via measuring and analyzing the sensitivity rate. As a result, there are some factors which a patient supporting center can take them into consideration while predicting the patients' treatment costs and expenses.

Keyword

Data mining, Prediction Model, Rare Disease, Decision Tree, Sensitivity Analysis, WEKA

1. Operation Management: Research (PhD) Ferdowsi University of Mashad – Iran (mohsened@gmail.com)
2. Phd student of psychology, Lorestan University, Iran (faribaebrahimbabaie@yahoo.com)