

شناسایی عوامل تأثیرگذار بر سکتة قلبی در بیماران دیابتی با استفاده از الگوریتم C&R

نویسندگان: حکیمه عامری*^۱، سمیه عزیزاده و اکبر برزگری^۲

۱- کارشناس ارشد فناوری اطلاعات - تجارت الکترونیک، دانشکده مهندسی صنایع دانشگاه
خواجه نصیرالدین طوسی تهران، ایران

۲- استادیار دانشگاه خواجه نصیرالدین طوسی تهران، دانشکده مهندسی صنایع دانشگاه خواجه
نصیرالدین طوسی تهران، ایران

۳- پزشک، دانشگاه علوم پزشکی گرگان، ایران

E-mail: hameri@mail.kntu.ac.ir

* نویسنده مسئول: حکیمه عامری

چکیده

مقدمه و هدف: بیماری‌های قلبی، شایع‌ترین علت مرگ‌ومیر در کشورهای توسعه‌یافته و همچنین در کل دنیا هستند و طبق پیش‌بینی سازمان بهداشت جهانی، عامل اصلی مرگ‌ومیر در سراسر دنیا در سال ۲۰۲۰ خواهند بود. طبق آخرین گزارش‌های سازمان بهداشت جهانی از هر ۲۰ مرگ‌ومیر، یکی به دلیل دیابت است. بیماری‌های قلبی و سکتة قلبی از مهم‌ترین عوارض دیابت به‌شمار می‌آیند. در این تحقیق با استفاده از الگوریتم‌های داده‌کاوی، احتمال بروز سکتة قلبی در بیماران دیابتی را با دقتی قابل‌قبول پیش‌بینی کرده، عوامل مؤثر در بروز سکتة قلبی شناسایی شده‌اند.

مواد و روش‌ها: در این مطالعه که به‌صورت گذشته‌نگر انجام شد، ۸۵۶ پرونده سال ۱۳۸۸ شمسی مرکز دیابت گرگان بررسی شده‌اند. اطلاعات موجود در پرونده بیماران با استفاده از روش‌های داده‌کاوی با نرم‌افزار SPSS CLEMENTINE تجزیه و تحلیل شدند. برای شناسایی عوامل مؤثر بر بروز سکتة قلبی از الگوریتم‌های دسته‌بندی داده‌کاوی استفاده شد.

نتایج: با استفاده از الگوریتم درخت تصمیم C&R، مدلی با دقت ۹۴ درصد معرفی شده‌است. براساس درخت تصمیم C&R، سابقه فشارخون، شاخص BMI، فشارخون سیستولیک و دیاستولیک، چربی خون با چگالی پایین، میزان فعالیت روزانه و سن از مهم‌ترین عوامل مؤثر در [بروز] سکتة قلبی در بیماران دیابتی شناسایی شده‌اند.

نتیجه‌گیری: می‌توان با استفاده از قوانین ایجادشده و شناسایی ویژگی‌های تأثیرگذار و کنترل عوامل مؤثر در بروز سکتة قلبی در بیماران دیابتی، میزان مرگ‌ومیر ناشی از این عارضه را تا حدی کاهش داد.

واژگان کلیدی: داده‌کاوی، دیابت، سکتة قلبی، الگوریتم درخت تصمیم.

دانشور

پزشکی

دوماهنامه علمی-پژوهشی
دانشگاه شاهد
سال بیست‌ویکم-شماره ۱۱۲
شهریور ۱۳۹۳

دریافت: ۱۳۹۳/۰۳/۲۱

آخرین اصلاح‌ها: ۱۳۹۳/۰۵/۱۸

پذیرش: ۱۳۹۳/۰۵/۲۲

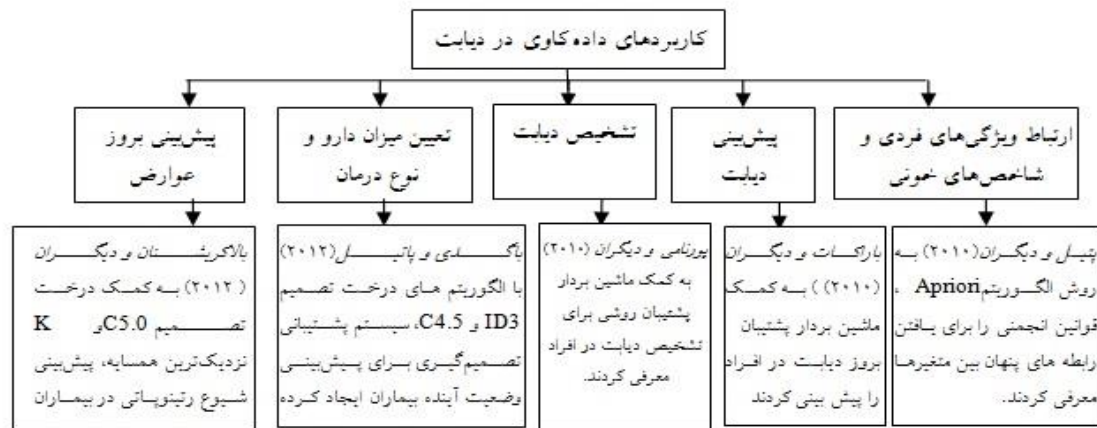
مقدمه

بروز دیابت در ده سال اخیر در سطح جهان دو برابر شده است و حدود ۲۰۰ میلیون نفر به این بیماری مبتلا هستند و سالانه شیوع دیابت در جهان، حدود ۶ درصد افزایش می‌یابد. در مطالعه‌ای که در ایران انجام شده است، گزارش شده که ۷/۷ درصد بالغان ۲۵ تا ۶۴ ساله که حدود ۲ میلیون نفرند، به دیابت، مبتلا بوده، ۱۶/۸ درصد بالغان معادل با ۴ میلیون نفر در وضعیت عدم تحمل گلوکز قرار دارند که تعداد زیادی از این بیماران در آینده به دیابت مبتلا خواهند شد (۱). براساس جدیدترین آمار ارائه شده، هر ۱۰ ثانیه، ۱ نفر به دلیل عوارض دیابت در جهان می‌میرد و هر ۱۰ ثانیه ۲ نفر به این بیماری مبتلا می‌شوند. از هر ۲۰ مرگ و میر، یکی به دلیل دیابت است. که معادل ۸۷۰۰ مرگ در هر روز، یا شش مورد مرگ در هر دقیقه است. افراد مبتلا به دیابت، دو تا چهار برابر بیشتر احتمال دارد به بیماری‌های قلبی - عروقی، مبتلا شوند (۲).

داده‌کاوی از رشته‌هایی جدید است که با به‌کارگیری و استفاده از داده‌های آماری، به استخراج اطلاعات و الگوهای مفید می‌پردازد. داده‌کاوی نشان‌دهنده پیشرفتی قابل توجه در انواع ابزار تحلیلی در دسترس است و به‌عنوان روشی معتبر، حساس و قابل اعتماد برای کشف الگوها و روابط میان آنها در نظر گرفته می‌شود (۳). یکی از زمینه‌های که می‌توان این دانش را به‌نحوی مؤثر استفاده کرد و نتایجی قابل توجه به دست آورد، داده‌های پزشکی است. خواجه‌وی و جایلاکشی، افزایش دقت تشخیص، کاهش هزینه‌ها و کاهش منابع انسانی را به‌عنوان مزایای معرفی داده‌کاوی در تجزیه و تحلیل پزشکی ثابت کرده‌اند (۴ و ۵). دسته‌بندی داده‌کاوی در پزشکی، عبارت است از: «بررسی میزان تأثیر دارو بر

بیماری، شناسایی عوارض جانبی داروها، تعیین نوع درمان، تجزیه و تحلیل داده‌های موجود در پرونده الکترونیک سلامت (EHR)، تشخیص و پیش‌بینی انواع بیماری‌ها مانند سرطان، تحلیل عکس‌های پزشکی مانند ماموگرافی، التراسونیک، اشعه X و MRI، ارائه مدل‌های توصیفی بر روی داده‌های پزشکی، کنترل عفونت بیمارستان و بهره‌برداری از خدمات سلامت» (۶)؛ برای نمونه، زایری و همکاران (۱۳۹۱)، مطالعه‌ای روی عوامل تأثیرگذار بر مرگ و میر پس از عمل جراحی کرونری بای پاس در بیماران غیروابسته به دیالیز انجام داده‌اند. داده‌ها از یک مطالعه مقطعی از ۱۳۹۰ بیمار در مدت ۳ سال از بیمارستان قلب شریعتی تهران جمع‌آوری شدند؛ در این تحقیق برای پیش‌بینی میزان مرگ و میر از مدل‌های رگرسیون لجستیک و رده‌بندی درختی با نرم‌افزارهای SPSS و CART استفاده شده است و نتایج با یکدیگر مقایسه شدند. در مدل رده‌بندی با دقت ۹۰ درصد، متغیرهای نارسایی کلیوی شدید، کارگذاری بالن پمپ داخل آئورتی حین و پس از عمل، تنفس طولانی مدت از راه دستگاه، مدت زمان خون‌رسانی قلب حین عمل بیش از ۱۶۰ به‌عنوان زیرگروه‌های دارای خطر بالای مرگ و افراد با عوارض قلبی بطنی پس از عمل جراحی به‌عنوان زیرگروه با خطر متوسط مرگ معرفی شدند؛ شاخص‌های حساسیت و ویژگی برای این مدل، به ترتیب ۸۲ درصد و ۸۹ درصد و برای مدل لجستیک ۸۰/۴ درصد و ۸۸ درصد به دست آمده‌اند (۷).

مطالعاتی بسیار در زمینه دیابت نوع ۲ با استفاده از روش‌های داده‌کاوی انجام شده‌اند. مطالعات انجام گرفته را می‌توان به پنج گروه کلی که در شکل ۱ نمایش داده شده است، تقسیم کرد (۸-۱۲).



شکل ۱. کاربردهای داده کاوی در دیابت

متغیرهای دموگرافیک، سابقه بیماری قلبی- عروقی، نمایه توده بدنی، LDL، HDL، تری‌گلیسیرید، قند خون ناشتا و ۲ ساعتی، سیگار، فشار خون سیستولیک، دیاستولیک و دور کمر مورد تحلیل قرار گرفته‌اند. مدلی با استفاده از نرم‌افزار MATLAB براساس الگوریتم‌های SCG، OSS و RP، به ترتیب با صحت پیش‌بینی ۷۸، ۷۶ و ۷۶ برای داده‌های آزمون و ۷۸/۳۷، ۷۴/۳۲ و ۷۵/۶۷ درصد برای داده‌های اعتبارسنجی به‌منظور پیش‌بینی سندروم متابولیک و شاخص HOMA-IR طراحی کرده‌اند.

هدف از انجام این تحقیق، بررسی علت بروز عارضه سکتة قلبی در بیماران دیابتی است. در این تحقیق با استفاده از روش‌های درخت تصمیم، قوانینی برای تشخیص سکتة قلبی در بیماران دیابتی ارائه شده‌اند. لازم به یادآوری است که ویژگی‌های مورد استفاده در روش پیشنهادی ما، ویژگی‌های بیماران دیابتی هستند و شاخص‌های مستقیم بیماری‌های قلبی بررسی نشده‌اند.

روش کار

روش‌هایی مختلف برای پیاده‌سازی و اجرای طرح‌های داده‌کاوی وجود دارند. یکی از روش‌های بسیار قوی، متدولوژی CRISP (Cross Industry Process For Data Mining) است (۱۸). در این مقاله، مدل پیشنهادی براساس CRISP ارائه شده‌است و چهار مرحله اصلی

مطالعاتی محدود در زمینه پیش‌بینی بروز عوارض در بیماران دیابتی انجام شده‌اند. تحقیق‌ها نشان داده‌اند که از ۸۰ درصد عوارض مزمن دیابت نوع ۲ می‌توان با شناسایی‌های اولیه افراد در معرض خطر جلوگیری کرد یا آنها را به‌تعویق انداخت؛ برای نمونه، هوان‌کو و همکاران با استفاده از دسته‌بندی ماشین بردار پشتیبانی (SVM) و با روش انتخاب ویژگی و تجسم‌سازی، وجود نروپاتی در بیماران دیابتی را پیش‌بینی کرده‌اند (۱۳). سجادی و همکاران با استفاده از روش‌های خوشه‌بندی در داده‌کاوی، عوامل خطر ساز بروز عوارض قلبی در بیماران دیابتی را شناسایی کرده‌اند؛ این عوامل عبارت‌اند از: افزایش فشار خون، افزایش شاخص توده بدنی و اختلال چربی‌های خون (۱۴). پارتیبیان و همکاران در سال ۲۰۱۱ با استفاده از شبکه بیزین، احتمال بروز مشکلات قلبی را در بیماران دیابتی پیش‌بینی کرده‌اند و برای انجام این عمل از مؤلفه‌هایی مانند سن، جنسیت، فشارخون و میزان قند خون استفاده کرده‌اند؛ دقت مدل به‌دست آمده، ۷۴ درصد بود (۱۵). پارتیبیان و همکاران در سال ۲۰۱۲ با استفاده از ماشین بردار پشتیبانی، مدلی برای پیش‌بینی احتمال بروز مشکلات قلبی در بیماران دیابتی با دقت مدل ۹۴/۶۰ درصد ایجاد کردند (۱۶). سدهی و همکاران (۱۷) با استفاده از روش شبکه‌های عصبی مصنوعی، مدلی برای پیش‌بینی توأم سندروم متابولیک و مقاومت به انسولین طراحی کرده‌اند؛

شناخت سیستم، شناخت داده‌ها و آماده‌سازی آنها، مدل‌سازی و ارزیابی و توسعه را دربرمی‌گیرد؛ هر یک از این فازها خود شامل زیربخش‌هایی می‌شوند. حرکت رو

به جلو و عقب میان فازهای مختلف، نیازاست، زیرا ورودی هر فاز به خروجی فاز مرحله پیشین، وابسته است؛ هر یک از این چهار فاز در شکل ۲ نشان‌داده شده‌اند.



شکل ۲. مدل پیشنهادی

درپی‌دارد(۱۹). علل بروز دیابت نامشخص‌اند؛ بااین‌حال به‌نظر می‌رسد، عوامل ارثی و عوامل محیطی مؤثرند. عوامل خطر ابتلا به دیابت نوع ۲ عبارت‌ند از: سن بالاتر از ۴۵ سال، اضافه وزن، داشتن پدر و مادر یا خواهر و برادر مبتلا به دیابت (وراثت خانوادگی)، داشتن فشار خون بالا (۱۶۰/۹۰ یا بالاتر)، داشتن کلسترول بالا (HDL ۳۵ یا کمتر و تری‌گلیسیرید ۲۵۰ یا بالاتر) و استرس زیاد (۲۰).

با توجه به رشد روزافزون بیماری دیابت و عدم وجود درمان قطعی برای این بیماری و تأثیرها و عوارضی شدید که روی اعضای حیاتی بدن در درازمدت برجای می‌گذارد، بررسی داده‌های جمع‌آوری شده درخصوص این بیماری می‌تواند مفید باشد. دیابت، سومین علت مرگ‌ومیر در جهان و مرگ‌ومیر ناشی از بیماری‌های قلبی، اولین دلیل مرگ‌ومیر بر اثر بیماری در جهان معرفی شده‌اند. سکتة‌های قلبی با پیشگیری از

در ادامه به بررسی هر یک از فازهای این مدل پیشنهادی متناسب با مطالعه موردی بیماری دیابت نوع ۲ می‌پردازیم.

شناخت سیستم

در این فاز به شناخت سیستم مورد نظر پرداخته می‌شود و سپس اهداف مدنظر و عوامل موفقیت کلیدی سیستم تعیین شده، دوباره بازنگری می‌شود. یکی از شایع‌ترین بیماری‌های متابولیسم، دیابت است که با سطوح بالای قند خون و اختلال متابولیسم کربوهیدرات، چربی و پروتئین مشخص می‌شود و با فقدان مطلق یا نسبی انسولین همراه است. افزایش مزمن قند خون در درازمدت صدمه‌رسانی به ارگان‌های مختلف حیاتی از جمله سیستم قلبی-عروقی، چشم‌ها، کلیه‌ها و سیستم عصبی و درنهایت، اختلال در کارکرد آنها را

شده‌اند. وایتن و همکاران در ثابت‌کردن که حذف عاقلانه، روشی کارآمد به‌جای جایگزین کردن ارزش‌ها با روش‌هایی مانند میانگین، انتساب تصادفی، انتساب رگرسیون و مدل‌های ییزی است (۲۲).

بعضی از اطلاعات موجود در پرونده، مانند نام و نام‌خانوادگی، شماره پرونده بیمار و نشانی حذف‌شدند؛ در مرحله بعدی، پرونده بیمارانی را که فقط یک‌بار مراجعه‌داشته‌اند، کنارگذاشتیم زیرا اطلاعاتی کامل از آزمایش‌ها و عوارض آنها در دسترس نبود.

بعضی از فیلدها به‌تنهایی اهمیتی نداشتند (مانند BUN (Blood urea) و CR (Chromium))؛ اگر این نسبت، میان ۱۰ تا ۲۰ باشد، نرمال و بیشتر از ۲۰ معنی خون‌ریزی گوارشی یا انسداد دستگاه ادرار است؛ نسبت این فیلدها نشان‌دهنده احتمال وجود عارضه کلیوی است و یا قد و وزن که به‌تنهایی اهمیتی ندارند، بلکه شاخص BMI آنها تاثیرگذار است؛ در نتیجه، این فیلدها حذف شدند و به‌جای آنها از شاخص‌های مرتبط استفاده شده‌است. شاخص BMI به این صورت محاسبه می‌شود:

$$BMI = \text{Weight}(kg) / (\text{Height}(cm))^2$$

متغیرهای کلسترول، لیپوپروتئین با چگالی بالا، لیپوپروتئین با چگالی پایین، فشار خون دیاستولیک، فشار خون سیستولیک، تری‌گلیسرید، شاخص توده بدنی، قند خون ناشتا، قند خون ۲ ساعت پس از غذا و اوره به کراتینین که دارای ارزش‌های عددی به‌صورت دامنه‌ای بودند، براساس منابع و سایت‌های معتبر علمی و پزشکی و تأیید پزشک متخصص به‌صورت کدبندی شده استفاده شده‌اند.

در نتیجه، پس از پالایش داده‌ها به رکوردهایی با مشخصات جدول ۱ رسیدیم.

عوامل خطرزا قابل‌کنترل است. با توجه به اینکه سکتة قلبی از عوارض بسیار شایع در بیماری دیابت است، پیدا کردن روشی برای شناسایی علت بروز این عارضه در بیماران دیابتی و کنترل این ویژگی‌ها می‌تواند در کاهش سکتة قلبی در بیماران، بسیار مؤثر باشد.

شناخت داده‌ها و آماده‌سازی آنها

در این فاز به جمع‌آوری داده‌های اولیه، توصیف داده‌ها، بازرسی و بررسی داده‌ها و اعتبارسنجی کیفیت داده‌ها می‌پردازیم. اطلاعات این تحقیق از یک مرکز درمان دیابت در شمال ایران جمع‌آوری شده‌اند. اطلاعات پرونده‌ها به سال ۱۳۸۸ مربوط‌اند. ۸۵۶ رکورد اولیه از بیماران وجودداشت که پس از پالایش و حذف رکوردهایی که اطلاعات اصلی آنها وجودنداشت، به ۲۵۲ رکورد نهایی می‌رسیم. میانگین سن بیماران ۵۳ سال و ۳۰ درصد آنها مرد و مابقی زن هستند. ۷۰ درصد بیماران دارای سابقه خانوادگی در دیابت هستند. ویژگی‌های آزمایشگاهی بیماران در این مرحله بررسی و شناسایی شدند.

در مهم‌ترین گام تحقیق (آماده‌سازی داده‌ها یا پیش‌پردازش داده‌ها) به بررسی پرونده بیماران پرداخته شده‌است. در جهان واقعی، داده همیشه کامل نیست و درخصوص اطلاعات پزشکی، این موضوع همیشه درست است. برای حذف تعدادی از تناقض‌ها و داده‌های ناقص درباره داده‌ها از پردازش داده استفاده شده‌است. چین و همکاران و هن و همکاران در بسیاری از روش‌های پردازش داده را ارائه کرده‌اند (۱۸ و ۲۱). در این تحقیق مواردی که ارزش ۰ (صفر) برای ویژگی‌های فشار خون و (Fasting blood Sugar) (2 Hour Post Prandial Blood) 2HDD-BS (glucose) و (Triglycerides) TG و BMI داشتند، حذف

جدول ۱. داده‌ها و ارزش‌های مربوط پس از پیش‌پردازش

ویژگی	علامت استفاده‌شده و مقیاس محاسبه	ارزش	نوع
سن براساس سال	Age	کمی گسسته	فاصله‌ای
جنسیت	Sex	زن=0; مرد=1	اسمی
شاخص توده بدنی	BMI index	کمبود وزن و لاغری) کمتر از: 1= ۱۸/۵; (وزن طبیعی) میان ۱۸/۵ و ۲۴/۹=2; (اضافه وزن) میان ۲۵ تا ۲۹/۹=3; (چاق) بیشتر از ۳۰=4	فاصله‌ای
قند خون ناشتا	Fbs(Fasting blood sugar in mg/dl)	1=70-100; 2=101-126; 3= more than>= 127	فاصله‌ای
قند خون ۲ ساعت پس از غذا	2hpp-bs(2 Hour PostPrandial - blood sugar in mg/dL)	1; Normal <=140; 2; Hidden Diabetes 140-199; 3; Diabetes >=200	فاصله‌ای
کلسترول	Chol(Serum cholesterol in mg)	1= optimal <= 200; 2=between200-239; 3= high above240	فاصله‌ای
لیپوپروتئین با چگالی بالا	HDL(high-density lipoprotein in mg/dL)	1= best>=60; 2= poor(HDL<= 40 and sex=1) or (HDL<= 50 and sex=0); 3= better level(between 40-59 for men and 50-59 for women)	فاصله‌ای
لیپو پروتئین با چگالی پایین	LDL(low-density lipoprotein in mg/dL)	1= optimal<=100; 2= near optimal100-129; 3= border line high130-159; 4= high160-189; 5= very high>=190	فاصله‌ای
تری گلیسرید	Trig(Triglyceride in mg/dl)	1=less than 150;normal 2=150-199; Slightly above normal 3= 200-499; high 4= >= 500; Very high	فاصله‌ای
اوره به کراتینین	Bun/cr (blood urea nitrogen (BUN) (mg/dL) / serum creatinine (mg/dL)	1= normal(after the kidney) =>10 and <=20 2= Prerenal (before the kidney) >20 3= Internal (within kidney) <10	فاصله‌ای
سابقه خانوادگی دیابت	Family history Dm(Diabetes mellitus)	سابقه خانوادگی نوع ۱ و ۲ دیابت دارد=1 سابقه خانوادگی دیابت ندارد=0	اسمی
سابقه فشار خون	Hyp(Hypertension in mmHg)	بلی=1 خیر=0	اسمی
اختلال چربی خون	Dislip(Dyslipidemia)	بلی=1 خیر=0	اسمی
سیگاری	Smoking(Smoker or not)	سیگاری=1 غیر سیگاری=0	اسمی
فشار خون سیستولیک	Sys bp(Systolic blood pressure Mm/Hg)	1=hypotension <=90; 2=desirable=90-119; 3=borderlinehypertension=120-139; 4=hypertension >=140	فاصله‌ای
فشار خون دیاستولیک	Diabp(Diastolic blood pressure Mm/Hg)	1= hypotension <60=; 2=desirable =61-79; 3=border line hypertension =80-89; 4= hypertension >=90	فاصله‌ای
میزان فعالیت بدنی	Sport	1= more than 3 days per week 2= less than 3 days per week 3= more than 30 minutes per day 4= daily	رتبه‌ای
میزان تحصیلات	Education Level	2= diploma 1= university 3= pre high school; 4= not educated	رتبه‌ای
سکتة قلبی	Heart failure	1= yes 0= no	اسمی

مدل‌سازی و ارزیابی

ابتدا لازم است داده‌های موجود را به سه بخش آموزش، آزمایش و اعتبارسنجی تقسیم کنیم؛ داده‌های بخش آموزش، درخت را تولید می‌کنند و داده‌های بخش آزمایش با کمک تعدادی رکورد، درخت تولید شده را آزمایش و برچسب مربوط به رکوردهای یاد شده را تعیین می‌کنند؛ داده‌های بخش اعتبارسنجی نیز صحت مدل تولید شده را بررسی می‌کنند. شاخص‌هایی مختلف برای ارزیابی صحت روش‌های دسته‌بندی وجود دارند که می‌توان حساسیت (Sensitivity)، شفافیت (Specificity)، دقت (Precision) و صحت (Accuracy) را نام برد. میزان صحت یک روش دسته‌بندی روی مجموعه داده‌های آموزشی، درصد مشاهداتی از مجموعه آموزش است که به درستی، توسط روش مورد استفاده دسته‌بندی شده است؛ برای محاسبه این شاخص از داده‌های آزمون استفاده می‌شود؛ همچنین می‌توان نرخ خطا (Error Rate) یا دسته‌بندی نادرست (Misclassification rate) را براساس شاخص صحت محاسبه کرد (۱۸ و ۲۷).

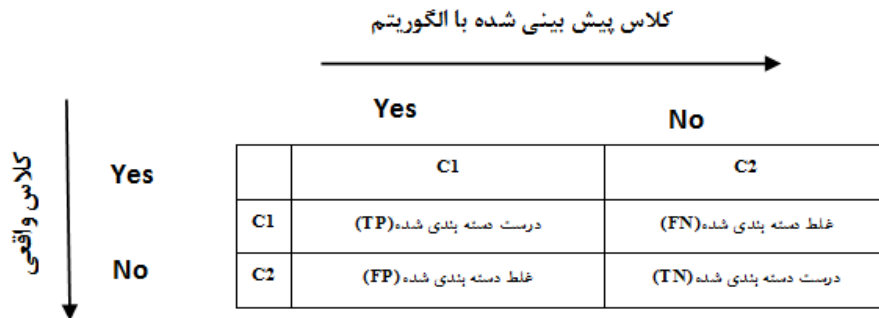
$$(۲) \quad \text{صحت} = 1 - \text{نرخ خطا}$$

برای محاسبه میزان صحت مدل می‌توان ماتریس اغتشاش (Confusion Matrix) را به کار برد؛ این ماتریس، ابزاری مفید برای تحلیل چگونگی عملکرد روش دسته‌بندی در تشخیص داده‌ها یا مشاهدات دسته‌های مختلف است؛ اگر داده‌ها در M دسته قرار گرفته باشند، یک ماتریس دسته‌بندی جدولی با حداقل اندازه $M \times M$ است. حالت مطلوب، این است که بیشتر داده‌های مرتبط به مشاهدات، روی قطر اصلی ماتریس قرار گرفته باشند و سایر مقادیر ماتریس ۰ (صفر) یا نزدیک به ۰ باشند (۱۸ و ۲۴).

روش‌های داده‌کاوی بسیاری برای مدل‌سازی وجود دارند. در این فاز با استفاده از روش‌های مختلف داده‌کاوی به پیدا کردن مدل و الگوی بهینه می‌پردازیم. یکی از روش‌های متداول، دسته‌بندی درخت تصمیم است. درخت تصمیم، توصیفی صریح از شاخه‌زنی با استفاده از الگوریتم است. از روش درخت طبقه‌بندی استفاده شده تا بهترین نسبت میان فیلدهای مختلف به دست آید؛ این درخت، ساختاری شبیه به فلوجارت دارد که بالاترین گره، ریشه درخت است و هر شاخه بیانگر خروجی‌های آزمایش بوده، گره‌های برگ، دسته یا توزیع دسته‌ها را نمایش می‌دهند (۱۸ و ۲۳).

قوانین ایجاد شده توسط درخت تصمیم به صورت «اگر» و «آنگاه» بیان می‌شوند؛ در این مرحله، درخت‌های تصمیم C5.0، QUEST، CHAID، C&R و شبکه عصبی مورد آزمایش قرار گرفته‌اند. از الگوریتم‌های پرکاربرد درخت تصمیم، الگوریتم C&R است؛ این درخت از دسته‌بندی و رگرسیون برای پیش‌بینی استفاده می‌کند. هر گره ابتدا فیلدهای ورودی را برای یافتن بهترین تجزیه آزمایش می‌کند تا شاخص ناخالصی حاصل از تجزیه کمترین مقدار باشد. تمام تجزیه‌ها دودویی هستند و تا زمانی ادامه می‌یابند که یکی از معیارهای توقف برآورده شود. درک مدل‌های این درخت و تفسیر قوانین استخراج شده نسبت به سایر مدل‌ها ساده‌تر است. برای آموزش درخت تصمیم مدل C&R، یک متغیر طبقه‌ای باید فیلد خروجی باشد و یک یا تعداد بیشتری فیلد ورودی وجود داشته باشند (۲۴).

پس از مدل‌سازی باید به ارزیابی نتایج حاصل از مدل‌سازی پرداخت. نتایج ارزیابی باعث بهبود مدل شده، مدل را قابل استفاده می‌کنند. برای بررسی صحت مدل،



شکل ۳. ماتریس اغتشاش

$$\text{حساسیت} = \frac{\text{تعداد داده های برجسب مثبتی که درست دسته بندی شده اند}}{\text{کل تعداد داده های مثبتی}} = TP/pos \quad (۳)$$

$$\text{شفافیت} = \frac{\text{تعداد داده های برجسب منفی که درست دسته بندی شده اند}}{\text{کل تعداد داده های منفی}} = TN/neg \quad (۴)$$

$$\text{دقت} = TP/(TP + FP) \quad (۵)$$

$$\text{صحت} = \frac{pos}{(pos + neg)} \text{حساسیت} + \frac{neg}{(pos + neg)} \text{شفافیت} \quad (۶)$$

توسعه

ساخت مدل، پایان یک طرح نیست و هدف از طرح های داده کاوی، کشف دانش و استفاده از دانش کشف شده در آینده است. دانش کشف شده باید سازماندهی شده، به شکلی قابل استفاده برای دیگران نیز درآید؛ ما در این فاز، علاوه بر تهیه گزارش تلاش کردیم تا نشان دهیم که میزان تأثیرگذاری مؤلفه های مختلف روی سکتة قلبی چه میزان است. می توان از این دانش برای پیش بینی وضعیت بیماران جدید استفاده کرده، در جهت کنترل سکتة قلبی در بیماران، همگام با دانش کشف شده از داده های پیشین گام برداشت.

یافته ها

هدف از داده کاوی استخراج دانش از اطلاعات ذخیره شده در پایگاه داده و ایجاد شرحی روشن و قابل فهم از الگوهاست. از الگوریتم های مختلف درخت تصمیم برای پیدا کردن بهترین نتیجه استفاده شده است. در جدول ۲ میزان صحت، دقت، شفافیت و حساسیت را برای درخت های تصمیم استفاده شده با هم مقایسه کرده ایم.

جدول ۲. مقایسه شاخص های دقت، صحت، حساسیت و شفافیت برای درخت های تصمیم مورد بررسی

درخت تصمیم	حساسیت	شفافیت	دقت	صحت	نرخ خطا
C5.0	۹۷ درصد	۵۸ درصد	۹۰ درصد	۹۲ درصد	۰/۰۸
CHAID	۹۸ درصد	۲۰ درصد	۸۳ درصد	۹۰ درصد	۰/۱
QUEST	-	-	-	۸۶ درصد	۰/۱۴
NURAL NETWORK	-	-	-	۸۶ درصد	۰/۱۴
C&R	۹۶ درصد	۶۰ درصد	۹۱ درصد	۹۴ درصد	۰/۰۶

مهم‌ترین اطلاعات مرتبط با درخت تصمیم را دربردارند. جدول ۳، مجموعه قوانین ایجادشده توسط گره درخت C&R را نمایش می‌دهد. معیار اطمینان (Confidence) برای قوانین ایجادشده، ۶۰ درصد تعیین شده است؛ با استفاده از این قوانین، می‌توان ارتباطی میان ویژگی‌های مؤثر بر سکت قلبی بیماران دیابتی را شناسایی و در جهت کنترل این عوامل خطر ساز تلاش کرد.

مقادیر حساسیت، شفافیت و دقت برای الگوریتم‌های QUEST و شبکه عصبی به این دلیل که درختی مناسب ایجاد نشده بود، غیر قابل محاسبه است. بهترین نتایج با استفاده از گره درخت C&R به دست آمده‌اند. با بررسی درخت تصمیم C&R، ویژگی‌های تأثیرگذار بر سکت قلبی در بیمار شناسایی شدند که عبارت‌اند از: سابقه فشار خون بالا، شاخص BMI، مقدار کنترل نشده برای لیپو پروتئین با چگالی پایین (کلسترول مضر)، بالا بودن سن و نداشتن تحرک زیاد. اغلب، مجموعه‌های قوانین،

جدول ۳. قواعد ایجادشده توسط درخت تصمیم C&R

دقت و اطمینان	قوانین ایجادشده
۱/۰	اگر قند خون، ۲ ساعت پس از خوردن غذا در بازه نرمال یا دیابت پنهان باشد، آنگاه بیمار به سکت قلبی، دچار نمی‌شود.
۰/۸۶	اگر قند خون، ۲ ساعت پس از خوردن غذا در بازه دیابتی باشد و شاخص BMI نشان‌دهنده وزن نرمال یا اضافه وزن باشد ولی سابقه فشارخون نداشته باشد، آنگاه بیمار به سکت قلبی، دچار نمی‌شود.
۰/۷۴	اگر قند خون، ۲ ساعت پس از خوردن غذا در بازه دیابتی باشد و بیمار، سابقه فشارخون داشته باشد و قند خون ناشتا بیشتر از ۱۲۷ باشد و تری‌گلیسیرید در بازه نرمال یا بالا باشد، آنگاه بیمار به سکت قلبی، دچار نمی‌شود.
۱/۰	اگر قند خون، ۲ ساعت پس از خوردن غذا در بازه دیابتی باشد و بیمار، سابقه فشارخون داشته باشد و قند خون ناشتا بیشتر از ۱۲۷ باشد، تری‌گلیسیرید در بازه بالاتر از نرمال یا خیلی بالا و کلسترول بالاتر از ۲۴۰ باشد، آنگاه بیمار به سکت قلبی، دچار نمی‌شود.
۰/۷۳	اگر قند خون، ۲ ساعت پس از خوردن غذا در بازه دیابتی باشد و بیمار، سابقه فشارخون داشته باشد و قند خون ناشتا میان ۷۰ تا ۱۲۶ باشد، آنگاه بیمار به سکت قلبی، دچار می‌شود.
۰/۷۱	اگر قند خون، ۲ ساعت پس از خوردن غذا در بازه دیابتی باشد و بیمار، سابقه فشارخون داشته باشد و قند خون ناشتا بیشتر از ۱۲۷ باشد، تری‌گلیسیرید در بازه بالاتر از نرمال یا خیلی بالا و کلسترول خیلی بالا باشد (بالاتر از ۱۴۰)، آنگاه بیمار به سکت قلبی، دچار می‌شود.

مطابق جدول شماره ۴، براساس الگوریتم C&R، مهم‌ترین عوامل مؤثر بر سکت قلبی بیماران دیابتی معرفی شده‌اند.

جدول ۴. عوامل تأثیرگذار بر سکت قلبی براساس الگوریتم C&R

ویژگی	وزن اهمیت
سابقه فشارخون بالا	۰/۳۶۸
شاخص BMI	۰/۱۹۸
فشارخون سیستولیک	۰/۰۸۱
فشارخون دیاستولیک	۰/۰۸۱
چربی خون با چگالی پایین	۰/۰۸۱
سن	۰/۰۸۱
میزان فعالیت بدنی	۰/۰۸۱
قندخون ناشتا	۰/۰۱۸
کاسترول	۰/۰۱۱

آن برابر ۹۱ درصد و صحت مدل ۹۴ درصد است؛ به این معنی که با استفاده از قوانین یافته شده، می توان احتمال بروز سکت قلبی را در بیمار دیابتی براساس ویژگی های دموگرافیک و نتایج آزمایشگاهی بیمار با دقت ۹۱ درصد پیش بینی کرد. بیشترین مؤلفه های تأثیرگذار بر سکت قلبی در بیماران دیابتی، سابقه فشار خون بالا، شاخص BMI، مقدار کنترل نشده برای کلسترول مضر، بالا بودن سن و نداشتن تحرک زیاد شناخته شده اند. با استفاده از قوانین ایجاد شده، برای یک نمونه جدید با ویژگی های مشخص، می توان پیش بینی کرد که «این فرد احتمال دارد [در آینده] به سکت قلبی، دچار شود یا خیر؟». با کنترل عوامل تأثیرگذار در هر بیمار، می توان امیدوار بود از بروز عارضه تاحدی اجتناب کرد و به این صورت، میزان مرگ و میر ناشی از سکت های قلبی را تاحدی کاهش داد.

در ادامه، نتایج کار خود را با کارهای مشابه در این حوزه در جدول ۵ مقایسه کرده ایم.

افراد متخصص نیز، ارزیابی قوانین را انجام دادند. با توجه به قواعد تولید شده، مؤلفه هایی که بیشترین تأثیر را بر سکت قلبی دارند، شناسایی و تأیید شده اند. طبق نظر کارشناسان می توان از مدل ایجاد شده، نتایج زیر را گرفت. با کنترل میزان قند خون، فشار خون و چاقی می توان از سکت قلبی اجتناب کرد.

بالا بودن قند خون و فشار خون بالا همراه با تری گلیسرید و کلسترول بالا، احتمال بروز سکت قلبی را بسیار افزایش می دهد. با کم کردن کلسترول و کنترل فشار خون بروز سکت قلبی در بیماران دیابتی را می توان کاهش داد.

بحث و نتیجه گیری

در این تحقیق با استفاده از الگوریتم های داده کاوی تلاش کردیم که احتمال بروز سکت قلبی در بیماران دیابتی را با بهره گیری از ویژگی های تشخیص دیابت پیش بینی کنیم. از میان الگوریتم های مورد استفاده، بهترین نتایج از الگوریتم درخت C&R به دست آمد که دقت مدل

جدول ۵. مقایسه نتایج کار با کارهای انجام شده پیشین

نویسندگان و سال ارائه	روش انتخابی	تأثیرگذارترین ویژگی ها	صحت مدل	دقت مدل	ارائه قوانین
سجادی و همکاران (۲۰۰۵)	خوشه بندی	افزایش فشار خون، افزایش شاخص توده بدنی و اختلال چربی های خون	-	-	☒
پارتیبان و همکاران (۲۰۱۱)	شبکه بیزین	سن، جنسیت، فشار خون و میزان قند خون	۷۴ درصد	۷۱ درصد	☒
پارتیبان و همکاران (۲۰۱۲)	ماشین بردار پشتیبانی (svm)	مشخص نشده است	۹۴/۶۰ درصد	۹۴/۰۶ درصد	☒
مدل انتخابی ما	C&R	فشار خون بالا، شاخص BMI، مقدار کنترل نشده برای کلسترول مضر، بالا بودن سن و نداشتن تحرک زیاد	۹۴ درصد	۹۱ درصد	✓

توده بدنی در هر سه مورد، مطالعاتی انجام شده، به عنوان عواملی با بیشترین تأثیرگذاری شناخته شده اند. ویژگی هایی مانند افزایش شاخص توده بدنی و فشار خون بالا را می توان با رژیم های غذایی و انجام حرکات

با مرور کارهای پیشین در این زمینه، مؤلفه های شناخته شده تأثیرگذار بر سکت قلبی تحقیق های انجام شده با ویژگی های یافته شده توسط کار ما منطبق هستند. ویژگی های سن، فشار خون بالا، افزایش شاخص

سپاس و قدردانی

این مطالعه با همکاری پزشکان و کارکنان محترم مرکز دیابت گرگان، زیر نظر دانشگاه علوم پزشکی استان گلستان انجام شده است، محققان بر خود لازم می دانند تا از مساعدت و همکاری یکایک این عزیزان تشکر و قدردانی کنند.

منابع

1. Neisani TR. Micronutrients are effective in reducing blood pressure in type II diabetic. Iranina's Students New Agency. [cited 2009 sep 24]; Available from (url) : [http://isna.ir/fa/news/8907-13779.166745]
2. World Health Organization(WHO). International Diabetes. The World health report: 2002: Reducing the risks, promoting healthy life. Switzerland: FederationWorld Health Organization; 2002.
3. Al Jarullah, Asma A. Decision tree discovery for the diagnosis of type II diabetes. Innovations in Information Technology (IIT), 2011 International Conference on IEEE. 2011; 303-307.
4. Khajehei M, Etemady F. Data mining and medical research studies.cimsim, Second International Conference on Computational Intelligence, Modelling and Simulation 2010; 119-122.
5. Jayalakshmi T, Santhakumaran A. A Novel classification method for diagnosis of diabetes mellitus using Aartificial neural networks. Data Storage and Data Engineering (DSDE)2010 International Conference on IEEE. 2010; 159-163.
6. Ameri H. Using data mining in diabetes, Master of Science Seminar in Information Technology (Ecommerce), K. N. Toosi University of Technology 2013[Persian]
7. Zayeri F, Sadeghi Nejad R, Noorkojuri H, Bagheri J, Ghazanfari E. Application of classification tree model for determining the effective factors of mortality after coronary bypass surgery in dialysis-independent patients. Daneshvar(medicine) Shahed University. 2012; 98:15-24[Persian]
8. Patil B, Durga T. Association rule for classification of type-2 diabetic patients. In Machine Learning and Computing (ICMLC), 2010 Second International Conference on IEEE. 2010; 330-334.
9. Bagdi R,Patil P. Diagnosis of diabetes using OLAP and data mining integration. International Journal of Computer Science & Communication Networks, 2012; 3: 314-322.
10. Vimala B, Shakouri M, Hoodeh H, Loo, Huck-Soo. Predictions using data mining and case-based reasoning: A case study for retinopathy. International Conference on Information Technology.2012; 63:573-576.
11. Barakat N, Bradley AP. Intelligible support vector machines for diagnosis of diabetes mellitus. Information Technology in Biomedicine, IEEE .2010; 14:1114-1120.
12. Purnami SW, Zain JM, Embong A. A new expert system for diabetes disease diagnosis using modified spline smooth support vector machine. Computational Science and Its Applications-ICCSA 2010. Springer Berlin Heidelberg. 2010; 6019:83-92.
13. Cho BH, Yu H, Kim KW, Kim TH, Kim IY, Kim SI. Application of irregular and unbalanced data to predict diabetic nephropathy using visualization and feature selection methods. Artificial Intelligence in Medicine 2008; 42:37-53.
14. Sajjadi F, Mohammadifard N,Ghaderian N, Alikhasi H,Maghroon M. Clustering of cardiovascular risk factors in diabetic and glucose intolerant cases. The Journal of Qazvin University of Medical Sciences. 2005; 9:35-43.
15. Parthiban G, Rajesh A, Srivatsa SK. Diagnosis of heart disease for diabetic patients using naive bayes method. International Journal of Computer Applications ,2011; 24:0975-8887.
16. Parthiban G, Srivatsa SK. Applying machine learning methods in diagnosing heart disease for diabetic patients. International Journal of Applied Information Systems (IJAIS). 2012; 3:2249-0868.
17. Mehrabi Y, Sedehi M, Kazemnejad A, Joharimajd V, Hadaegh F. Artificial neural network for joint prediction of metabolic syndrome and HOMA-IR.Tehran Lipid and Glucose Study (TLGS) . 2010; 17:29-38[Persian]
18. Han J, Kamber M. chapter 1: Introduction: Data Mining: Concepts and Techniques. 2nd ed., San Francisco: Morgan Kaufman Publisher. 2006.

19. Ahmadi K. Guideline & book review. The internal (endocrine and lung). Tehran: Ahmadi Cultural Institute. 2009 [Persian].
20. Franciosi M, De Berardis G, Rossi MC, Sacco M, Belfiglio M, Pellegrini F and et al. Use of the diabetes risk score for opportunistic screening of undiagnosed diabetes and impaired glucose tolerance the IGLOO (Impaired Glucose Tolerance and Long-Term Outcomes Observational) study. Diabetes Care. 2005; 28:1187-1194.
21. Newman DS, Hettich J, Blake CLS, Merz CJ. UCI Repository of machine learning databases, Irvine, CA: University of California. Department of Information and Computer Science. 1998.
22. Chen G, Astebro T. How to deal with missing categorical data: test of a simple Bayesian method. Organizational Research Methods. 2003; 6: 309-327.
23. Alizadeh S, Ghazanfari M, Teimorpour B. Data Mining and Knowledge Discovery, Tehran: Publication of Iran University of Science and Technology. 2nd ed. 2011 [Persian].
24. Alizadeh S, Malekmahmudi S. Data Mining and Knowledge Discovery, step by step, Tehran: publication of KNTU university. 2011 [Persian].

Daneshvar
Medicine

*Scientific-Research
Journal of Shahed
University
21st Year, No.112
September- October,
2014*

Received: 11/06/2014

Last revised: 09/08/2014

Accepted: 13/08/2014

Identification of influencing factors for heart attack in diabetic patients using C & R algorithm

Hakimeh Ameri^{1*}, Somayeh Alizadeh¹, Akbar Barzegari²

1. Department of Industrial Engineering, KN Toosi University Of Technology, Tehran, Iran

2. Golestan University of Medical Sciences, Gorgan, Iran.

* E-mail: hameri@mail.kntu.ac.ir

Abstract

Background and Objective: Cardiovascular disease is the most common cause of death in developed countries and in the whole world, and according to the World Health Organization prediction, will be the major cause of morbidity throughout the world in 2020. According to the recent World Health Organization report from each 20 deaths, one is due to diabetes. Heart disease and heart attack are the most important complications of diabetes. In this study, data mining algorithms were used to predict the risk of heart attack in diabetic patients with acceptable accuracy and identify the factors that affect the incidence of heart attack.

Materials and Methods: The study was performed retrospectively on 856 patients in 2009 from Gorgan diabetic center. Clinical data of patients using data mining methods were analyzed in the SPSS software. To identify the influencing factors on incidence heart attack, classification data mining algorithms were used.

Results: A model with 94 percent accuracy is identified using the C&R decision tree algorithm. According to the C&R Tree hypertension, index BMI, systolic and diastolic blood pressure, LDL, daily activity level and age are identified as the most important factors of heart disease in diabetic patients

Conclusion: With the use of Created rules and identifying effective features and controlling effective factors on diabetic patients, the mortality rate of this complication was somewhat reduced.

Keywords: Data mining, Diabetes, Heart attack, Decision tree algorithm