پیش بینی میزان اثربخشی داروهای پره اکلامپسی با داده کاوی بر اساس دوز و روش مصرف دارو

نجمه جوریان[۱]، مریم عاشوری[۲*]

۱. کارشناس ارشد مدیریت اجرایی، بیمارستان حکیم، دانشکده علوم پزشکی و خدمات درمانی نیشابور، نیشابور، ایران.
۲. کارشناس ارشد مهندسی فناوری اطلاعات، دانشکده مهندسی صنایع، دانشگاه خواجه نصیرالدین طوسی، تهران، ایران.

تاریخ دریافت: ۱۳۹۳/۴/۲۵ **تاریخ پذیرش:** ۱۳۹۳/۹/۱

خلاصه

مقدمه: مسمومیت بارداری یا پره اکلامپسی، یکی از اختلالات خطرناک دوران بارداری است که حتی می‌تواند به خاتمه بارداری و سقط جنین ختم شود. مطالعه حاضر با هدف تولید یک مدل پیش‌بینی کننده کارا به منظور پیش بینی دقیق میزان اثربخشی داروهای پره اکلامپسی انجام شد.

روش کار: این مطالعه توصیفی- مقطعی در سال ۹۲-۱۳۹۰ بر روی ۷۷۴ بیمار در بیمارستان حکیم نیشابور انجام شد. سه الگوریتم متفاوت از الگوریتم های استنتاج قانون به نام‌های C5.0، C&R Tree و CHAID، روی داده های عارضه پره اکلامپسی اعمال شد و سپس صحت مدل های تولید شده به دست آمد. در نهایت برای تأیید صحت نتایج حاصل از پیش‌بینی از خوشه بندی استفاده شد. در این مطالعه جهت تحلیل داده ها و اجرای الگوریتم های داده کاوی از نرم افزار کلمنتاین (نسخه ۱۲/۰) استفاده شد.

یافته‌ها: مقادیر به دست آمده برای صحت مدل های ایجاد شده از اجرای الگوریتم‌های C5.0، C&R Tree و CHAID روی مجموعه داده های آموزش و آزمایش برای هر سه الگوریتم ۹۹/۶۳ و ۹۹/۱۴ درصد بود. صحت بالای مدل ها عملکرد خوب این الگوریتم ها را نشان داد. حساسیت ۱۰۰ درصد و دقت ۹۹/۵، برتری الگوریتم C&R Tree نسبت به سایر الگوریتم ها را تأیید کرد. نتایج حاصل از خوشه‌بندی روی درخت بهتر نشان داد که در مواقعی که میزان اثربخشی دارو افزایش یافت، میانگین سنی بیمار هم بالاتر بود و دلیل آن، رابطه مستقیم ایجاد عارضه پره اکلامپسی با افزایش سن بود.

نتیجه‌گیری: شواهد به دست آمده نشان می دهند که با افزایش سن احتمال وجود این عارضه افزایش و از داروهایی با میزان اثربخشی بالاتر برای درمان باید استفاده نمود. بنابراین در انتخاب داروی مصرفی با میزان اثربخشی متفاوت دقت بیشتری باید صورت گیرد. پیش بینی میزان اثربخشی دارو با هدف دستیابی به مواردی چون: کمک به پزشک برای افزایش صحت تشخیص و جلوگیری از تشخیص نادرست در دوز مصرفی دارو برای بیمار، تشخیص شدت پره اکلامپسی و مهار عوارض خطرناک مصرف بیش از حد دارو توسط بیمار، پیش بینی ذخیره داروی کافی و جلوگیری از عوارض جانبی ناشی از کمبود دارو صورت می‌گیرد.

کلمات کلیدی: پره اکلامپسی، خوشه بندی، درخت تصمیم، دسته بندی

* **نویسنده مسئول مکاتبات:** مریم عاشوری؛ دانشکده مهندسی صنایع، دانشگاه خواجه نصیرالدین طوسی، تهران، ایران. **تلفن:** ۰۵۱-۳۷۲۵۰۶۵۲؛ **پست الکترونیک:** maryam.ashoori@gmail.com

مقدمه

امروزه فناوری اطلاعات و کامپیوتر از طریق توموگرافی به کمک کامپیوتر تا سیستم های پایگاه داده در حوزه پزشکی و بهداشت و درمان نفوذ کرده است. پیچیدگی اطلاعات پزشکی و فناوری کامپیوتر، یک حوزه باز برای اکتشافات آینده باقی گذاشته است (۱). از طرفی حجم داده های پزشکی که به صورت الکترونیکی ذخیره می شوند، روز به روز در حال افزایش است. اما مجموعه عظیم داده های خام، به خودی خود کاربردی ندارند. برای معنی بخشیدن به این داده ها باید آن ها را تحلیل و تبدیل به اطلاعات و یا بهتر از آن، یعنی تبدیل به دانش کرد (۲). در دسترس بودن حجم بزرگی از داده ها اگر چه فرصت و مجال است، اما دانشی که در این پایگاه ها وجود دارد، معمولاً نهان و استفاده نشده است، از این رو ابزارهای داده کاوی می‌توانند در به دست آوردن این اطلاعات نهان به کار گرفته شوند (۳).

داده کاوی، تشخیص الگو را به صورت خود به خود انجام داده و تلاش می‌کند الگوهایی را در داده ها کشف کند که با روش های آماری مرسوم به سختی قابل کشف هستند (۴). استخراج دانش از میان حجم انبوه داده های مرتبط با سوابق بیماری و پرونده‌های پزشکی افراد و با استفاده از فرآیند داده کاوی می‌تواند منجر به شناسایی قوانین حاکم بر ایجاد، رشد و تسریع بیماری ها شده و اطلاعات ارزشمندی را به منظور شناسایی علل رخداد بیماری ها، پیش بینی کند و درمان بیماری ها را با توجه به عوامل محیطی حاکم، در اختیار متخصصین و دست اندرکاران حوزه سلامت قرار دهد. بنابراین پیچیدگی اطلاعات پزشکی و وجود ابزارهای داده کاوی باعث می شود که داده کاوی بر روی داده های پزشکی و سلامت از اهمیت بالایی برخوردار شود (۲).

پره اکلامپسی، عارضه اصلی تأثیرگذار در حداقل ۳ یا ۴ درصد تمام بارداری ها است و در سطح جهانی باعث مرگ و میر ۵۰۰۰۰ نفر از مادران باردار می شود (۵). پره اکلامپسی نه تنها یک بیماری دوران بارداری است، بلکه می تواند یک بیماری با پیامد های مادام العمر برای مادر و نوزاد در نظر گرفته می‌شود (۶) که در صورت ابتلاء مادر به پره اکلامپسی، جنین نیز در معرض خطرهایی نظیر تأخیر رشد داخل رحمی، نارس بودن، مرگ نوزادی و جنینی قرار می‌گیرد (۷). پره اکلامپسی به افزایش فشار خون پس از هفته ۲۰ بارداری همراه با پروتئینوری با ترشح بیش از ۳۰۰ میلی گرم پروتئین در ادرار ۲۴ ساعته و یا نسبت ۰/۳> (کراتینین/ پروتئین) و یا وجود ۳۰ میلی گرم در دسی لیتر پروتئین ثابت در نمونه های ادراری ساده گفته می‌شود (۸). در صورت عدم تشخیص به موقع پره اکلامپسی، تشنج عمومی بروز می‌کند که به آن، اکلامپسی گفته می‌شود (۷). در حال حاضر آزمونی به حد کافی قابل اعتماد برای تشخیص زود هنگام پره اکلامپسی که اجازه مداخله به موقع را داشته باشد وجود ندارد (۹).

عوامل خطر پره اکلامپسی شامل: چاقی، سابقه پره اکلامپسی (به خصوص در سنین پایین بارداری و سندرم HELLP)، نولی پار بودن، سن بالا، دیابت قندی، سابقه خانوادگی پره اکلامپسی، بارداری چندقلو، بیماری قلبی و عروقی، فشار خون مزمن، بیماری مزمن کلیوی، طولانی بودن پیشرفت زایمان می باشد (۵). با توجه به اهمیت فوق العاده پره اکلامپسی و عوارض ناشی از آن در مادر و نوزاد، کنترل دقیق فشار خون طی بارداری اهمیت بسزایی دارد. برای کنترل فشار خون در این بیماران لازم است از داروهایی نظیر هیدرالازین، لابتالول، نیفیدپین و غیره استفاده شود (۱۰). هیدرالازین در دوزهای ۱۰-۵ میلی گرمی با فواصل ۲۰-۱۵ دقیقه‌ای تجویز می‌شود تا اینکه پاسخ رضایت بخش حاصل شود. پاسخ رضایت بخش قبل از زایمان یا هنگام زایمان، به صورت کاهش فشار دیاستولی تا حد ۱۰۰-۹۰ میلی متر جیوه تعریف می شود. اما فشار خون دیاستولی نباید از این حد کمتر شود تا خون رسانی جفت مختل نشود (۱۱).

در این مقاله با پیاده سازی الگوریتم های دسته بندی C5.0، C&R Tree و CHAID از مجموعه رویکردهای پیش بینی در داده کاوی به پیش بینی میزان اثربخشی داروهای مورد استفاده در پره اکلامپسی پرداخته شد و سپس الگوریتم کارآتر برای تولید مدل پیش بینی کننده انتخاب شد. در ادامه علل کاهش

صحت و کارآیی مدل‌های پیش بینی کننده با توجه به ضریب اطمینان به دست آمده برای قوانین تولید شده، ریشه یابی شد. با کشف کارآترین الگوریتم ها برای تشخیص و تلاش برای افزایش کارآیی و میزان درستی آن ها با توجه به عوامل دخیل از جمله انتخاب زیر مجموعه صفات مناسب و نوع داده مورد استفاده، می‌توان به سمت ایجاد سیستم های مکانیزه با قابلیت اعتماد بالا با توانایی کشف الگوهای پیچیده و پیش بینی روندهای آینده برای مواجه با انواع بیماری ها گام برداشت (۲). سپس بر روی نتایج حاصل از دسته بندی، رویکرد خوشه بندی به منظور بررسی و تأیید نتایج حاصل از پیش بینی اجرا شد.

دو مسئله در خصوص پیش بینی میزان تأثیر داروهای مصرفی در پره اکلامپسی مطرح است. اول اینکه تجویز بیش از حد لازم و کمتر از میزان مورد نیاز دارو می تواند مشکلات بسیاری برای بیمار به وجود آورد. با توجه به اینکه پره‌اکلامپسی در دوران بارداری اتفاق می‌افتد، تجویز بیش از حد دارو می‌تواند اثرات نامطلوبی بر روی جنین بگذارد. تجویز کمتر از میزان مورد نیاز نیز باعث ایجاد شرایط روانی نامطلوب برای مادر تا رسیدن به پاسخ رضایت بخش می شود که این شرایط نیز بر روی جنین تأثیرگذار است. دوم اینکه با در نظر گرفتن این مطلب که پره اکلامپسی، یک عارضه خطرناک است، جهت تأمین و توزیع داروی مورد نیاز می‌بایست برنامه‌ریزی مناسب صورت گرفته و ذخیره مناسبی در این خصوص از طرف سازمان های مسئول تأمین دارو در نظر گرفته شود. بنابراین با در اختیار داشتن پایگاه داده های مربوط به مشخصات بیماران می‌توان میزان داروی پره اکلامپسی مورد نیاز در یک منطقه را پیش بینی کرد. با این روش، احتمال مواجه بیماران با کمبود دارو و مشکل در توزیع کاهش می یابد که این دو مسئله ضرورت و اهمیت مطالعه حاضر را مشخص می‌کند. در عین حال با در نظر گرفتن عوارض مصرف کم و یا بیش از حد دارو برای بیماران، ضرورت ایجاد یک سیستم تصمیم یار پزشک به خوبی احساس می‌شود که می‌توان با استفاده از بانک های اطلاعات این بیماران، این سیستم را ایجاد نمود.

بنابراین مطالعه حاضر با هدف کلی کشف دانش نهفته در داده های موجود در پایگاه داده های بیماری پره اکلامپسی انجام شد. دانش حاضر می‌تواند علاوه بر یاری رساندن به مسئولین برای پیش بینی میزان تأثیر داروی مصرفی، در ارائه یک دید مناسب از شرایط مصرف دارو به مسئولین ذی ربط کمک شایانی کند.

روش کار

مجموعه داده های پره اکلامپسی متعلق به بیمارستان حکیم نیشابور بود. این مجموعه داده شامل ۷۷۴ نمونه با ۷ صفت بود که از خرداد ماه سال ۱۳۹۰ تا تیر ماه سال ۱۳۹۲ جمع آوری شد. این یافته ها با مراجعه مستقیم پژوهشگر به بیمارستان حکیم نیشابور و استخراج داده‌ها به صورت خروجی اکسل از سیستم نرم افزاری صورت گرفت و محتوای داده ها مورد تأیید مسئولین بیمارستان می‌باشد. روایی روش جمع آوری اطلاعات توسط اساتید امر مورد تأیید قرار گرفت. صفات مربوط به مجموعه داده مورد بررسی سن، نیتروگلیسرین، هیدرالازین ۵۰، هیدرالازین ۲۵، هیدرالازین ۱۰ و هیدرالازین ۲۰ درصد بودند که نوع داده‌ی این صفات عدد بود. داروهای هیدرالازین ۵۰، هیدرالازین ۲۵ و هیدرالازین ۱۰ به صورت قرص، هیدرالازین ۲۰ درصد به صورت آمپول و نیتروگلیسرین به صورت سرم استفاده می شوند.

در این مطالعه جهت تحلیل داده ها و اجرای الگوریتم های داده کاوی از نرم افزار کلمنتاین (نسخه ۱۲/۰) استفاده شد. پنج الگوریتم استنتاج قانون مختلف در کلمنتاین (نسخه ۱۲/۰) شامل C5.0، CHAID، QUEST، C&R Tree و لیست تصمیم است (۱۲، ۱۳). تمامی الگوریتم های نامبرده درخت تصمیم تولید نمودند که به ایجاد مجموعه ای از قوانین برای توصیف بخش های مجزایی از داده های مرتبط با فیلد خروجی کمک نمود (۱۲). درخت تصمیم برای دسته بندی مورد استفاده قرار می‌گیرد. دسته بندی به عنوان یکی از شناخته شده ترین روش های داده کاوی از دو مرحله تشکیل می‌شود. هدف مرحله اول که مرحله استنتاج می باشد، کشف مدلی برای تعریف دسته های از پیش

مشخص شده داده ها است. در مرحله دوم که پیش بینی نام دارد، برای نمونه هایی که تعلق آن ها به دسته خاصی مشخص نیست، بر اساس مدل استنتاج شده می‌توان تعلق آن ها را پیش بینی کرد (۱۴). بنابراین مدل های پیش بینی کننده، پیش بینی داده های ناشناخته را با استفاده از مقادیر شناخته شده انجام می‌دهند (۱۵). روش های دسته بندی برای پیش بینی یا توصیف مجموعه های داده با طبقات دودویی یا اسمی مناسب‌تر هستند (۱۶).

اجرای الگوریتم های C5.0، C&R Tree و CHAID روی داده های موجود با هدف پیش بینی مقدار اثربخشی داروهای پره اکلامپسی صورت می‌گیرد. برای سنجش مقدار اثربخشی داروها با نظر متخصصین حوزه داروسازی و زنان و زایمان، یک متغیر اسمی بر اساس نوع و روش مصرف دارو و با در نظر داشتن جرم مصرف یکسان برای تمام بیماران در نظر گرفته شد. سپس متغیر اسمی انتخاب شده برای نگهداری میزان اثربخشی داروی مورد استعمال معرفی و به عنوان خروجی در نظر گرفته شد. جهت آزمون روایی، داده‌ها به دو دسته داده های آموزش و آزمایش تقسیم شدند که خروجی های نهایی بررسی و روایی پژوهش تأیید شد. داده‌های نمونه به دو زیر گروه (۷۰ درصد برای آموزش و ۳۰ درصد برای آزمایش) تقسیم و سپس اجرای الگوریتم های مورد نظر بر روی داده‌ها صورت گرفت. هنگام اجرای الگوریتم ها، ابتدا مدل با استفاده از داده های آموزش تولید شد و در گام بعدی با داده-های آزمایش مورد آزمون قرار گرفت. داده‌های آزمون که دسته آن ها مشخص بود به مدل داده شد و خروجی مدل با دسته‌های مشخص توسط مدل مقایسه گردید و سپس دقت کل مدل استخراج شد. در واقع، برچسب شناخته شده از نمونه آزمون با نتایج دسته بندی مقایسه شد (۲۰). همچنین پر واضح است تا زمانی که صحت مدل تعیین نشود، نمی‌توان درباره کارایی مدل قضاوت کرد (۱۳). بنابراین ابتدا صحت مدل‌های تولید شده مورد بررسی قرار گرفت و پس از آن، علل کاهش صحت مدل ها ریشه یابی شدند.

در مرحله بعد برای تأیید نتایج پیش بینی از خوشه بندی استفاده شد. خوشه بندی روش یادگیری غیر نظارتی می باشد که مفهوم فیلد خروجی در آن وجود ندارد (۱۲). برای تأیید نتایج پیش بینی روی مدل بهتر تولید شده، خوشه بندی K-Means را اجرا و نتایج حاصل از خوشه بندی تحلیل شدند. خوشه بندی با ۲، ۳، ۴، ۵، ۶ و ۷ خوشه روی مدل تولید شده اجرا شد. سپس تعداد بهینه خوشه بر روی مدل با بهره‌گیری از شاخص دان[1] بر اساس (۱) محاسبه شد. هدف شاخص دان، ماکزیمم نمودن فاصله درون خوشه ای و در ضمن، مینیمم کردن فاصله برون خوشه ای است (۱۷).

یافته ها

از درخت های تصمیم به وجود آمده می‌توان قوانین را استنتاج کرد. قوانین تولید شده توسط الگوریتم C5.0 شامل ۴ مجموعه قانون برای میزان اثربخشی خیلی زیاد، زیاد، متوسط و کم می‌باشد. علی رغم اینکه داروهای مورد بررسی ۵ نوع مختلف بودند، اما تنها ۴ مجموعه قانون استخراج شد و میزان اثربخشی یک دارو در نظر گرفته نشد. زیرا داروی مورد نظر دارای تعداد رکورد کمی در مجموعه داده مورد بررسی بود و عملاً وجود نمونه‌های مربوط به این دارو تأثیر چندانی بر قوانین تولید شده و ساختار درخت تصمیم نمی‌گذارد. درخت تولید شده با ریشه میزان اثربخشی دارو توسط این الگوریتم، یک درخت دودویی مورب چپ با عمق زیاد می‌باشد. در زیر چند نمونه از قوانین ایجاد شده که در هنگام تولید، رکوردهای بیشتری را در برگرفتند، آمده است.

- اگر داروی مورد استفاده هیدرالازین ۲۰ درصد باشد، آن گاه اثربخشی دارو زیاد می‌شود.
- اگر داروی مورد استفاده هیدرالازین ۲۰ درصد نباشد و نیتروگلیسرین باشد، آن‌گاه اثربخشی دارو خیلی زیاد می‌شود.

قوانین تولید شده توسط الگوریتم C&R Tree شامل ۴ مجموعه قانون برای میزان اثربخشی خیلی زیاد، زیاد، متوسط و کم می‌باشد. درخت تولید شده توسط این

[1] Dunn

الگوریتم نیز دودویی و نامتوازن می‌باشد. در زیر چند نمونه از قوانین ایجاد شده که در هنگام تولید، رکورد های بیشتری را در برگرفتند، آمده است.

- اگر روش مصرف دارو آمپول وریدی باشد، آن گاه اثربخشی دارو زیاد می‌شود.
- اگر روش مصرف دارو سرم یا قرص باشد و داروی مورد استفاده نیتروگلیسرین باشد، آن‌گاه اثربخشی دارو خیلی زیاد می‌شود.

قوانین تولید شده توسط الگوریتم CHAID نیز شامل ۴ مجموعه قانون برای میزان اثربخشی خیلی زیاد، زیاد، متوسط و کم می‌باشد. درخت تولید شده توسط این الگوریتم غیر دودویی و کم عمق می‌باشد. در زیر چند نمونه از قوانین ایجاد شده که در هنگام تولید، رکورد های بیشتری را در برگرفتند، آمده است.

- اگر روش مصرف دارو آمپول وریدی باشد، آن‌گاه اثر بخشی دارو زیاد می‌شود.
- اگر روش مصرف دارو سرم باشد، آن‌گاه اثربخشی دارو خیلی زیاد می‌شود.

درخت ایجاد شده توسط الگوریتم C5.0 پیش بینی را تنها بر اساس نوع داروی مورد استعمال انجام می‌دهد، در حالی که درخت های ایجاد شده توسط الگوریتم های C&R Tree و CHAID علاوه بر نوع داروی مورد استفاده، روش مصرف دارو را نیز در نظر می گیرند. همچنین بررسی مجموعه داده‌های ورودی نشان می‌دهد که سن بیمار، یکی از ورودی های مدل های تولید شده است، ولی درخت تصمیم های تولید شده، عامل سن را در میزان اثربخشی دارو نادیده گرفته‌اند. این نکته بیانگر این مطلب است که سن بیمار در نوع داروی انتخاب شده برای مصرف یا میزان تأثیر دارو اهمیت چندانی ندارد. جدول ۱ مقدار ضریب اطمینان برای قوانین حاصل شده در الگوریتم های مختلف را نشان می‌دهد. یافته های جدول ۱ نشان می‌دهد که عملکرد مدل های تولید شده برای میزان اثربخشی داروی مصرفی یکسان می‌باشد.

جدول ۱- مقدار ضریب اطمینان قوانین به دست آمده

نام الگوریتم	میزان تأثیر داروی مصرفی	ضریب اطمینان قانون تولید شده
C5.0	کم	۱
	متوسط	۰/۷۱۴
	زیاد	۱
	خیلی زیاد	۱
C&R Tree	کم	۱
	متوسط	۰/۷۱۴
	زیاد	۱
	خیلی زیاد	۱
CHAID	کم	۱
	متوسط	۰/۷۱۴
	زیاد	۱
	خیلی زیاد	۱

برای تعیین بهترین مدل پیش بینی کننده میزان اثربخشی داروهای مصرفی، نیاز به بررسی صحت مدل های تولید شده می‌باشد. شاخص های مختلفی برای ارزیابی صحت روش های دسته بندی وجود دارد که می‌توان حساسیت، شفافیت، دقت و صحت را نام برد. میزان صحت یک روش دسته بندی بر روی مجموعه داده های آموزشی، درصد مشاهداتی از مجموعه آموزشی است که به درستی توسط روش مورد استفاده، دسته بندی شده اند. مقدار شاخص های صحت، حساسیت، شفافیت و دقت بر اساس روابط (۱) تا (۴) محاسبه می‌شوند.

(۱) صحت = (TP+ TN) / (Pos+ Neg)

(۲) حساسیت = (TP) / (Pos)

(۳) شفافیت = (TN) / (Neg)

(۴) دقت = TP/ (FP+ TP)

TP= تعداد داده های برچسب مثبتی که درست دسته بندی شدند.

TN= تعداد داده های برچسب منفی که درست دسته بندی شدند.

FN= تعداد داده های برچسب مثبتی که نادرست دسته بندی شدند.

FP= تعداد داده های برچسب منفی که نادرست دسته بندی شدند.

Pos (TP+ FN) = کل تعداد داده های برچسب مثبتی که درست یا نادرست دسته بندی شدند.

Neg (TN+ FP)= کل تعداد داده های برچسب منفی که درست یا نادرست دسته بندی شدند (۱۷).

بر اساس نتایج به دست آمده، صحت مدل های ایجاد شده از اجرای الگوریتم های C5.0، C&R Tree و CHAID بر روی مجموعه داده آموزشی و آزمون برابر و به ترتیب مقادیر ۹۹/۶۳ درصد و ۹۹/۱۴ درصد بود. از طرفی مقادیر حاصل شده برای صحت مدل ها، نشان دهنده عدم دسته بندی صحیح برخی مقادیر در جای مناسب خود بود. با توجه به ضرایب اطمینان به دست آمده در جدول ۱ برای قوانین مختلف، می‌توان دلیل پایین بودن صحت را توجیه کرد. قوانینی که تأثیر متوسط داشتند، همگی در جای مناسب خود قرار نگرفتند و باعث پایین آمدن ضریب اطمینان مدل به وجود آمده و در نهایت پایین آمدن صحت کلی مدل شدند. با توجه به یکسان بودن میزان صحت مدل ها، درخت بهتر بر اساس نوع درخت انتخاب شد. الگوریتم مناسب برای تولید درخت تصمیم بهتر الگوریتم C&R Tree است که به دلایل زیر انتخاب شد.

- درخت تصمیم حاصل دودویی بوده و نقطه شکست آن در هر مرحله، داده های مورد بررسی را به دو زیر مجموعه تقسیم نمود. همچنین عمق زیاد درخت نشان می‌دهد که زیر مجموعه های به وجود آمده با جزئیات بیشتری در هر مرحله مورد بررسی قرار گرفتند. بنابراین درخت تولید شده نسبت به درخت غیر دودویی دقیق تر است. این عامل باعث برتری این مدل بر مدل الگوریتم CHAID شد.

- مجموعه قوانین تولید شده توسط این درخت علاوه بر نوع دارو، روش مصرف دارو را نیز مدنظر قرار داد که این عامل باعث برتری آن بر مدل الگوریتم C5.0 شد.

حساسیت ۱۰۰ درصد و دقت ۹۹/۵۰ درصد برای الگوریتم C&R Tree برتری الگوریتم انتخابی را نسبت به سایر الگوریتم ها تأیید می‌کند. پس از انتخاب الگوریتم مناسب، به اجرای خوشه بندی و محاسبه مقدار شاخص دان برای مدل بهتر پرداخته شد. یافته های حاصل از خوشه بندی و محاسبه مقدار شاخص دان برای مدل تولید شده توسط الگوریتم C&R Tree در نمودار ۱ نشان می‌دهند که تعداد بهینه خوشه عدد ۵ است. زیرا هر چه مقدار به دست آمده از شاخص دان بزرگتر باشد، بهتر است و تعداد خوشه ای که مقدار این شاخص را زیادتر کند، مقدار بهینه تعداد خوشه ها است (۱۷). جدول ۲ (پیوست ها) توضیحات کامل تری را درباره شاخص دان نشان می‌دهد.

جدول ۲: مقادیر شاخص دان برای خوشه بندی K-Means

	تعداد خوشه=۲	تعداد خوشه= ۳	تعداد خوشه= ۴	تعداد خوشه= ۵	تعداد خوشه= ۶	تعداد خوشه= ۷
خوشه ۱	*۱,۰۲۲	۳,۸۵۲	۳,۹۴۱	۴,۰۳۲	۴,۰۳۲	۴,۰۳۲
خوشه ۲	۱,۰۸۰	۱,۱۷۷	۱,۶۲۶	۴,۹۱۰	۴,۹۱۰	۴,۹۱۰
خوشه ۳		۳,۶۵۰	۳,۷۴۲	۳,۸۱۶	۱,۲۹۶	۱,۲۹۶
خوشه ۴			۳,۲۵۵	†۳,۴۸۲	۳,۴۸۲	۱,۳۶۸
خوشه ۵				۴,۸۶۰	۴,۸۶۰	۴,۸۶۰
خوشه ۶					۱,۲۱۴	۱,۲۱۴
خوشه ۷						۲,۵۱۵

*انتخاب کمترین مقدار هر ستون که بصورت زیر خط دار در جدول ۲ نشان داده شده‌اند.

†انتخاب بیشترین مقدار بین مقادیر زیرخط دار که بصورت تیره و زیرخط دار نشان داده شده است. این مقدار عدد خوشه بهینه می‌باشد.

نمودار ۱- تعیین تعداد خوشه بهینه برای مدل C&R Tree

یافته های حاصل از خوشه بندی روی نتایج درخت C&R Tree نشان می‌دهد که خوشه اول شامل ۸۷ نمونه داده است که میانگین سنی ۳۰/۳۹ داشتند و همگی داروی نیتروگلیسرین را به روش سرم دریافت کرده و میزان اثربخشی دارو برای آن ها خیلی زیاد بود. خوشه دوم شامل ۵ نمونه داده است که میانگین سنی ۲۷/۴ داشتند و همگی داروی هیدرالازین ۵۰ را به صورت قرص دریافت نمودند. میزان اثربخشی دارو برای افراد این خوشه متوسط بود. خوشه سوم شامل ۴۰۳ نمونه است که میانگین سنی ۲۹/۷۹ داشتند و همگی داروی هیدرالازین ۲۰٪ را به روش آمپول دریافت کردند. میزان اثربخشی دارو برای افراد این خوشه زیاد بود. خوشه چهارم شامل ۴۵ نمونه داده است که میانگین سنی ۲۹/۱۵ داشتند و همگی داروی هیدرالازین ۲۵ را به صورت قرص دریافت کرده و میزان اثربخشی دارو برای آن ها کم بود. خوشه پنجم نیز شامل ۲ نمونه داده است که میانگین سنی ۲۲/۵ داشتند و همگی داروی هیدرالازین ۱۰ را به صورت قرص دریافت کرده و میزان اثربخشی دارو برای آن‌ها ناچیز بود.

بحث

نتایج مطالعه حاضر بر اساس جدول ۱ نشان می‌دهد که در مدل های تولید شده، اگر میزان اثربخشی دارو متوسط باشد، ضریب اطمینان مدل برای قرار گرفتن نمونههای با میزان اثربخشی متوسط در مکان درست، پایین می آید و به همین دلیل، صحت کلی مدل پایین می‌آید. زیرا مقدار ضریب اطمینان به دست آمده در جدول ۱ برای این میزان اثربخشی کمتر از ۱ است و این نشان می‌دهد که داروهایی با این اثربخشی در جای مناسب خود در درخت تصمیم تولید شده قرار نگرفته اند. از طرف دیگر ضریب اطمینان ۱ برای میزان اثربخشی کم، زیاد و خیلی زیاد نشان می‌دهد که مصرف کنندگان داروهایی با این میزان اثربخشی در جای مناسب خود قرار گرفته اند. مقدار صحت برابر ۳ الگوریتم روی داده های آموزشی و آزمون دلالت بر انتخاب درست الگوریتم ها دارد، اما برای تعیین الگوریتم بهتر جهت نحوه عملکردش روی مجموعه داده مورد بررسی نوع درخت مورد بررسی قرار گرفت. درخت دودویی با عمق زیاد می‌تواند جزئیات بیشتری را در قوانین تولید شده دخیل سازد و از طرفی وارد کردن جزئیات بیشتری در تولید قوانین، مقادیر ۱۰۰ و ۹۹/۵ درصد برای حساسیت و دقت مؤکد این مطلب است که درخت به وجود آمده با الگوریتم C&R Tree مدل بهتری برای بررسی و پردازش بیشتر می‌باشد.

یافته‌های حاصل از خوشه بندی روی درخت C&R Tree نشان می‌دهد که خوشه بندی بر اساس نوع داروی مصرفی صورت گرفته است و عامل سن بیمار در خوشه بندی تأثیر چندانی نداشته است، همان گونه که در تولید درخت تصمیم سن بیمار تأثیرگذار نبود. اما ذکر این نکته ضروری است که سن بیمار در نوع داروی مصرفی بی‌تأثیر نیست، زیرا نتایج خوشه بندی نشان می‌دهد که در مواقعی که میزان اثربخشی دارو افزایش یافته است، میانگین سنی بیمار هم بالاتر بوده است.

خوشه ۱ و ۳ با میزان اثربخشی خیلی زیاد و زیاد نشان می‌دهند که داده‌های این دو خوشه بالاترین میانگین سنی و بیش ترین جمعیت مصرف کننده داروهای پره اکلامپسی را دارا بوده‌اند. شاید دلیل این مطلب رابطه مستقیم ایجاد عارضه پره‌اکلامپسی با افزایش سن باشد که با افزایش سن، احتمال وجود این عارضه در دوران بارداری افزایش یافته و از طرفی برای درمان بیماران با سن بیشتر، از داروهایی با میزان اثربخشی بالاتر استفاده می شود. خوشه های ۲، ۴ و ۵ نمونه هایی را شامل می‌شود که همگی به روش مصرف قرص، عارضه پره اکلامپسی را درمان کرده و میانگین سنی آن‌ها نسبت به دو خوشه ۱ و ۳ پایین تر است. بنابراین می توان نتیجه گرفت که عارضه پره اکلامپسی در سنین پایین‌تر با مصرف قرص قابل درمان است. برای تأیید نتایج حاصل از پیش بینی با توجه به جدول ۱ و خوشه بندی روی نتایج درخت C&R Tree می‌توان گفت نمونه هایی که باعث کاهش صحت کلی مدل شدند، همگی در خوشه ۲ قرار گرفتند و نمونه هایی که صحت مدل را بالا می‌برند، در خوشه های مجزای ۱، ۳ و ۴ قرار گرفتند. همچنین نمونه های خوشه ۵ به دلیل کم بودن تعداد در درخت تصمیم جای نگرفتند.

بررسی نمونه مطالعات مشابه که شامل موارد کاربرد روش های داده کاوی در زمینه بیماری پره اکلامپسی می‌باشد، نشان می‌دهد که از نظر هدف مطالعه حاضر در مطالعات خارجی به پیش بینی میزان اثربخشی داروهای پره اکلامپسی توجهی نشده است. به عنوان مثال فاچی[1] و همکاران (۲۰۰۹) به ارائه چارچوبی برای توصیف کاربرد الگوریتم‌های درخت تصمیم روی پره اکلامپسی پرداختند. چارچوب پیشنهادی به بررسی صدک وزن تولد اصلاح شده و هفته تولد نوزاد می پردازد (۱۸). غربالگری پره اکلامپسی با استفاده از داده کاوی پیشرفته به منظور شناسایی سرم جدید شاخص های بیولوژیک نیز انجام شده است که پنینگز[2] و همکاران (۲۰۱۱) از یک استراتژی فیلتر کننده سه مرحله‌ای شامل انتخاب ژن های خاص بافت، متن کاوی برای اولویت گذاری ژن آینده و شناسایی شاخص های قابل کشف خون استفاده کردند (۹). همچنین در مطالعات داخلی به عارضه پره اکلامپسی از دید داده کاوی پرداخته نشده است.

بر اساس نتایج مطالعه لیو[3] و همکاران (۲۰۱۲) و نیز بررسی مقالات موجود در زمینه داده کاوی بین سال های ۲۰۰۰ تا ۲۰۱۱ مشخص شد که تعداد ۱۷ مقاله به دسته بندی و ۹ مقاله به خوشه بندی از مجموع ۱۸۸ مقاله پرداختند (۱۹). بنابراین از نظر نوع تکنیک مورد استفاده در مطالعه حاضر با توجه به بکارگیری بیشتر تکنیک های دسته بندی در مطالعات پیشین و نیز تفسیر پذیری و قابل فهم بودن نتایج حاصل از درخت تصمیم (یک تکنیک دسته بندی)، این تکنیک به عنوان روش پیش بینی انتخاب شد. محدودیت موجود در مطالعه حاضر، نوع رشته ای خروجی است که استفاده از تعداد بیشتری از الگوریتم های پیش بینی را محدود می‌سازد. همچنین کم بودن تعداد صفات برای انتخاب زیر مجموعه‌ای از صفات جهت رسیدن به الگوریتم هایی با کارایی بالا نیز جزء محدودیت های مطالعه حاضر محسوب می شود.

در این مقاله با استفاده از روش‌های پیش بینی به استخراج قواعد مرتبط با این نوع سیستم های خبره پرداخته شد و این سیستم ها علاوه بر کمک به پزشکان جهت پیشگیری از تجویز اشتباه دارو، می‌توانند در پیش‌بینی میزان داروی مورد نیاز مراکز درمانی با توجه به شمار بیمارانی که به طور عمده به این مراکز مراجعه می‌کنند، مفید واقع شود. عدم قطعیت در پیش بینی تقاضای میزان دارو، یکی از مشکلات تأمین دارو می‌باشد که با انتخاب سطح موجودی مناسب، می‌توان احتمال مواجه با کمبود دارو در مراکز درمانی را کاهش داد. داده های سیستم های خبره تصمیم یار پزشک، به این نوع پیش بینی‌ها نیز کمک می‌کند.

نتیجه گیری

شواهد به دست آمده نشان می‌دهند که با افزایش سن

[1] Fiaschi
[2] Pennings
[3] Liao

احتمال وجود این عارضه افزایش و از داروهایی با میزان اثربخشی بالاتر برای درمان باید استفاده نمود. بنابراین در انتخاب داروی مصرفی با میزان اثربخشی متفاوت دقت بیشتری باید صورت گیرد.

شناخت میزان اثربخشی داروهای پره اکلامپسی، بیمارستان ها و مراکز درمانی را قادر می‌سازد تا برنامه ریزی سیستماتیکی برای سازماندهی داروهای مورد نیاز و پر مصرف به شیوه‌ای مؤثر انجام دهند. یافته های مقاله نشان می‌دهد که درخت های تصمیم تولید شده برای پیش بینی میزان اثربخشی داروها توسط الگوریتم های C5.0، C&R Tree و CHAID دارای صحت برابر بوده و به همین خاطر برای تعیین مدل بهتر تولید شده، از جهت ساختار نوع درخت را مد نظر قرار دادیم. زیرا هرچه مدل تولید شده دارای ساختار بهتر و جزئیات بیشتری باشد، دارای کارایی بهتری روی مجموعه داده مورد بررسی می‌باشد، به همین دلیل برای پیش بینی مناسب تر است. نتایج خوشه بندی روی مدل بهتر به دست آمده نشان می‌دهد نمونه هایی که باعث پایین آمدن صحت مدل می‌شوند، همگی در یک خوشه قرار می‌گیرند.

تشکر و قدردانی

بدین وسیله از پرسنل محترم بیمارستان حکیم نیشابور جهت همکاری در تهیه این مقاله تشکر و قدردانی می شود.

منابع

1. Li, B.N., Dong, M.C. Banking on Blood. IET Computing & Control Engineering, August/September 2006.
2. Ashoori,M., Naji Moghaddam,V., Alizadeh,S., Safi,M.Classification and Clustering Algorithm Application for Prediction of Tablet Numbers: Case Study Diabetes Disease. Health Information Management 1392; 10(5).
3. Rashidmehrabadi,A., Pedram,M. Classification and identification of blood donors tend to donate blood in the future. The Fourth Iran Data Mining Conference, Sharif University of Technology,2010,Tehran.
4. Ashoori, M., Taheri, Z. Using Clustering Methods for Identifying Blood Donors Behavior. 5th Iranian Conference on Electrical and Electronic Engineering (ICEEE), 20-22 August 2013, Gonabad, Iran, PP: 4055-4057.
5. Andrea L. Tranquilli, Beatrice Landi, Stefano R. Giannubilo, Baha M. Sibai. Preeclampsia: No longer solely a pregnancy disease, Pregnancy Hypertension: An International Journal of Women's Cardiovascular Health, (2012), PP: 350–357.
6. Cathrine Staff,N., Circulating predictive biomarkers in preeclampsia, Pregnancy Hypertension: An International Journal of Women's Cardiovascular Health, (2011), PP: 28–42.
7. Nikpour, S. Atarodi Kashani, Z. Mokhtarshahi, Sh. Parsay, S. Nooritajer, M. Haghani, H. Study of the Correlation of the Consumption of Vitamin C-Rich Foods with Preeclampsia and Eclampsia in Women Referred to Shahid Akbar Abadi Hospital in Tehran, Journal of Medical Sciences, 2004,Volume XIV / Issue 54.
8. Tabandeh,A., Argangi , H., Arabi, M . Serum hCG β in women with mild preeclampsia in pregnant women, Journal of Laboratory Medicine, 2013; Volume VII (1): 58-55.
9. Pennings, J.L.A., Kuc, S., Rodenburg, W., Koster, M.P.H., Schielen, P.C.J.I., Vries, A.D. Integrative data mining to identify novel candidate serum biomarkers for pre-eclampsia screening. Prenatal Diagnosis Journal, Vol. 31, No. 12, 2011, PP: 1153-1159.
10. Tabasi Z, Oghbaee N, Samimi M, Sadat Z. Comparison between Effects of Intravenous Labetalol and Hydralazine on Control of Hypertension and Maternal and Neonatal utcomes in Severe Preeclamptic Patients: A Randomized Clinical Trial. Qom Univ Med Sci J 2013;6(4):44-49.
11. Zabihi Mahmoodabadi A, Nobakht AR, Behrashi M, Musavi GH. Furosemide versus hydralazine for managing post partumhy pertension in severe preeclampsia: a comparative study. J Shahid Sadoughi Univ Med Sci 2012; 20(4): 482-88.
12. Modeling techniques in Clementine. Chapter11; Available from: URL:https://fhss.byu.edu/SPSS%20Modeler/Chapter%2011.pdf.
13. Rule Induction. Chapter12; Available from: URL: https://fhss.byu.edu/SPSS%20Modeler/Chapter%2012.pdf.
14. Khalilinezad,M., Minaee bidgoli,B. Data mining in medicine. The Third Iran Data Mining Conference,2009.
15. Bhardwaj, A., Sharma, A., and Shrivastava, V. Data Mining Techniques and Their Implementation in Blood Bank Sector –A Review. International Journal of Engineering Research and Applications (IJERA), Vol. 2, No. 4, July-August 2012, PP: 1303-1309.

16. Tan P.N, Steinbach M, Kumar V. Introduction to Data Mining. USA: Addison-Wesley Longman; 2005.
17. Ghazanfari,M., Alizade,S., Teymoorpour,B. Data Mining and Knowledge Discovery,2008,Iran University of Science and Technology.
18. Fiaschi, L., Garibaldi, J.M., Krasnogor, N. A Framework for the Application of Decision Trees to the Analysis of SNPs Data. IEEE conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB'09), 2009, PP: 106-113.
19. Liao S.-H, Chu P.-H, Hsiao P.-Y. Data mining techniques and applications – A decade review from 2000 to 2011. Expert Systems with Applications Journal, Vol. 39, 2012, PP: 11303-11311.
20. Toloie eshlaghy, A., Poor ebrahimi, A., Ebrahimi, M., Ghasem ahmad, L. Using Data Mining Techniques for Prediction Breast Cancer Recurrence. Iranian Journal of Diseases Breast, Vol. 5, No. 4, 2013, PP: 23-34.