استفاده از الگوریتم‌های دسته‌بندی و خوشه‌بندی برای پیش‌بینی تعداد قرص مصرفی: موردکاوی بیماری دیابت*

مریم عاشوری[1]، وجیهه ناجی مقدم[2]، سمیه علیزاده[3]، مهسا صفی[4]

مقاله پژوهشی

چکیده

مقدمه: امروزه با شیوع بیماری دیابت پیش‌بینی تعداد قرص مصرفی Glibenclamid و Metformin روزانه برای بیماران به پزشکان در جهت تشخیص تعداد قرص مصرفی بیمار و هم‌چنین مهار عوارض شدید و خطرناک مصرف بیش از حد دارو کمک می‌نماید. از این‌رو در پژوهش حاضر به‌منظور پیش‌بینی تعداد قرص مصرفی روزانه‌ی بیماران دیابتی، از تکنیک‌های داده‌کاوی استفاده شد.

روش بررسی: مطالعه‌ی حاضر به روش توصیفی- مقطعی صورت گرفت. نمونه‌گیری به روش سرشماری بود و تمامی بیماران (۲۷۸۳ بیمار) را در فاصله‌ی زمانی فروردین ۸۷ تا خرداد ۹۱ در برگرفت. جامعه‌ی پژوهش متشکل از داده‌های مرکز تحقیقات دیابت یزد وابسته به دانشگاه علوم پزشکی شهید صدوقی یزد بود. در مرحله‌ی پیش پردازش داده‌ها تعداد بیماران تحت بررسی به ۷۴۰ مورد رسید. روایی و پایایی روش جمع‌آوری داده‌ها مورد تأیید قرار گرفت. در این مطالعه جهت تحلیل داده‌ها و اجرای الگوریتم‌های داده‌کاوی از نرم‌افزار Clementine 12.0 استفاده شد. دو الگوریتم متفاوت از الگوریتم‌های استنتاج قانون به نام‌های C5.0 و CHAID روی داده‌ها اعمال گردید و سپس صحت مدل‌های تولید شده به‌دست آمد. در نهایت برای تأیید صحت مدل‌های تولید شده از خوشه‌بندی استفاده گردید.

یافته‌ها: مقادیر به‌دست آمده برای صحت مدل‌های ایجاد شده از اجرای الگوریتم‌های C5.0 و CHAID روی مجموعه داده‌های تحت بررسی ۴۵/۵۲ و ۲۸/۳۸ درصد بود. خوشه‌بندی نتایج به‌دست آمده از اجرای الگوریتم C5.0 تعداد قرص مصرفی ۳، ۵، ۶ و ۷ با صحت مقدار پیش‌بینی شده‌ی به ترتیب ۴۶/۸۳، ۳۶/۳۶، ۵۵/۷۱ و ۱۵ درصد را در یک خوشه قرار داد، زیرا نمونه داده‌هایی که دارای صحت پایینی در پیش‌بینی تعداد قرص مصرفی بود و یا تعداد نمونه داده‌ی کمی داشت، در یک خوشه قرار گرفتند. هم‌چنین خوشه‌بندی نتایج اجرای الگوریتم CHAID نیز تعداد قرص مصرفی ۵ با صحت مقدار پیش‌بینی شده‌ی ۲۰/۹۳ را در یک خوشه قرار داد.

نتیجه‌گیری: در مراکز تحقیقات دیابت وجود رویکرد سازمان‌دهی شده جهت پیش‌بینی تعداد قرص مصرفی به‌منظور کمک به پزشک برای افزایش صحت تشخیص و جلوگیری ازعوارض جانبی ناشی از تشخیص نادرست در تعداد قرص ضروری است. با توجه به لزوم استفاده از فن‌آوری‌های رایانه‌ای، اینترنت و نرم‌افزارهای تحلیلی و به‌منظور مهار اثرات خطرناک بیماری، بهتر است اقدامات لازم جهت ابداع رویکردهای پیشنهادی با مشاوره‌ی متخصصان انجام شود.

واژه‌های کلیدی: دیابت؛ درخت تصمیم؛ دسته‌بندی؛ خوشه‌بندی؛ شاخص Dunn

دریافت مقاله: ۹۱/۶/۲۷ اصلاح نهایی: ۹۲/۳/۶

پذیرش مقاله: ۹۲/۴/۱۸

ارجاع: عاشوری مریم، ناجی مقدم وجیهه، علیزاده سمیه، صفی مهسا. **استفاده از الگوریتم‌های دسته‌بندی و خوشه‌بندی برای پیش‌بینی تعداد قرص مصرفی: موردکاوی بیماری دیابت.** مدیریت اطلاعات سلامت ۱۳۹۲؛ ۱۰(۵): ۷۴۹-۷۳۹.

* این مقاله حاصل یک طرح تحقیقاتی داده‌کاوی در دانشگاه صنعتی خواجه نصیرالدین طوسی می‌باشد.

۱- دانشجوی کارشناسی ارشد، مهندسی فن‌آوری اطلاعات تجارت الکترونیک، دانشگاه صنعتی خواجه نصیرالدین طوسی، تهران، ایران (نویسنده‌ی مسؤول)
Email: maryam.ashoori@gmail.com

۲- دانشجوی کارشناسی ارشد، مهندسی فن‌آوری اطلاعات تجارت الکترونیک، دانشگاه صنعتی خواجه نصیرالدین طوسی، تهران، ایران

۳- استادیار، مهندسی صنایع، دانشگاه صنعتی خواجه نصیرالدین طوسی، تهران، ایران

۴- دانشجوی کارشناسی ارشد، مهندسی صنایع، دانشگاه صنعتی خواجه نصیرالدین طوسی، تهران، ایران

مقدمه

در دنیای پزشکی داده‌های مربوط به علایم بیماران مبتلا به بیماری‌های گوناگون و روش‌های کمکی برای تشخیص این بیماری‌ها بسیار وسیع و گسترده می‌باشد تا جایی‌که معمولاً تحلیل و در نظر گرفتن همه جانبه‌ی تمامی عوامل دخیل توسط یک فرد دشوار به نظر می‌رسد (۱). استخراج دانش از میان حجم انبوه داده‌های مرتبط با سوابق بیماری و پرونده‌های پزشکی افراد با استفاده از فرایند داده‌کاوی می‌تواند منجر به شناسایی قوانین حاکم بر ایجاد، رشد و تسریع بیماری‌ها گردیده و اطلاعات ارزشمندی را به‌منظور شناسایی علل رخداد بیماری‌ها، پیش‌بینی و درمان بیماری‌ها با توجه به عوامل محیطی حاکم در اختیار متخصصین و دست‌اندرکاران حوزه‌ی سلامت قرار دهد (۲). دیابت، بیماری مزمنی است که در نتیجه‌ی اختلال در تولید و عملکرد انسولین در بدن به وجود می‌آید. دیابت را برحسب علل ایجاد بیماری به دو نوع تقسیم می‌کنند: دیابت نوع اول، دیابت وابسته به انسولین یا دیابت جوانان نامیده می‌شود. دیابت نوع دوم، دیابت غیروابسته به انسولین یا دیابت بزرگسالان نامیده می‌شود (۳).

در این مقاله با توجه به اهمیت موضوع میزان مصرف قرص توسط بیماران دیابتی، محقق قصد دارد با پیاده‌سازی الگوریتم‌های دسته‌بندی C5.0 و CHAID (CHi-squared Automatic Interaction Detection) از مجموعه رویکردهای پیش‌بینی در داده‌کاوی به تعیین تعداد قرص بهینه‌ی مصرفی روزانه‌ی بیماران دیابتی پرداخته، سپس به معرفی الگوریتم کاراتر برای تولید مدل پیش‌بینی کننده بپردازد و در ادامه علل کاهش صحت و کارایی مدل پیش‌بینی کننده را ریشه‌یابی کند. با کشف کاراترین الگوریتم‌ها برای تشخیص میزان درستی آن‌ها با توجه به عوامل دخیل از جمله انتخاب زیرمجموعه صفات مناسب و نوع داده‌ی مورد استفاده، می‌توان به سمت ایجاد سیستم‌های مکانیزه با قابلیت اعتماد بالا با توانایی کشف الگوهای پیچیده و پیش‌بینی روندهای آینده برای مواجهه با انواع بیماری‌ها گام برداشت (۱). سپس برای تأیید صحت الگوریتم‌های پیش‌بینی کننده از خوشه‌بندی استفاده شد.

قرص‌های خوراکی ضد دیابت در ۵ دسته طبقه‌بندی می‌شوند. بعضی از انواع آن‌ها شامل Glibenclamid، Metformin، Acarbose، Pioglitazone، Repaglinide و غیره می‌باشند. هم‌چنین لازم به ذکر است که تنها مصرف قرص نمی‌تواند دیابت را ریشه‌کن کند. این قرص‌ها برای کمک به کار لوزالمعده و پایین آوردن قند خون تجویز می‌شوند. قرص‌های ضد دیابت مانند هر داروی دیگری به‌صورت بالقوه دارای عوارض جانبی می‌باشند و ممکن است در صورت مصرف زیاد دارو و یا به تأخیر انداختن وعده‌ی غذایی و حتی حذف یک وعده‌ی غذایی، بیمار با کاهش شدید قند خون مواجه شود. قرص‌های خوراکی مثل Glibenclamid و Metformin با تحریک سلول‌های تولیدکننده‌ی انسولین در لوزالمعده و یا کاهش مقاومت بدن نسبت به انسولین، سطح قند خون را پایین می‌آورند. قرص Metformin باعث کاهش اشتها و تثبیت وزن در بیماران دیابتی می‌شود. این دارو برای بیماران چاق و افراد دارای چربی خون بالا مناسب‌تر است. هم‌چنین افراد دارای بیماری‌های کبدی، کلیوی و نارسایی قلبی یا تنفسی نباید از Metformin استفاده کنند. با شروع مصرف این دارو ممکن است مشکلات گوارشی مثل تهوع و اسهال رخ دهد. با توجه به توضیحات داده شده درمی‌یابیم که در تجویز و میزان داروهای خوراکی باید دقت لازم اتخاذ شود (۴).

دو مسأله در خصوص پیش‌بینی تعداد قرص‌های مصرفی در بیماری دیابت مطرح است؛ اول اینکه تجویز بیش از حد لازم و کم‌تر از میزان مورد نیاز می‌تواند مشکلات بسیاری برای بیمار به‌وجود می‌آورد، دوم اینکه با در نظر گرفتن اینکه دیابت یک بیماری مزمن است، جهت تأمین و توزیع داروی مورد نیاز می‌بایست برنامه‌ریزی مناسب صورت گرفته و ذخیره‌ی مناسبی در این خصوص از طرف سازمان‌های مسؤول تأمین دارو در نظر گرفته شود. بنابراین با در اختیار داشتن پایگاه داده‌های مربوط به مشخصات بیماران می‌توان میزان مورد نیاز داروی بیماران دیابتی در یک منطقه را پیش‌بینی نمود. به این روش احتمال مواجهه‌ی بیماران با کمبود دارو و مشکل در توزیع کاهش می‌یابد که این دو مسأله ضرورت و اهمیت تحقیق حاضر را مشخص می‌کند. در عین حال با در نظر گرفتن

عوارض مصرف کم یا بیش از حد نیاز دارو برای بیماران دیابتی، ضرورت ایجاد یک سیستم تصمیم‌یار پزشک به‌خوبی احساس می‌شود که می‌توان با استفاده از بانک‌های اطلاعات این بیماران، این سیستم را ایجاد نمود.

امروزه حجم داده‌هایی که به‌صورت الکترونیکی در حوزه‌ی پزشکی ذخیره می‌شوند روز به روز در حال افزایش است. برای معنی بخشیدن به این داده‌ها باید آن‌ها را تحلیل و تبدیل به دانش کرد. با در نظر گرفتن چنین حجمی از الگوها و استفاده از انسان به‌عنوان تشخیص‌دهنده‌ی الگوها و تحلیل‌گر داده‌ها، پاسخ‌گویی به چنین حجم بالایی از اطلاعات امکان‌پذیر نیست و به همین دلیل داده‌کاوی در حوزه‌ی پزشکی از اهمیت بالایی برخوردار است. داده‌کاوی در پزشکی در پیش‌گیری و یا تشخیص نوع بیماری‌ها و انتخاب روش‌های درمان بیماری‌ها کاربرد دارد. مهم‌ترین خدمات قابل ارایه در پزشکی با استفاده از داده‌کاوی عبارتند از:

- بررسی میزان تأثیر دارو بر بیماری و اثرات جانبی آن
- تشخیص و پیش‌بینی انواع بیماری‌ها مانند تشخیص و یا پیش‌بینی انواع سرطان‌ها
- تعیین روش درمان بیماری‌ها
- پیش‌بینی میزان موفقیت اقدامات پزشکی مانند اعمال جراحی
- تجزیه و تحلیل داده‌های موجود در سیستم‌های اطلاعات سلامت (۲).

از این‌رو نظام اطلاعات بالینی دیابت به‌منظور شناسایی بیماران مبتلا به دیابت و گروه‌های مستعد در معرض خطر دیابت، بررسی چگونگی روند بیماری و ارایه‌ی طرح‌های مراقبت بهداشتی مورد نیاز، ایجاد ارتباط بین سایر ارایه‌دهندگان مراقبت بهداشتی و در نهایت بهبود مستمر کیفیت مراقبت از بیماران مبتلا به دیابت و هزینه‌های دیابت، داده‌های بیماران مبتلا به دیابت را گردآوری و پردازش و در قالب اطلاعات ارایه می‌دهد (۵). تاکنون مطالعات بسیاری در خصوص استفاده از داده‌کاوی روی داده‌های بیماران دیابتی انجام گرفته است. برای نمونه Su و همکاران برای تشخیص دیابت نوع ۲ از داده‌های AnthropometricalScanning استفاده نمودند. در این تحقیق فاکتورهای ریسک در داده‌های Anthropometrical در خصوص دیابت نوع ۲ مورد بررسی قرار گرفتند. در این مطالعه تکنیک‌های شبکه‌های عصبی، رگرسیون لجستیک، درخت تصمیم و Roughset برای پیش‌بینی دیابت استفاده شدند (۶).

طبق یافته‌های Liao و همکاران تکنیک‌های داده‌کاوی به‌عنوان کاربردی گسترده در زمینه‌های تخصصی برای کسب بینش جدید در زمینه‌ی مورد نظر مطرح گردیده‌اند. با بررسی مقالات موجود در زمینه‌ی داده‌کاوی بین سال‌های ۲۰۰۰ تا ۲۰۱۱ مشخص گردیده است که تعداد ۱۷ مقاله به دسته‌بندی و ۹ مقاله به خوشه‌بندی از مجموع ۱۸۸ مقاله پرداخته‌اند. الگوریتم CHAID در سال ۲۰۰۲ توسط Rygielski و همکاران، الگوریتم K-Means در سال ۲۰۰۷ توسط Adderley و همکاران و الگوریتم C5.0 در سال ۲۰۱۰ توسط Marx مورد استفاده قرار گرفته‌اند (۷). هم‌چنین با بهره‌گیری از تکنیک‌های داده‌کاوی برنامه‌ریزی دوز مصرفی برای بیماران دیابتی توسط Y1ld1r1m و همکاران صورت گرفته است. روش‌های ANFIS (Adaptive Neuro Fuzzy Inference System) و Rough Set روی داده‌های ۸۹ بیمار مختلف اعمال گردیده است که روش ANFIS در مقایسه با روش Rough Set موفق‌تر و قابل اعتمادتر بوده است (۸). بررسی مطالعات خارجی نشان می‌دهد که تاکنون مطالعات تحقیقی برای پیش‌بینی تعداد قرص مصرفی صورت نگرفته است. مطالعات کیوان پور و همکاران روی داده‌های دیابت با هدف تعیین بهترین الگو برای تشخیص بیماری صورت گرفته است. کیوان پور به این نکته اشاره می‌نماید که هیچ الگوریتمی وجود ندارد که همواره دارای کارایی بیشینه باشد و عوامل متعددی از جمله نوع داده‌ی مجموعه‌ی داده و انتخاب زیرمجموعه‌ی صفات در تغییر کارایی الگوریتم‌ها مؤثر هستند (۱).

هدف کلی از انجام تحقیق حاضر کشف دانش نهفته در داده‌های موجود در یکی از مراکز درمانی بیماران دیابتی می‌باشد. دانش حاضر می‌تواند علاوه بر یاری رساندن به پزشکان در تجویز دارو، در ارایه‌ی یک دید مناسب از شرایط مصرف دارو توسط بیماران دیابتی به مسؤولین ذی‌ربط کمک شایانی نماید.

روش بررسی

مطالعه‌ی حاضر از نوع توصیفی- مقطعی بوده و مجموعه داده‌های آن متعلق به مرکز تحقیقات دیابت یزد است. نمونه‌گیری به روش سرشماری بوده و تمامی بیماران (۲۷۸۳ بیمار) را در فاصله‌ی زمانی فروردین ۸۷ تا خرداد ۹۱ در برمی‌گیرد. در مرحله‌ی پیش پردازش داده‌ها جهت پاک‌سازی مجموعه داده، با نظر افراد خبره رکوردهایی که برخی فیلدهای آن خالی از مقدار بود، حذف شده و تعداد بیماران تحت بررسی به ۷۴۰ مورد رسید. این مجموعه داده شامل ۷۴۰ نمونه با ۸ صفت می‌باشد که بین سال‌های ۸۷ تا ۹۱ با بهره‌گیری از سیستم تحت وب جمع‌آوری گردیده است. این یافته‌ها با مراجعه‌ی مستقیم پژوهش‌گر به مرکز تحقیقات دیابت یزد و استخراج داده‌ها به‌صورت خروجی اکسل از سیستم تحت وب مذکور صورت گرفت و محتوای داده‌ها مورد تأیید مسؤولین مرکز دیابت می‌باشد. روایی روش جمع‌آوری اطلاعات توسط اساتید امر مورد تأیید می‌باشد. در این پژوهش جهت تحلیل داده‌ها و اجرای الگوریتم‌های داده‌کاوی از نرم‌افزار Clementine 12.0 استفاده شده است. جدول ۱ صفات مرتبط با این مجموعه داده به همراه توصیف مختصری از آن‌ها را نشان می‌دهد.

جدول ۱: مجموعه داده‌ی دیابت

نام فیلد	نوع داده
سن	عددی
جنسیت	رشته‌ای
تعداد مصرف روزانه‌ی قرص Glibenclamid	عددی
مدت مصرف Glibenclamid به روز	عددی
تعداد مصرف روزانه‌ی قرص Metformin	عددی
مدت مصرف Metformin به روز	عددی
گلوکز قبل از صبحانه	عددی
گلوکز ۲ ساعت پس از صبحانه	عددی

با توجه به مجموعه داده‌ی تحت بررسی، انتخاب الگوریتم مناسب برای اعمال روی داده‌ها ضروری به‌نظر می‌رسد. بنابراین ابتدا به بررسی الگوریتم‌های موجود برای انتخاب الگوریتم مناسب پرداخته می‌شود. پنج الگوریتم استنتاج قانون مختلف در Clementine 12.0 شامل C5.0، CHAID، QUEST، C&R Tree و لیست تصمیم می‌باشند (۱۰، ۹). از همه‌ی الگوریتم‌های نامبرده، درخت تصمیم یا مجموعه‌ای از قوانین برای توصیف بخش‌های مجزایی از داده‌های مرتبط با فیلد خروجی منشعب می‌گردد. خروجی مدل، دلیل هر قاعده را نشان داده و سپس برای درک فرایند تصمیم‌گیری مورد استفاده قرار می‌گیرد (۹). این الگوریتم‌ها درخت تصمیم را با تقسیم بازگشتی داده به زیرمجموعه‌هایی که توسط فیلدهای پیش‌بینی کننده تعریف می‌گردند که به نتیجه وابسته هستند، می‌سازد (۱۰). درخت تصمیم برای دسته‌بندی مورد استفاده قرار می‌گیرد. دسته‌بندی به‌عنوان یکی از شناخته شده‌ترین روش‌های داده‌کاوی از دو مرحله تشکیل می‌شود. در مرحله‌ی اول که مرحله‌ی استنتاج می‌باشد، هدف کشف مدلی برای تعریف دسته‌های از پیش مشخص شده‌ی داده‌ها است. مدل براساس نمونه‌های آموزشی ارایه شده به سیستم ایجاد می‌شود. الگوریتم استنتاج با استفاده از مقادیر مشخصه‌های نمونه‌هایی که به هر دسته تعلق دارند، تعریفی برای آن دسته‌ی خاص ایجاد می‌کند. در مرحله‌ی دوم که پیش‌بینی نام دارد، برای نمونه‌هایی که تعلق آن‌ها به دسته‌ی خاصی مشخص نیست، براساس مدل استنتاج شده می‌توان تعلق آن‌ها را پیش‌بینی نمود (۲). تکنیک‌های دسته‌بندی برای پیش‌بینی یا توصیف مجموعه‌های داده با طبقات دودویی یا اسمی مناسب‌تر هستند (۱۱).

الگوریتم‌های استنتاج قانون تفاوت‌هایی دارند که برای کاربران مهم هستند. در زیر تفاوت‌های این الگوریتم‌ها مورد بررسی قرار می‌گیرد.

- نوع خروجی: C5.0، QUEST و لیست تصمیم فیلد خروجی سمبلیک (از نوع رشته‌ای از رشته‌های مرتب) را مورد استفاده قرار می‌دهند. درخت C&R و CHAID قادر به تولید خروجی سمبلیک و عددی هستند و لیست تصمیم نتیجه دودویی را پیش‌بینی می‌نماید.
- نوع تقسیم‌بندی: زمانی‌که مجموعه داده به‌صورت بازگشتی به زیرگروه‌هایی تقسیم می‌گردد، درخت C&R و QUEST فقط تقسیم‌بندی به دو زیر گروه (زیر گروه آموزش و زیر گروه آزمون) را پشتیبانی می‌نمایند در حالی‌که CHAID، C5.0 و لیست تصمیم تقسیم‌بندی به بیش از دو زیر گروه (زیر گروه آموزش، زیر گروه آزمون و زیر گروه اعتبارسنجی) را پشتیبانی می‌نمایند.
- رشد سریع درخت و هرس: سه الگوریتم QUEST، C5.0 و درخت C&R درختانی با رشد سریع بوده، هرس نمودن آن‌ها رو به عقب است که یک روش شناخته‌شده‌ی اثربخش می‌باشد، اما معیارهای هرس متفاوتی دارند. C5.0 شامل صحت (بیش‌ترین صحت روی نمونه‌ی آموزشی) و عمومیت (نتایج برای دیگر داده‌ها عمومیت می‌یابند) است.
- نتایج: همه‌ی الگوریتم‌ها می‌توانند یک مدل در نتیجه‌ی مجموعه قوانین برای یک خروجی سمبلیک نمایش دهند. مجموعه قوانین می‌توانند به‌سادگی نسبت به درخت‌های تصمیم پیچیده تفسیر گردند. درخت تصمیم برای هر رکورد داده یک دسته‌بندی یکتا ارایه می‌دهد، در حالی‌که بیش از یک قانون در مجموعه‌ی قانون ممکن است به‌کار گرفته شود. زمانی‌که یک رکورد داده چندین قانون ارایه می‌دهد، به رکورد مورد نظر اولین قانون تعلق می‌گیرد (۱۰).

با توجه به تفاوت‌های ذکر شده می‌توان علت انتخاب دو الگوریتم C5.0 و CHAID را چنین ذکر نمود.

- خروجی تولید شده توسط این دو الگوریتم سمبلیک بوده و دودویی نمی‌باشد که این مسأله با توجه به نمونه داده‌ی تحت بررسی، اهمیت پیدا می‌کند.

پس از انتخاب الگوریتم‌های مناسب نوبت به اجرای الگوریتم‌ها روی داده‌های تحت بررسی می‌رسد. اجرای الگوریتم‌های C5.0 و CHAID روی داده‌های موجود با هدف تعیین مجموع تعداد قرص Glibenclamid و Metformin مصرفی توسط بیماران، صورت می‌گیرد که تحقق این امر با اندازه‌گیری میزان گلوکز قبل و دو ساعت پس از صرف صبحانه میسر می‌شود. در این راستا ابتدا دو متغییر جدید برای نگهداری مجموع قرص Glibenclamid و Metformin و مجموع گلوکز قبل و دو ساعت پس از صبحانه معرفی می‌گردد. سپس داده‌های موجود براساس فیلد مجموع قرص Glibenclamid و Metformin به‌صورت صعودی مرتب می‌شود و این فیلد به‌عنوان خروجی در نظر گرفته می‌شود. در ادامه مدل‌های تولید شده از اجرای الگوریتم‌های C5.0 و CHAID مورد ارزیابی قرار می‌گیرند. پر واضح است تا زمانی‌که صحت مدل تعیین نگردد، نمی‌توان درباره‌ی پایایی مدل قضاوت نمود (۱۰).

در مرحله‌ی بعد برای تأیید صحت مدل‌های تولید شده از خوشه‌بندی استفاده شد. خوشه‌بندی روش یادگیری غیرنظارتی می‌باشد که مفهوم فیلد خروجی در آن وجود ندارد (۹). برای تأیید صحت مدل‌های تولید شده، خوشه‌بندی K-Means را اجرا و نتایج حاصل از خوشه‌بندی تحلیل می‌شود. خوشه‌بندی برای تعداد ۲، ۳، ۴، ۵، ۶ و ۷ خوشه روی مدل‌های تولید شده‌ی حاصل از اجرای الگوریتم‌های C5.0 و CHAID اجرا می‌گردد. سپس تعداد بهینه‌ی خوشه روی هر مدل با بهره‌گیری از شاخص Duun طبق رابطه‌ی (۱) محاسبه می‌شود. هدف از شاخص Duun ماکزیمم نمودن فاصله‌ی درون خوشه‌ای در ضمن مینیمم کردن فاصله‌ی برون خوشه‌ای است (۱۲).

(رابطه ۱)

$$D = \min_{i=1\dots n_c}\left\{\min_{j=i+1\dots n_c}\left(\frac{d(c_i,c_j)}{\max_{k=1\dots n_c}(diam(c_k))}\right)\right\}$$

که $d(c_i, c_j)$ و $diam(c_i)$ طبق روابط (۲) و (۳) محاسبه می‌گردند.

$$d(c_i, c_j) = \min_{x\in c_i, y\in c_j}\{d(x,y)\}$$ (رابطه ۲)

$$diam(c_i) = \max_{x,y\in c_i}\{d(x,y)\}$$ (رابطه ۳)

جهت ارزیابی عملکرد الگوریتم مورد استفاده، شاخص صحت استفاده می‌شود. این شاخص نشان دهنده‌ی میزان پیش‌بینی صحیح الگوریتم دسته‌بندی می‌باشد. جهت محاسبه‌ی این شاخص ذکر برخی تعاریف ضروری است. «صحیح مثبت» داده‌های مثبتی هستند که به‌درستی توسط الگوریتم پیش‌بینی شده‌اند، در حالی‌که «صحیح منفی» داده‌های منفی هستند که به‌درستی پیش‌بینی شده‌اند. «غلط مثبت» داده‌های منفی هستند که به اشتباه پیش‌بینی شده‌اند. «غلط‌های منفی» داده‌های مثبتی هستند که به اشتباه پیش‌بینی شده‌اند. به این ترتیب میزان صحت یک الگوریتم دسته‌بندی به‌صورت ذیل محاسبه می‌شود: (۱۳)

$$\text{صحت} = \frac{\text{صحیح مثبت} + \text{صحیح منفی}}{\text{داده منفی} + \text{داده مثبت}} \quad \text{(رابطه ۴)}$$

یافته‌ها

از درخت‌های تصمیم به‌وجود آمده می‌توان قوانین را استنتاج نمود. قوانین تولید شده توسط الگوریتم C5.0 شامل ۶ مجموعه قانون برای تعداد ۲، ۳، ۴، ۵، ۶ و ۷ قرص مصرفی می‌باشد. تعداد زیاد قوانین تولید شده توسط الگوریتم C5.0 نشان‌دهنده‌ی عمق زیاد درخت می‌باشد. هم‌چنین درخت ایجاد شده توسط الگوریتم C5.0 دودویی و با ریشه‌ی مجموع ۴ قرص مصرفی روزانه می‌باشد. در زیر چند نمونه از قوانین ایجادشده که هنگام تولید رکودهای بیش‌تری را در بر گرفته‌اند، آمده است.

- اگر مدت مصرف Glibenclamid به روز ≥۹۰۸ باشد و مجموع مقادیر گلوکز قبل و ۲ ساعت پس از صرف صبحانه < ۴۸۳ و مدت مصرف Metformin به روز ≥۹۰۸ باشد، آن‌گاه مجموع تعداد قرص مصرفی= ۲ می‌شود.
- اگر مدت مصرف Glibenclamid به روز ≥۵۴۴ و < ۱۸۰ باشد و مجموع مقادیر گلوکز قبل و ۲ ساعت پس از صرف صبحانه ≥۴۸۳ و سن <۴۰ باشد آن‌گاه مجموع تعداد قرص مصرفی =۳ می‌شود.

قوانین تولید شده توسط الگوریتم CHAIDشامل ۲ مجموعه قانون برای تعداد ۴ و ۵ قرص مصرفی می‌باشد. تعداد کم قوانین تولید شده توسط الگوریتم CHAID نشان‌دهنده‌ی عمق کم درخت می‌باشد. درخت ایجاد شده توسط این الگوریتم غیر دودویی و با ریشه‌ی مجموع ۴ قرص مصرفی روزانه می‌باشد. در زیر قوانین تولید شده، آمده است.

- اگر مدت مصرف Glibenclamid به روز ≥۱۴۵۶ و مدت مصرف Metformin به روز ≥۷۲۸ باشد آن‌گاه مجموع تعداد قرص مصرفی =۴ می‌شود.
- اگر مدت مصرف Glibenclamid به روز ≥۱۴۵۶ و مدت مصرف Metformin به روز <۷۲۸ باشد آن‌گاه مجموع تعداد قرص مصرفی =۴ می‌شود.
- اگر مدت مصرف Glibenclamid به روز <۲۱۸۴ باشد آن‌گاه مجموع تعداد قرص مصرفی =۴ می‌شود.
- اگر مدت مصرف Glibenclamid به روز <۱۴۵۶ و ≥۲۱۸۴ باشد آن‌گاه مجموع تعداد قرص مصرفی =۵ می‌شود.

هم‌چنین یافته‌ها نشان می‌دهند که صحت مدل ایجاد شده از اجرای الگوریتم C5.0 روی مجموعه داده‌ی تحت بررسی مقدار ۴۵/۵۲ می‌باشد. در عین حال درصد صحت مدل ناشی از اجرای الگوریتم CHAID، ۲۸/۳۸ درصد به‌دست آمد. نتایج الگوریتم C5.0 نسبت به الگوریتم CHAID پایایی بالاتری را نشان داده و عملکرد بهتری دارد. مقادیر حاصل شده برای صحت نشان‌دهنده‌ی عدم دسته‌بندی صحیح برخی مقادیر در جای مناسب خود هستند. با این شرایط برای ریشه‌یابی عوامل مؤثر در کاهش صحت مدل به مقایسه‌ی مقادیر پیش‌بینی شده با مقادیر واقعی داده‌ها پرداخته می‌شود. جدول ۲ میزان صحت مقدار پیش‌بینی شده برای تعداد قرص مصرفی در مدل تولید شده توسط الگوریتم C5.0 را نمایش می‌دهد.

جدول۲: میزان صحت مقدار پیش‌بینی شده‌ی تعداد قرص مصرفی (الگوریتم C5.0)

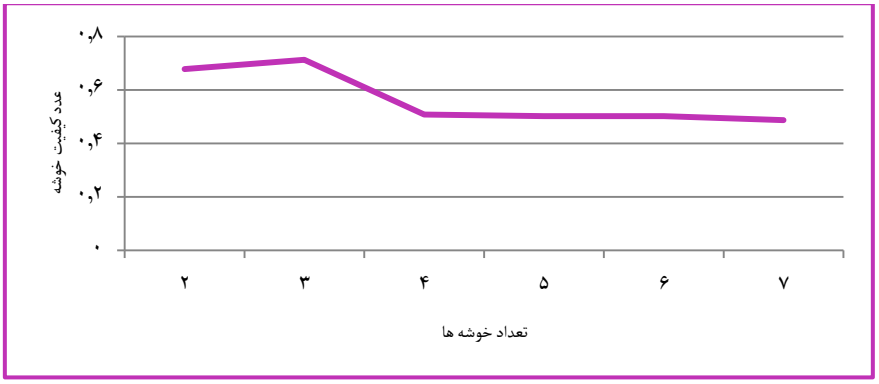
صحت مقدار پیش‌بینی شده در برابر مقدار واقعی	تعداد قرص مصرفی
۵۵/۲۲۴ درصد	۲
۴۶/۸۳۵ درصد	۳
۶۶/۱۶۵ درصد	۴
۳۶/۳۶۴ درصد	۵
۵۵/۷۱۴ درصد	۶
۱۵درصد	۷

جدول ۳ میزان صحت مقدار پیش‌بینی شده برای تعداد قرص مصرفی در مدل تولید شده توسط الگوریتم CHAID را نمایش می‌دهد.

جدول ۳: میزان صحت مقدار پیش‌بینی شده‌ی تعداد قرص مصرفی (الگوریتم CHAID)

صحت مقدار پیش‌بینی شده در برابر مقدار واقعی	تعداد قرص مصرفی
۹۰/۹۷۲ درصد	۴
۲۰/۹۳۰ درصد	۵

جدول ۲ نشان می‌دهد که در مدل تولید شده توسط الگوریتم C5.0 اگر تعداد قرص‌ها ۳، ۵ یا ۷ عدد باشد، مدل از ضریب اطمینان پایینی جهت قرار گرفتن این تعداد قرص در محل مناسب برخوردار است و به همین علت صحت کلی مدل پایین می‌آید. زیرا میزان صحت مقدار پیش‌بینی شده‌ی تعداد قرص مصرفی با مقدار واقعی در این دو حالت کم‌ترین مقدار است. جدول ۳ نشان می‌دهد که در مدل تولید شده توسط الگوریتم CHAID نیز اگر تعداد قرص‌ها ۵ عدد باشد، صحت مقدار پیش‌بینی شده در برابر مقدار واقعی ۲۰/۹۳۰ درصد می‌گردد که این مقدار نشان از ضریب اطمینان پایین قرار گرفتن تعداد قرص ۵ در محل مناسب خود در درخت تولید شده دارد. هم‌چنین یافته‌های به‌دست آمده از اجرای خوشه‌بندی و محاسبه‌ی مقدار شاخص Duun برای مدل تولید شده توسط الگوریتم C5.0 در نمودار ۱ نشان می‌دهد که تعداد بهینه‌ی خوشه عدد ۳ می‌باشد. زیرا هر چه مقدار به‌دست آمده از شاخص Duun بزرگ‌تر باشد، بهتر است و تعداد خوشه‌ای که مقدار این شاخص را زیادتر نماید، مقدار بهینه‌ی تعداد خوشه‌ها است (۱۲).



نمودار ۱: تعیین تعداد خوشه‌ی بهینه برای مدل C5.0

خوشه‌ی اول به‌دست آمده روی مدل C5.0 شامل ۲۴۵ نمونه داده می‌باشد. در این خوشه ۳۲/۶۵ درصد از نمونه‌ها را مجموع تعداد قرص ۳، ۳۱/۰۲ درصد از نمونه‌ها را مجموع تعداد قرص ۵، ۳۰/۲ درصد از نمونه‌ها را مجموع تعداد قرص ۶ و ۶/۱۲ درصد از نمونه‌ها را مجموع تعداد قرص ۷ تشکیل می‌دهد. خوشه‌ی دوم به‌دست آمده شامل ۱۸۴ نمونه داده می‌باشد که ۱۰۰درصد نمونه‌ها را مجموع تعداد قرص ۴ تشکیل می‌دهد و خوشه‌ی سوم شامل ۹۶ نمونه داده می‌باشد که ۱۰۰ درصد نمونه داده‌های این خوشه را نیز مجموع تعداد قرص ۲ تشکیل می‌دهد. به همین روش برای مدل CHAID مقدار ۲، عدد بهینه‌ی تعداد خوشه‌ها می‌باشد. خوشه‌ی اول در این حالت شامل ۷۱ نمونه داده با مجموع تعداد قرص ۵ می‌باشد و خوشه‌ی دوم نیز شامل ۴۵۴ نمونه داده با مجموع تعداد قرص ۴ می‌باشد.

بحث

یافته‌های پژوهش حاضر نشان می‌دهند که زیاد بودن تعداد قوانین تولید شده و درصد صحت مقادیر پیش‌بینی شده با مقادیر واقعی می‌توانند علت بالا بودن صحت مدل تولید شده توسط الگوریتم C5.0 نسبت به مدل تولید شده توسط الگوریتم CHAID را شرح دهند. در مدل تولید شده توسط الگوریتم

CHAID به‌خاطر وجود قوانین کم‌تر و هم‌چنین صحت پایین مقدار پیش‌بینی شده با مقدار واقعی برای تعداد ۵ قرص، صحت کلی مدل به شدت کاهش می‌یابد. به‌طور کلی مدل ساخته شده توسط الگوریتم C5.0 به‌دلیل صحت بیش‌تر و تولید مجموعه قانون کامل‌تر، از عملکرد بهتری روی داده‌های موجود برخوردار است. بررسی جدول ۲ نشان می‌دهد که مجموع تعداد قرص ۳، ۵ و ۷ که دارای صحت پایینی در پیش‌بینی مقادیر تعداد قرص مصرفی هستند و سبب کاهش صحت مدل می‌گردند، همگی با اجرای خوشه‌بندی در یک خوشه واقع شده‌اند. در عین حال مجموع تعداد قرص مصرفی ۶ به دلیل کم بودن تعداد نمونه‌ها (۷۴ نمونه) نسبت به نمونه داده‌های مجموع تعداد قرص مصرفی ۲ و ۴ و با وجود بالا بودن مقدار پیش‌بینی شده برای مقدار تعداد قرص مصرفی، در خوشه‌ای قرار گرفته که تعداد قرص ۳، ۵ و ۷ در آن واقع شده‌اند. به‌طور کلی این خوشه شامل مقادیری است که دارای صحت پایینی در پیش‌بینی تعداد قرص مصرفی بوده و صحت مدل را پایین می‌آورند. با توجه به جدول ۳ و خوشه‌بندی صورت گرفته روی مدل CHAID خوشه‌ی اول (مجموع تعداد قرص ۵) شامل نمونه داده‌هایی می‌باشد که صحت مقدار پیش‌بینی شده‌ی آن برای مجموع تعداد قرص مصرفی پایین است و سبب پایین آمدن صحت مدل می‌گردد. خوشه‌ی دوم (مجموع تعداد قرص ۴) شامل نمونه داده‌هایی است که صحت مقدار پیش‌بینی شده‌ی بالایی دارند. بنابراین با توجه به نظر کارشناسان پزشکی و مجموعه داده‌های تحت بررسی، خوشه‌بندی صورت گرفته روی مدل C5.0 به دلیل صحت بالاتر آن، بهتر است.

بررسی نمونه مطالعات مشابه که شامل موارد کاربرد تکنیک‌های داده‌کاوی در زمینه‌ی بیماری دیابت می‌باشد، نشان می‌دهد که از نظر هدف تحقیق تاکنون در مطالعات خارجی به پیش‌بینی تعداد قرص مصرفی بیماران توجهی نشده است. به‌عنوان نمونه Su و همکارانش برای پیش‌بینی دیابت نوع ۲ از تکنیک‌های شبکه‌های عصبی، رگرسیون لجستیک، درخت تصمیم وRoughset استفاده نمودند (۶). هم‌چنین برنامه‌ریزی دوز مصرفی برای بیماران دیابتی با بهره‌گیری از تکنیک‌های داده‌کاوی توسط Y1ld1r1m و همکاران صورت گرفته و روش‌های ANFIS (Adaptive Neuro Fuzzy Inference System) و Rough Set روی داده‌های ۸۹ بیمار مختلف اعمال گردیده است (۸). در نمونه‌ای داخلی ناجی مقدم و همکاران تعداد قرص مصرفی بیماران دیابت را با بهره‌گیری از الگوریتم‌های دسته‌بندی C5.0 و CHIAD پیش‌بینی نموده‌اند (۱۴).

طبق یافته‌های Liao و همکاران و نیز بررسی مقالات موجود در زمینه‌ی داده‌کاوی بین سال‌های ۲۰۰۰ تا ۲۰۱۱ مشخص گردیده است که تعداد ۱۷ مقاله به دسته‌بندی و ۹ مقاله به خوشه‌بندی از مجموع ۱۸۸ مقاله پرداخته‌اند (۷). از طرفی مطالعات کیوان پور و همکاران روی داده‌های دیابت با هدف تعیین بهترین الگو برای تشخیص بیماری نیز صورت گرفته و نتایج نشان می‌دهد که هیچ الگوریتمی وجود ندارد که همواره دارای کارایی بیشینه باشد و عوامل متعددی از جمله نوع داده‌ی مجموعه‌ی داده و انتخاب زیرمجموعه‌ی صفات در تغییر کارایی الگوریتم‌ها مؤثر هستند (۱). بنابراین از نظر نوع تکنیک مورد استفاده در پژوهش حاضر با توجه به به‌کارگیری بیش‌تر تکنیک‌های دسته‌بندی در مطالعات پیشین و نیز تفسیرپذیری و قابل فهم بودن نتایج حاصل از درخت تصمیم (یک تکنیک دسته‌بندی)، این تکنیک به‌عنوان روش پیش‌بینی انتخاب شد و برای شناسایی بهترین الگوریتم، الگوریتم‌هایی که با توجه به متغیر هدف مورد استفاده (تعداد قرص) قابل استفاده بود بر روی داده‌ها آزمون شد تا بهترین الگوریتم شناسایی شود. محدودیت موجود در پژوهش حاضر نوع سمبلیک خروجی است که استفاده از تعداد بیش‌تری از الگوریتم‌های پیش‌بینی را محدود می‌سازد. هم‌چنین کم بودن تعداد صفات برای انتخاب زیرمجموعه‌ای از صفات به جهت رسیدن به الگوریتم‌هایی با کارایی بالا نیز جزو محدودیت‌های پژوهش به حساب می‌آید.

با توجه به نتایج پژوهش، در مرکز تحقیقات دیابت وجود رویکرد سازمان‌دهی شده جهت پیش‌بینی تعداد قرص مصرفی بیمار به‌منظور کمک به پزشک برای افزایش صحت تشخیص و جلوگیری ازعوارض جانبی ناشی از تشخیص نادرست تعداد قرص خوراکی برای بیمار، ضروری است. نظام اطلاعات بالینی دیابت به‌منظور شناسایی بیماران مبتلا به دیابت و گروه‌های مستعد در معرض خطر دیابت، بررسی چگونگی روند بیماری و ارایه‌ی طرح‌های مراقبت بهداشتی مورد نیاز، ایجاد ارتباط بین

سایر ارایه‌دهندگان مراقبت بهداشتی و در نهایت بهبود مستمر کیفیت مراقبت از بیماران مبتلا به دیابت و هزینه‌های دیابت، داده‌های بیماران مبتلا به دیابت را گردآوری و پردازش و در قالب اطلاعات ارایه می‌دهد (۵). از جمله نقاط قابل بهبود در این حیطه پیش‌بینی میزان داروی مورد نیاز مراکز درمانی به جهت حفظ رفاه حال مراجعه کنندگان و ایجاد سیستم‌های پزشک‌یار جهت کمک به پزشکان در تجویز اولیه دارو می‌باشد.

نتیجه‌گیری

در این مقاله الگوریتم‌های C5.0 و CHAID روی مجموعه داده‌های بیماران دیابتی پیاده‌سازی و درخت تصمیمی برای پیش‌بینی تعداد بهینه‌ی قرص مصرفی روزانه‌ی بیماران دیابتی اتخاذ و سپس عمل خوشه‌بندی صورت گرفته است. شواهد به‌دست آمده نشان می‌دهند که در تعیین تعداد قرص مصرفی دقت بیش‌تری باید صورت گیرد. پیش‌بینی تعداد قرص مصرفی روزانه‌ی بیماران دیابتی می‌تواند با هدف دست‌یابی به مواردی چون؛ کمک به پزشک برای افزایش صحت تشخیص و جلوگیری از تشخیص نادرست در تعداد قرص خوراکی برای بیمار، تشخیص شدت دیابت و مهار عوارض خطرناک مصرف بیش از حد نیاز دارو توسط بیمار صورت گیرد، زیرا مصرف بیش از حد نیاز به قرص Glibenclamid سبب کاهش شدید قند خون و اختلالات الکترولیتی در سرم می‌گردد و نیز افراد دارای بیماری‌های کبدی، کلیوی و نارسایی قلبی یا تنفسی نباید از Metformin استفاده کنند. به همین خاطر باید در تجویز و میزان داروهای مصرفی دقت لازم صورت گیرد. درخت تصمیم تولید شده برای پیش‌بینی تعداد قرص مصرفی توسط الگوریتم C5.0 از صحت بالاتری برخوردار بوده و برای پیش‌بینی مناسب‌تر می‌باشد. هم‌چنین نتایج خوشه‌بندی صورت گرفته روی مدل‌های تولید شده نشان می‌دهد که نمونه داده‌هایی که سبب کاهش صحت مدل می‌گردند و دارای صحت پایینی در پیش‌بینی تعداد قرص مصرفی بوده و یا تعداد نمونه داده‌ی کمی دارند، در یک خوشه قرار می‌گیرند.

پیشنهادها

در این مقاله با استفاده از روش‌های پیش‌بینی به استخراج قواعد مرتبط با این نوع سیستم‌های خبره پرداخته شد و این سیستم‌ها 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 به‌طور عمده به این مراکز مراجعه می‌کنند، مفید واقع شود.

عدم قطعیت در پیش‌بینی تقاضای میزان دارو یکی از مشکلات تأمین دارو می‌باشد که با انتخاب سطح موجودی مناسب می‌توان احتمال مواجهه با کمبود دارو در مراکز درمانی را کاهش داد. داده‌های سیستم‌های خبره‌ی تصمیم‌یار پزشک به این نوع پیش‌بینی‌ها نیز کمک می‌کند.

تشکر و قدردانی

بر خود لازم می‌دانیم از مرکز تحقیقات دیابت یزد برای همکاری در تهیه این مقاله تشکر نماییم. بی تردید ثمردهی این پروژه بدون همکاری پرسنل محترم این مرکز تحقق نمی‌یافت.

References

1. Keyvan poor MR, Khalatbari L. Classified Algoritms Comparision in Diagnosis of Diabetes and Cardio Deficiency. Proceeding of the 3rd Iran Data Mining Conference; 2010 Feb 15; Tehran, Iran; 2010. [In Persian]
2. Khalilinezhad M, Minaee Bidgoli B. Clinical Data Mining. Proceeding of the 3rd Iran Data Mining Conference; 2010 Feb 15; Tehran, Iran; 2010. [In Persian]
3. Endocrine and Metabolism Research Institute. Education of TypeI Diabetes [Online]. 2010; Available from: URL: http://emri.tums.ac.ir/upfiles/ 60782275.pdf
4. Endocrine and Metabolism Research Institute. Oral Tablets for lowering blood glucose [Online]. 2010; Available from: URL: http://emri.tums.ac.ir/upfiles/61459053.pdf.
5. Hossieni AS, Moghadasi H, Jahanbakhsh M. Diabetes Clinical Information System in Countries. Health Inf Manage 2006; 3(1): 33-9. [In Persian]

6. Su CT, Yang CH, Hsu KH, Chiu WK. Data Mining for the Diagnosis of Type II Diabetes from Three-Dimensional Body Surface Anthropometrical Scanning Data. Computers and Mathematics with Applications 2006; 51(6-7): 1075-92.
7. Liao SH, Chu PH, Hsiao PY. Data Mining Techniques and Applications – A Decade Review from 2000 to 2011. Expert Systems with Applications 2012; 39(12): 11303-11.
8. Yildirim EG, Karahoca A, Ucar T. Dosage Planning for Diabetes Patients Using Data Mining Methods. Procedia Computer Science 2011; 3: 1374-80.
9. Modeling Techniques in Clementine [online]; Available from: URL: https://fhss.byu.edu/SPSS%20Modeler/Chapter%2011.pdf.
10. Rule Induction [online]; Available from: URL: https://fhss.byu.edu/SPSS%20Modeler/Chapter%2012.pdf.
11.Tan PN, Steinbach M, Kumar V. Introduction to Data Mining. USA: Addison-Wesley Longman; 2005.
12. Ghazanfari M, Alizadeh S, Teymour poor B. Data Mining and Knowledge Retrieval. Iran: iust; 2008. [Book in Persian]
13. Han J, Kamber M, Pai J. Data Mining: Concepts and Techniques. USA: Morgan Kaufman; 2000: 360-2.
14. Naji Moghadam V, Ashoori M, Alizadeh S, Safi M. The Classification Algorithm for Number of Tablet Usage Prediction: Case Study Diabetes. Proceeding of the 6rd Iran Data Mining Conference; 2012 Dec 18; Tehran, Iran; 2012. [In Persian]

Classification and Clustering Algorithm Application for Prediction of Tablet Numbers: Case Study Diabetes Disease*

Maryam Ashoori[1]; Vajihe NajiMoghadam[2]; Somayeh Alizadeh[3]; Mahsa Safi[4]

Original Article

Abstract

Introduction: By diabetes outbreak in these days, prediction of tablet daily usage like Glibenclamid and Metformin helps doctors to recognize number of tablets. Also, it should be considered that the need of diabeticto drug is critical. So, in this paper we have used data mining techniques to predict the number of daily usage of tablets for diabetes.

Methods: This study done by descriptive-cross sectional method. It done by Census sampling method and contains all 2783 patients from March 2008 to May 2012. In data preprocessing step the number of patients reduced to 740 cases. Data gathering method validity confirmed by supervisor and specialists. Also reliability value has compared. In this study Clementine 12.0 has been used for data analysis and data mining algorithms application. Two different algorithms namely CHAID and C5.0 have been used on data and then the generated models accuracy has been achieved. At the end, to confirm the accuracy, we have used clustering method.

Results: The obtained values for generated models accuracy by C5.0 and CHAID algorithm's execution on dataset was 45/52 and 28/38 respectively. The clustering of obtained results of C5.0 algorithm executing, put 3, 5, 6 and 7 of tablet usage with 46/83, 36/36, 55/71 and 15 percent of predicted value accuracy, respectively, in one cluster because the cases which have low accuracy or have low samples will be located in the same cluster. Also the clustering of CHAID algorithm executing results put 5 of tablet usage with 20/93 percent of predicted value accuracy in a cluster.

Conclusion: In Diabetes Center, an organized approach to predict number of daily usage tablets and prediction from side effects of false recognition in number of tablets is necessary. In order to prevent dangerous effects of diabetes, it is better to invent novel approaches by the help of expert consultant and use of computerized technologies, internet and analytical softwares.

Keywords: Diabetes; Decision Tree; Classification; Clustering; Dunn Index

Received: 17 Sep, 2012 Accepted: 9 Jul, 2013

Citation: Ashoori M, NajiMoghadam V, Alizadeh S, Safi M. **Classification and Clustering Algorithm Application for Prediction of Tablet Numbers: Case Study Diabetes Disease.** Health Inf Manage 2013; 10(5): 749.

* This article is derived from a Data Mining research project in K. N. Toosi University of Technology, Tehran, Iran.

1- MSc student, Information Technology Engineering, K. N. Toosi University of Technology, Tehran, Iran (Corresponding Author) Email: Maryam.ashoori@gmail.com
2- MSc student, Information Technology Engineering, K. N. Toosi University of Technology, Tehran, Iran
3- Assistant Professor, Industrial Engineering, K. N. Toosi University of Technology, Tehran, Iran
4- MSc student, Industrial engineering, K. N. Toosi University of Technology, Tehran, Iran