مجله دیابت و متابولیسم ایران. بهمن و اسفند ۱۴۰۱؛ دوره ۲۲ (شماره ۶) ۴۲۰-۴۰۹

مقاله پژوهشی

مُدل سازی و شناسایی عوامل مؤثر بر وجود عوارض ناشی از دیابت با روش‌های داده‌کاوی

مصطفی بسکابادی[۱]، نجمه مهاجری[۲]، علی تقی پور[۳]، حبیب الله اسماعیلی[۴]، سید جواد حسینی[۲]، احسان موسی فرخانی[۳*]

چکیده

مقدمه: در ایران با پیشرفت فناوری و توسعه‌ی آمارهای ثبتی لزوم استفاده از روش‌های داده کاوی بیشتر مورد توجه محققین قرار گرفته است. درخت رگرسیون و طبقه‌بندی یکی از روش‌های مهم در مُدل‌بندی داده‌های حجیم است که برای کنترل جامعه و پیش‌بینی مورد توجه محققین زیادی قرار گرفته است. هدف این مطالعه تعیین متغیرهای تأثیرگذار بر فراوانی رخداد عوارض ناشی از دیابت است.

روش‌ها: این پژوهش از نوع مقطعی-تحلیلی است. در این پژوهش، اطلاعات تمام افراد مراجعه کننده‌ی دیابتی تحت پوشش دانشگاه علوم پزشکی مشهد در سال ۱۳۹۷ از سامانه‌ی سینا استخراج گردید. ۵۰۱۶ نفر از افراد وارد شده به مطالعه دارای عارضه‌ی دیابت و ۵۳۶۱۳ نفر نیز بدون عارضه بودند. روش برازش مدل درخت رگرسیون و طبقه‌بندی و معیار سنجش مدل ضریب تعیین و مساحت منحنی راک و نمودار Lift است.

یافته‌ها: منحنی راک برای مدل درختی برازش داده شده ۷۳/۸ درصد که نشان دهنده‌ی توان نسبتاً بالای مدل است. براساس نمودار Lift قدرت تصمیم‌گیری بروز عارضه‌ی دیابت برای فردی که مراجعه می‌کند ۳/۵ برابر افزایش می‌یابد.

نتیجه‌گیری: نتایج مدل رگرسیون و طبقه‌بندی درختی نشان داد که از متغیرهای کمی به‌ترتیب نزولی سن، عامل خطرسنجی، FBS، HbA1C، مجموع زمان فعالیت، کلسترول، FBS وHDL، بیماری قلبی و عروقی، سابقه‌ی سکته، فشار خون، کلسترول، تجویز استاتین، شغل با فعالیت فیزیکی سخت، منطقه‌ی زندگی، روغن مصرفی، پیاده‌روی، مصرف سبزی‌ها و جنسیت در فراوانی رخداد عارضه‌ی دیابت مؤثرتر از عوامل دیگر هستند.

واژگان کلیدی: درخت رگرسیون و طبقه‌بندی، عوارض دیابت، منحنی راک

۱- دانشکده‌ی بهداشت، دانشگاه علوم پزشکی مشهد، مشهد، ایران

۲- معاونت بهداشت، دانشگاه علوم پزشکی مشهد، مشهد، ایران

۳- گروه اپیدمیولوژی، دانشکده‌ی بهداشت، دانشگاه علوم پزشکی مشهد، مشهد، ایران

۴- گروه آمار زیستی، دانشکده‌ی بهداشت، دانشگاه علوم پزشکی مشهد، مشهد، ایران

***نشانی:** مشهد، میدان دانشگاه، دانشگاه علوم پزشکی مشهد، کُد پستی: ۹۱۳۷۶۷۳۱۱۹، نمابر: ۰۵۱۳۸۵۲۲۷۷۵، تلفن: ۰۵۱۳۸۵۱۴۵۴۸، پست الکترونیک: farkhanie@mums.ac.ir

تاریخ دریافت: ۱۴۰۱/۰۵/۰۱ **تاریخ درخواست اصلاح:** ۱۴۰۱/۰۸/۰۸ **تاریخ پذیرش:** ۱۴۰۱/۱۰/۱۰

مقدمه

دیابت یا بیماری قند یک اختلال سوخت‌وسازی (متابولیک) در بدن و یکی از بیماری‌های شایع در جهان است که تاکنون راه درمان قطعی برای آن یافت نگردیده است. در این بیماری توانایی تولید هورمون انسولین در بدن از بین می‌رود یا بدن در برابر انسولین مقاوم شده و بنابراین انسولین تولیدی نمی‌تواند عملکرد طبیعی خود را انجام دهد. این بیماری شایع‌ترین علت قطع اندام، نابینایی و نارسایی کلیوی و از عوامل خطر در ایجاد بیماری‌های قلبی است.

دیابت در بزرگسالان یک معضل سلامتی در جهان است. شیوع دیابت در جهان به‌طور نگران کننده‌ای در حال افزایش است. تعداد افراد مبتلا به دیابت از ۱۷۱ میلیون نفر به ۳۶۶ میلیون نفر در بازه‌ی زمانی سال ۲۰۰۰ تا ۲۰۳۰ تخمین زده شده است [۱]. از ۱۰ کشور با بالاترین نرخ دیابت نوع دو، پنج کشور در قاره‌ی آسیا واقع شده‌اند [۲]. هم‌اکنون بیش از سه میلیون نفر در ایران مبتلا به دیابت هستند که براساس برآورد سازمان جهانی بهداشت، چنانچه اقدامات مؤثری صورت نپذیرد، این تعداد تا سال ۲۰۳۰ به نزدیک ۷ میلیون نفر خواهد رسید [۳]. با شیوع بیماری دیابت، افراد مبتلا در معرض خطر پیشرفت عوارض بیماری هستند. میزان مرگ‌ومیر در افراد دیابتی ۱/۵–۲/۵ درصد بالاتر از جمعیت عمومی است. با توجه به ماهیت مزمن، غیرواگیر و پُرهزینه‌ی این بیماری برای بهداشت عمومی، بار مالی فراوانی را بر فرد، خانواده، جامعه و کشور وارد می‌کند [۴، ۵]. طول عمر افراد دیابتی به‌دلیل عوارض آن حدود ۱۰ سال کمتر از جمعیت عمومی است که بخشی از این سال‌ها به‌علت مرگ زودرس ناشی ازعوارض بیماری و بخشی به‌دلیل زندگی توأم با ناتوانی ناشی از دیابت است. بیش از ۸۰٪ هزینه‌های دیابت مربوط به بستری شدن به‌ویژه بستری به‌دلیل عوارض مزمن آن است [۶]. بار ناشی از دیابت و به‌عبارت دیگر تعداد سال‌های مفید که دیابت و عوارض آن از فرد و جامعه به هدر می‌رود، بسیار قابل توجه است [۷، ۸]. برآوردها نشان می‌دهند با رشد جمعیت از دو جنبه‌ی افزایش در تعداد و مسن‌تر شدن، تا سال ۲۰۲۰ تأثیر هزینه‌ی مستقیم این بیماری در آمریکا به ۱۳۸ بیلیون دلار می‌رسد [۹]. بنابراین دیابت به‌دلیل عوارض عدیده‌ی آن، یک بیماری پُرهزینه برای فرد، خانواده و جامعه قلمداد می‌شود [۱۰]. در ایران هزینه‌های پزشکی ناشی از نابینایی به‌علت دیابت تقریباً ۲۰۰۰ دلار در سال، نارسایی کلیوی ۴۵۰۰۰ و قطع عضو ۲۹۵۰۰ دلار در سال محاسبه شده است [۱۱]. در ایران سالیانه ۲۲۶/۲۸۲/۹۶۲/۵۰۰ ریال صرف هزینه‌های مستقیم دیابت می‌شود که این رقم براساس تعرفه‌های وزارت بهداشت و درمان است [۱۲]. سیستم‌های بهداشتی- درمانی و منابع آنها مسئول ارائه‌ی خدمات مناسب به بیماری‌های مزمن از جمله دیابت هستند [۱۳، ۱۴].

در سال‌های اخیر با توجه به دسترسی گسترده به مقادیر بسیار عظیمی از داده‌ها و نیاز قریب‌الوقوع برای تبدیل داده به اطلاعات مفید و دانش، روش‌های داده‌کاوی توجه زیادی را در علوم مختلف از جمله علوم پزشکی به خود جلب نموده‌اند. داده‌کاوی عبارت است از «کشف روش‌ها و الگوهای ویژه در پایگاه داده‌های بزرگ، برای هدایت تصمیم‌گیری در مورد فعالیت‌های آینده». در ایران نیز با توسعه‌ی آمارهای ثبتی، لزوم استفاده از روش‌های داده‌کاوی بیشتر مورد توجه محققین قرار گرفته است. در دانشگاه‌های علوم پزشکی کشور، سامانه‌ها و فرابرهای مختلف آماری (از جمله سامانه‌های سیب وزارت بهداشت و سینا در دانشگاه علوم پزشکی مشهد) به جمع‌آوری حجم زیادی از اطلاعات پرداخته و نیاز روز افزون برای تبدیل اطلاعات حجیم به دانش احساس می‌شود.

داده‌کاوی یکی از روش‌هایی است که می‌تواند ارتباطات و وابستگی‌های جدید و بدیعی را کشف کند که برای پزشکان مفید هستند [۱۵]. داده کاوی در حوزه سلامت کاربردهای فراوانی دارد که از جمله‌ی آنها میتوان به موارد ذیل اشاره کرد: تشخیص بیماری‌ها، دسته‌بندی بیماران در مدیریت بیماری، پیدا کردن الگوهایی برای تشخیص سریع‌تر بیماران و جلوگیری از بروز عوارض در آنها [۱۶، ۱۷]. در این تحقیق نیز با توجه به اهمیت عوارض بیماری دیابت و هزینه‌هایی که بر مردم و حوزه‌ی سلامت کشور وارد می‌کند این بیماری را از

این منظر مورد بررسی قرار خواهیم داد، بنابراین پیاده‌سازی روشی که بتواند امکان تشخیص صحیح و مؤثر عوامل خطر را در ابتلا به دیابت مشخص کند، می‌تواند گام مهمی در پیشگری و کنترل عوارض ناشی این بیماری باشد. هدف از این مطالعه بررسی عوامل مؤثر در فراوانی رخداد عوارض ناشی از بیماری دیابت است که از تکنیک‌های داده‌کاوی برای پیش‌بینی و مُدل‌بندی استفاده می‌شود.

روش‌ها

این پژوهش از نوع مقطعی-تحلیلی است. در این تحقیق، تمام افراد مراجعه کننده در سال ۱۳۹۷ به مراکز خدمات جامع سلامت که در سامانه‌ی پرونده‌ی الکترونیک سلامت (سینا) ثبت و جهت ایشان تشخیص دیابت براساس کُدهای ICD-10 درج گردیده است، استخراج گردید. بیماران ثبت شده براساس سیستم کُددهی بین‌المللی ICD-10 با کُد اختصاصی E11.1 تا E11.8 به‌عنوان فراد مبتلا به دیابت همراه با عارضه و افراد با کُد اختصاصی E11 و E10 به‌عنوان افراد مبتلا به دیابت بدون عارضه در نظر گرفته شدند.

تعداد افراد دیابتی دارای عوارض به تعداد ۵۰۱۶ نفر و افراد دیابتی که عوارض بیماری برای آنها ثبت نشده بود ۵۳۶۱۳ نفر بودند. داده‌های مربوط به میزان فعالیت بدنی، سابقه‌ی ابتلا به بیماری، سبک زندگی از اطلاعات ثبت شده براساس پروتکل مراقبتی وزارت بهداشت درمان و آموزش پزشکی ابلاغی جمع‌آوری گردیده است. برای این مجموعه داده با توجه به حجم بالای داده‌ها و تعداد داده‌های تکمیل نشده (Missing value) روش درخت رگرسیون و طبقه‌بندی Classification and Regression Trees (CART) جهت برازش مدل مناسب پیشنهاد می‌شود. روش CART روشی مهم در داده‌کاوی بوده که علاوه بر پیش‌بینی و مُدل‌بندی براساس پارتیشن‌هایی از عوامل خطر، اثر داده‌های تکمیل نشده را در مدل لحاظ می‌کند. این روش برای پیش‌بینی و دسته‌بندی بیماران نتایج کاربردی و ارزشمندی در علوم پزشکی دارد [۱۸، ۱۹]. تجزیه‌وتحلیل داده‌ها در این تحقیق با استفاده از نرم‌افزار R نسخه‌ی ۲-۱-۳ و JMP نسخه‌ی ۱۳ انجام شده است.

درخت رگرسیون و طبقه‌بندی

درخت تصمیم یکی از روش‌های ناپارمتری طبقه‌بندی کردن است که در داده‌کاوی بسیار مورد استفاده قرار می‌گیرد. با توجه به نوع متغیر به دو دسته‌ی طبقه‌بندی درختی برای متغیر گسسته و رگرسیون درختی برای متغیر پیوسته تقسیم می‌شود. مدل درختی CART نامی است که به هر دو روال بالا اطلاق می‌شود. نام CART سرنام کلمات درختان رگرسیون و طبقه‌بندی است. این روش در سال ۱۹۸۴ توسط بریمن [۲۰] معرفی شده است. برای توضیحات تکمیلی از نوع ساخت درخت به منبع شماره [۲۰] و برای انواع روش‌های دیگر درختی به منبع شماره [۲۱] مراجعه شود.

تعاریف و مفاهیم معیارهای استفاده شده

در این بخش به چند معیاری که در ادامه‌ی تحقیق مورد استفاده و ارزیابی قرار گرفته‌اند اشاره خواهیم کرد و مفاهیم آن‌ها را توضیح می‌دهیم:

۱- اگر X مجموعه‌ای از داده‌های مشاهده شده و Θ مجموعه‌ای از پارامترها باشد، آنگاه $L(X|\Theta)$ احتمال درستی Θ براساس X دانست. بزرگ بودن این تابع احتمال (Likelihood) نسبت به Θ نشان دهنده‌ی بهتر بودن مدل است. گاهی برای سهولت در محاسبات از لگاریتم این تابع استفاده می‌شود. لگاریتم تابع احتمال (Loglik-) را یک معیار برای سنجش نیکویی برازش مدل مورد استفاده قرار می‌دهیم.

۲- معیار اطلاعاتی آکائیکه (Akaike information criterion) یا به‌طور مخفف AIC معیاری برای سنجش نیکویی برازش مدل است. این معیار نشان می‌دهد که استفاده از یک مدل آماری به چه میزان باعث از دست رفتن اطلاعات می‌شود. به عبارت دیگر، این معیار تعادلی میان دقت مدل و پیچیدگی آن برقرار می‌کند. با توجه به داده‌ها، چند مدل رقیب ممکن است با توجه به مقدار AIC رتبه‌بندی شوند و مدل دارای کمترین AIC بهترین است.

۳- یک مدل رگرسیونی زمانی به‌عنوان یک مدل خوب برای پیش‌بینی در نظر گرفته می‌شود که قدرت توضیح دهندگی آن، که توسط ضریب تعیین (R2) اندازه‌گیری می‌شود، حتی الامکان بالا باشد. ضریب تعیین تغییرات متغیر وابسته به‌وسیله ی متغیرهای توضیحی در مدل را می‌سنجد. لذا این معیار شاخصی است که نشان می‌دهد تا چه اندازه معادله‌ی رگرسیونی داده‌ها را به نیکویی برازش می‌کند.

یافته‌ها

در این بخش جزییات همه‌ی متغیرهای مختلف را به‌صورت گزارش توصیفی بررسی می‌نماییم. در جمعیت مورد مطالعه ۶۸ درصد جمعیت را زنان و ۳۲ درصد جمعیت را مردان تشکیل داده‌اند. فراوانی خیلی بیشتر زنان در این جمعیت را نمی‌توان فراوانی واقعی در جامعه‌ی بیماران دیابتی دانشگاه علوم پزشکی مشهد تفسیر کرد، بلکه به این دلیل می‌توان دانست که مراجعین در مراکز و پایگاه‌های بهداشتی بیشتر بانوان بوده‌اند. میانگین و انحراف معیار سن افراد مراجعه کننده به‌ترتیب برابر ۵۷/۵ سال و ۱۲/۲ است.

جمعیت شهری مراجعین دیابتی ۳۸ درصد و روستایی ۲۱ درصد و جمعیت حاشیه شهر نیز ۴۱ درصد هستند. در این میان ۶۵ درصد با سواد و ۳۵ درصد بی‌سواد هستند و نسبت شغلی آنها ۶۶ درصد بیکار، ۱۲ درصد شغل آزاد و ۲۲ درصد استخدام است. می‌توان دلیل درصد بالای بی‌سواد و بیکار را در این جمعیت به‌دلیل این دانست که تعداد قابل توجهی از این مراجعین در روستاها و با میانگین سنی بالا هستند. افرادی که بیان کردند مصرف مواد افیونی غیرقانونی داشتند ۲/۶ درصد گزارش شده است. افرادی که حد BMI بالایی دارند ۷۵ درصد و BMI نرمال ۲۲ درصد و پایین نیز ۰/۷ درصد و حدود ۲/۱ درصد هم به این مورد پاسخ ندادند. دیگر خصوصیات افراد مراجعه کننده به‌صورت دسته‌بندی زیر توصیف می‌شوند:

۱- نوع تغذیه‌ی افراد مراجعه کنند به این صورت توزیع شده است:

متغیر تعداد واحد مصرفی معمول روزانه‌ی میوه‌ها ۱۳/۶ درصد دو سهم و بیشتر گزارش کرده و ۵/۶ درصد کمتر از دو سهم و ۰/۴ درصد به‌ندرت بوده‌اند، حدود ۸۰ درصد هم به این مورد پاسخ ندادند. متغیر تعداد واحد مصرفی معمول روزانه سبزی ها، سه سهم و بیشتر ۱۰/۲ درصد و کمتر از سه سهم ۸/۷ درصد و به‌ندرت ۰/۶ درصد و تکمیل نشده نیز ۸۰ درصد بودند. تعداد واحد مصرفی معمول روزانه‌ی لبنیات نیز ۱۳/۸ درصد دو سهم و بیشتر گزارش کرده و ۵/۵ درصد کمتر از دو سهم و ۳/۱ درصد به‌ندرت و تکمیل نشده نیز ۸۰ درصد بوده‌اند. درصد افرادی که از نمکدان با وعده‌های غذایی استفاده می‌کردند ۲/۱ و ۵۸ درصد تکمیل نشده بودند. مصرف معمول فست‌فود و نوشابه‌های گازدار گزارش شده به‌صورت ۳/۲ درصد بیشتر از هفته‌ای دوبار و ۶/۱ درصد یک یا دوبار در ماه و ۴۲ درصد استفاده به‌ندرت داشتند، حدود ۴۸ درصد هم به این مورد پاسخ ندادند. نوع روغن مصرفی مراجعه کنندگان دیابتی شامل ۱۰/۳ درصد روغن جامد، ۲۵/۷ درصد روغن مایع و ۱۶/۳ درصد تلفیق دو نوع روغن گزارش شده بود حدود ۴۸ درصد هم به این مورد پاسخ ندادند.

۲- درمورد فعالیت‌های فیزیکی مراجعین توزیعی به این صورت گزارش شده است:

افرادی که فعالیت بدنی مطلوب داشتند ۳۱ درصد در مقابل ۳۵ درصد فعالیت بدنی غیرمطلوب در مراجعین دیابتی که حدود ۳۴ درصد تکمیل نشده بودند. درصد مراجعین که پیاده‌روی را در طی روز انجام می‌دادند ۵۷ در مقابل ۳۶/۴ درصد بدون پیاده‌روی بودند و حدود ۶/۶ درصد تکمیل نشده بودند. انجام فعالیت شغلی سخت در بین داده‌ها ۱۹/۳ درصد در مقابل ۷۳/۱ درصد و حدود ۷/۶ درصد تکمیل نشده بودند. انجام تمرین‌های تفریحی ورزشی ۲/۴ درصد در مقابل ۹۰ درصد کسانی که تمرین‌های ورزشی مداوم نداشتند و حدود ۷/۶ درصد تکمیل نشده بود.

۳- سوابق و نوع بیماری مراجعین به این صورت توزیع شده‌اند:

متغیر سابقه‌ی سکته‌ی قلبی و مغزی حدود ۹۰ درصد تکمیل نشده داشت مابقی آن ۶/۷ درصد دارای سابقه‌ی سکته و ۳/۳ بدون سابقه بودند. ۱۰/۵ درصد از بیماران دارای سابقه‌ی بیماری قلبی و عروقی و ۲۶/۵ درصد فاقد سابقه هستند و در حدود ۶۳

درصد افراد این متغیر فاقد اطلاعات بوده است. ۰/۲ درصد از بیماران دارای سابقه‌ی بیماری کلیوی در بستگان و ۵/۵ درصد فاقد سابقه بودند و این اطلاعات در حدود ۹۴ درصد افراد وجود نداشت. کلسترول بالا در ۲۹ درصد بیماران و فشار خون در ۵۴ درصد از موارد گزارش شده است.

۴- میانگین و انحراف معیار نتایج آزمایشگاهی بیماران دیابتی در جدول ۱ آمده است.

جدول ۱- میانگین و انحراف معیار نتایج آزمایشگاهی

متغیر	تری‌گلیسرید	LDL	HDL	FBS	کلسترول	HbA1	گلوکز
میانگین	۱۸۴/۱۹	۱۰۶/۳	۴۵/۹۴	۱۵۴/۲۴	۱۸۴/۶۳	۷/۵۸	۲۲۶/۸۵
انحراف معیار	۷۸/۱۵	۳۰/۷۹	۱۰/۲۵	۵۲/۸۴	۳۸/۶۱	۱/۸۹	۸۳/۵۲

در این تحقیق، متغیر پاسخ افرادی هستند که حداقل یکی از عوارض ناشی از بیماری دیابت را دارند شامل عوراض قلبی، کلیوی، چشمی، عصبی و مغزی. افراد دیابتی دارای عوارض ۵۰۱۶ نفر و افراد دیابتی فاقد عوارض ۵۳۶۱۳ نفر بودند. فروانی رخداد عوارض ذکر شده در جدول ۲ مشاهده می‌شود.

جدول ۲- عوارض مختلف دیابت با بالاترین فراوانی رخداد

عوارض بیماری	مغزی و قلبی	عصبی و قلبی و چشمی	عصبی و قلبی	قلبی و چشمی و کلیوی	قلبی و کلیوی	چشمی و کلیوی	عصبی و چشمی	قلبی و چشمی	مغزی	کلیوی	عصبی	چشمی	قلبی
تعداد	۴۲	۴۳	۶۵	۶۸	۷۶	۱۰۷	۱۰۹	۳۴۹	۱۲۵	۲۶۲	۶۳۸	۱۰۹۵	۲۱۷۲
درصد	۰/۸۳۷	۰/۸۵۷	۱/۲۹۶	۱/۳۵۶	۱/۵۱۵	۲/۱۳۳	۲/۱۷۳	۶/۹۵۸	۲/۴۹۲	۵/۲۲۳	۷/۳۳۷	۲۱/۸۳	۴۳/۳۰۱

باید در این تحلیل در نظر گرفت به‌دلیل حجم بالای داده‌های مورد تحلیل و اینکه از روش‌های محدود تعیین حجم نمونه استفاده نمی‌شود، نمی‌توان از روش‌های کلاسیک آماری استفاده کرد و خطای ۵ درصد را مبنای تصمیم‌گیری قرار داد.

بحث

انتخاب متغیر تأثیرگذار بر مدل

از نتایج مقایسه‌ی دوبه‌دوی هر متغیر کیفی (دسته‌ای) با متغیر عارضه‌ی دیابت و برازش مدل تک متغیره‌ی درخت طبقه‌بندی با روش جداول توافقی برای انجام تأثیر هر متغیر بر فراوانی رخداد عارضه‌ی دیابت جدول ۳ استخراج شده که در آن نشان می‌دهد چه اندازه هر متغیر تغییرات متغیر پاسخ (عارضه‌ی دیابت) را بیان می‌کند. سپس با توجه به دو معیار ضریب تعیین R2 وLoglik می‌توان به‌صورت زیر تفسیر کرد:

با توجه به جدول ۳ برای متغیرهای کیفی مشاهده می‌شود عامل داشتن بیماری قلبی‌وعروقی با توجه به بیشترین ضریب تعیین ۰/۰۸۱ بیشترین تأثیر را بر احتمال فراوانی رخداد عارضه‌ی دیابت بر روی یک بیمار دیابتی دارد. سابقه‌ی سکته متغیر دوم از نظر بیان تغییرپذیری عارضه‌ی دیابت، با ضریب تعیین برابر ۰/۰۵۶ بوده و اهمیت زیادی بر روی متغیر هدف ما یعنی فراوانی رخداد عارضه‌ی دیابت دارد. بر همین مبنا متغیرهای فشارخون، کلسترول، تجویز استاتین، شغل با فعالیت فیزیکی سخت، منطقه‌ی زندگی، روغن مصرفی، پیاده‌روی، مصرف سبزیجها و جنسیت به ترتیب نزولی اهمیت بر فراوانی رخداد عارضه‌ی دیابت دارد.

در این مجموعه داده‌ها متغیرهای با ضریب تعیین (R2) کمتر از ۰/۰۰۱ به‌دلیل تأثیر خیلی کم بر روی متغیر هدف، وارد مدل نهایی نخواهیم شد. متغیر نارسایی کلیوی در بستگان نزدیک با ضریب تعیین برابر ۰/۰۰۱ تأثیر قابل توجهی بر عارضه‌ی دیابت ندارد. بعد از آن به‌ترتیب نزولی متغیرهای شغل، مصرف مواد افیونی، تحصیلات، مصرف لبنیات، مصرف معمول فست‌فود، مصرف میوه، ورزش و وجود نمکدان در وعده‌های غذایی با توجه به دو معیار ضریب تعیین و Loglik تحلیل بر عدم تأثیرگذاری این بر فراوانی رخداد عارضه‌ی دیابت دانست.

جدول ۳- تاثیرگذاری متغیرهای کیفی بر فراوانی رخداد عارضه‌ی دیابت

نام متغیر	مصرف سبزی‌ها	پیاده‌روی	روغن مصرفی	منطقه‌ی زندگی	فعالیت بدنی شغل سخت		تجویز استاتین	کلسترول بالا	فشارخون	سابقه‌ی سکته	بیمار قلبی عروقی
LogLike	۶/۸۷	۳۵/۲۳	۲۱/۰۲	۶۱/۵۴	۴۱/۲۸	۱۳۴/۷۱	۶۶/۵۰	۲۱۱/۴۳	۴۴۰/۰۰	۱۵۵/۲۷	۳۹۸/۰۴
ضریب تعیین	۰/۰۰۱۹	۰/۰۰۲۱	۰/۰۰۲۶	۰/۰۰۳۶	۰/۰۰۳۷	۰/۰۰۸۲	۰/۰۱۰۶	۰/۰۱۲۳	۰/۰۲۵۷	۰/۰۵۲۶	۰/۰۸۱۸
نام متغیر	مصرف نمک	ورزش	**BMI**	مصرف میوه	مصرف فست فود	مصرف لبنیات	تحصیلات	مصرف مواد افیونی	شغل	نارسایی کلیه بستگان	جنسیت
LogLike	۰/۳۰	۰/۹۲	۲/۳۱	۰/۶۵	۱/۴۰	۱/۷۰	۷/۹۱	۹/۲۹	۱۳/۰۲	۰/۷۲	۲۱/۴۶

متغیر سن بیشترین تأثیر را بر فراوانی رخداد عارضه‌ی دیابت دارد. با توجه به روش طبقه‌بندی رگرسیون لجستیک در برازش مدل درختی و با توجه به بیشترین ضریب تعیین ۰/۰۳ می‌توان تحلیل بر تأثیرگذار بودن زیاد این متغیر بر بروز عارضه دیابت دانست. سپس مطابق جدول ۴ در مرتبه‌ی دوم متغیر Risk Score با توجه به ضریب تعیین ۰/۰۲۳ تاثیرگذار بر فراوانی رخداد عارضه‌ی دیابت است. بر همین مبنا متغیرهای گلوکز، هموگلوبین (HbA1C)، مجموع زمان فعالیت، کمیت کلسترول، FBS و HDL به‌ترتیب نزولی اهمیت روی عارضه‌ی دیابت دارند.

در این مجموعه داده‌ها نیز مشابه حالت کیفی متغیرهای با ضریب تعیین کمتر از ۰/۰۰۱ را به‌دلیل تأثیر خیلی کم بر روی عارضه دیابت وارد مدل نهایی نخواهیم کرد. بنابرین متغیر LDL و تری‌گلیسرید تأثیر قابل توجهی بر عارضه دیابت ندارد.

جدول ۴: تأثیرگذاری متغیرهای کمی بر فراوانی رخداد عارضه دیابت

نام متغیر	تری‌گلیسرید	LDL	HDL	FBS	کلسترول	ارزیابی روانشناختی	هموگلوبین	گلوکز	Risk Score	سن
ضریب تعیین	۰	۰/۰۰۰۷	۰/۰۰۱۳	۰/۰۰۲۴	۰/۰۰۳۸	۰/۰۰۴۲	۰/۰۰۷۵	۰/۰۱۱۷	۰/۰۲۳۴	۰/۰۳۰۲
AIC	۲۸۱۹۴/۶	۲۴۷۷۹/۶	۲۶۷۱۷/۱	۳۱۵۰۵/۱	۲۹۷۴۱/۹	۱۰۵۱۱	۱۸۵۱۲/۶	۱۲۲۵۸/۷	۲۵۱۴۰/۷	۳۳۲۲۲/۹

برازش مدل

در این بخش به تعیین مدلی مناسب بر مجموعه داده‌های تعیین کننده‌ی فراوانی رخداد عارضه‌ی دیابت خواهیم پرداخت. ابتدا با توجه به نتایج بخش قبل متغیرهای تأثیرگذار بر متغیر عارضه‌ی دیابت را از متغیرهای بدون تأثیر یا کم تأثیر جدا کرده و فقط متغیرهای تأثیرگذار را وارد فرآیند مدل‌سازی خواهیم کرد. در از نتایج مقایسه‌ی دوبه‌دوی هر متغیر کمی با متغیر عارضه‌ی دیابت و برازش مدل تک متغیره‌ی درخت طبقه‌بندی با روش رگرسیون لجستیک برای انجام تأثیر هر متغیر بر بروز عارضه‌ی دیابت جدول ۴ استخراج شده که در آن مشخص می‌شود چه اندازه هر متغیر، تغییرات متغیر پاسخ (عارضهی دیابت) را بیان می‌کند. سپس با توجه به دو معیار ضریب تعیین R2 و آکاییک AIC می‌توان به‌صورت زیر تفسیر کرد:

مرحله‌ی بعد طبق روش‌های متداول داده‌کاوی مجموعه‌ی داده‌ها را به سه دسته به‌صورت ۷۵ درصد داده‌های آموزشی و ۱۵ درصد داده‌های اعتبارسنجی و ۱۰ درصد داده‌های آزمایشی به‌صورت تصادفی تقسیم می‌کنیم. این تقسیم‌بندی به‌دلیل این است که با داده‌های آموزشی مدل مناسب را برازش داده، با داده‌های اعتبارسنجی مدل را مورد ارزیابی قرار می‌دهیم و با داده‌های آزمایشی در آینده تحقیق از روش‌های دیگر اگر مدلی برازش شود با این داده‌ها مورد مقایسه‌ی بین مدل‌ها قرار خواهد گرفت.

با استفاده از روش درخت رگرسیون و طبقه‌بندی (CART)، مدل نهایی درختی به‌صورت شکل ۱ است. با روش بهینه‌سازی مقدار ضریب تعیین مدل، برای مدل رگرسیون درختی ۲۱ بار تقسیم شاخه درخت مناسب است. بنابراین با توجه به روش درختی CART با تقسیمات دوتایی ۴۲ گره در مدل نهایی خواهیم داشت.

تفسیر مدل درختی

در ریشه‌ی درخت اطلاعات فراوانی رخداد عارضه‌ی دیابت بدون هیچ اطلاعی از عوامل دیگر، قرار دارد. بدون هیچ اطلاعی از دیگر متغیرها دیده می‌شود یک بیمار دیابتی ۸/۵ درصد احتمال بروز انواع عارضه‌ی دیابت را دارد و ۹۱/۵ درصد بدون هیچ عارضه‌ای است. اولین تقسیم درخت در این مدل با متغیر سن انجام شده است نشان می‌دهد اگه فرد مراجعه کننده سنی بالای ۵۳ سال داشته باشد (شاخه‌ی ۱) احتمال رخداد دیابت آن تا ۱۰/۸درصد افزایش می‌یابد و در سن زیر ۵۳ سال (شاخه‌ی ۲) احتمال رخداد عارضه‌ی دیابت به ۳/۷ درصد کاهش می‌یابد. از تقسیم دوم می‌توان نتیجه گرفت که فردی که سن زیر ۵۳ سال دارد و بیماری قلبی‌وعروقی ندارد (شاخه‌ی ۳) احتمال رخداد عارضه‌ی دیابت آن به ۳/۰۲ درشد کاهش یافته و اگر بیماری قلبی‌وعروقی داشته باشد (شاخه‌ی ۴) احتمال رخداد عارضه‌ی دیابت آن به ۱۸/۳ درصد افزایش می‌یابد. از تقسیم سوم می‌توان نتیجه گرفت که فردی که سن بالای ۵۳ سال دارد و سابقه‌ی سکته‌ی قلبی ندارد (شاخه‌ی ۵) احتمال رخداد عارضه‌ی دیابت آن به ۹/۱۳ درشد کاهش یافته و اگر سابقه‌ی سکته‌ی قلبی داشته باشد (شاخه‌ی ۶) احتمال رخداد عارضه‌ی دیابت آن به ۲۷/۷ درصد افزایش می‌یابد.

به همین ترتیب همه‌ی شاخه‌بندی درخت را می‌توان تفسیر کرد تا به شاخه‌های نهایی یا برگ‌های درخت رسید. به‌عنوان نمونه شاخه‌ی نهایی (۴۲، ۴۱، ۲۲، ۲۱، ۱۲، ۱۱) را به‌صورت زیر تفسیر می‌کنیم:

۱- شاخه‌ی نهایی ۱۱: احتمال رخداد عارضه‌ی دیابت برای بیمار دیابتی که سن بالای ۵۳ سال و سابقه‌ی سکته‌ی قلبی دارد و کلسترول بالا هم دارد برابر ۳۴/۲ درصد است.

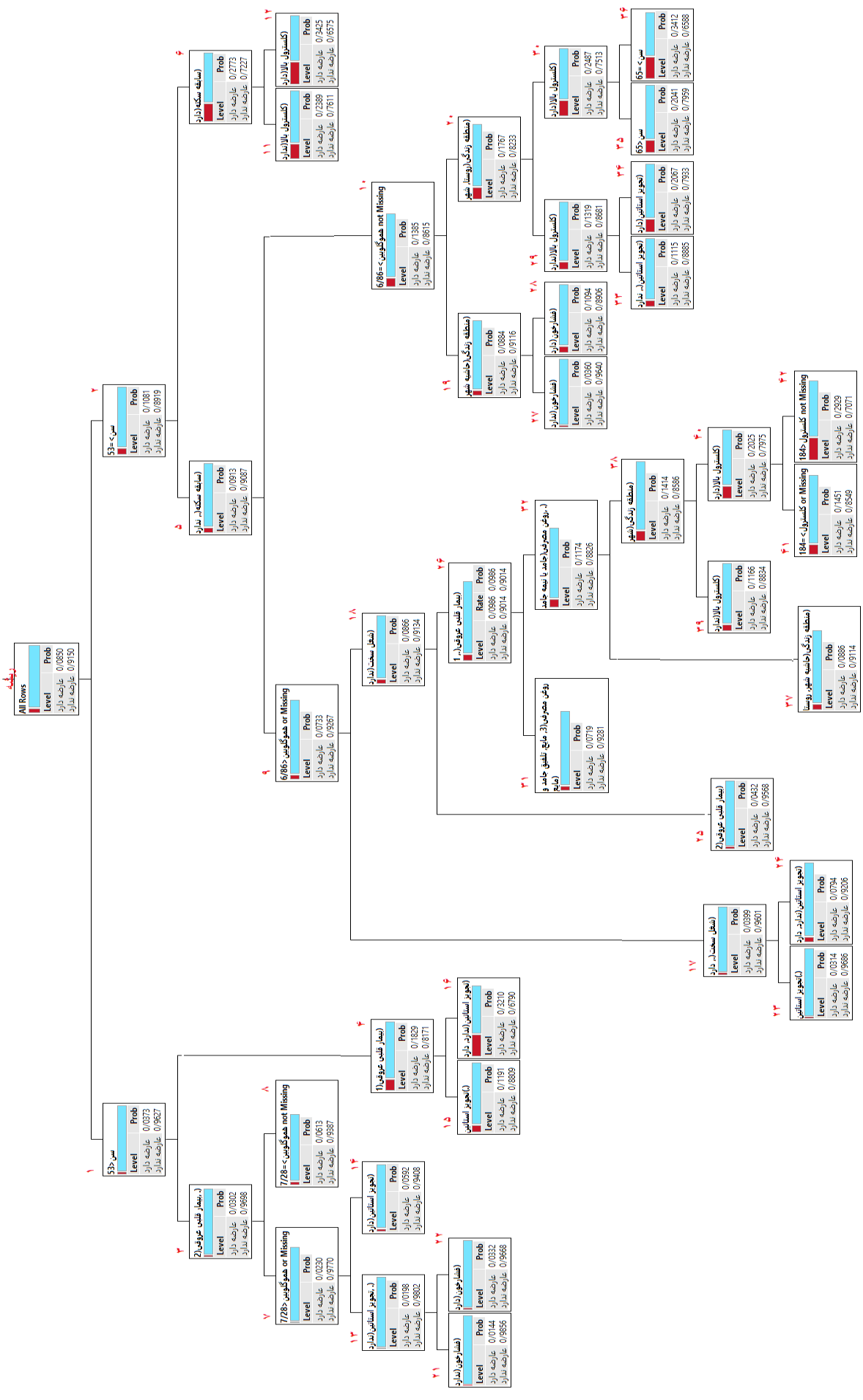
۲- شاخه‌ی نهایی ۲۲: احتمال رخداد عارضه‌ی دیابت برای بیمار دیابتی که سن بالای ۵۳ سال و سابقه‌ی سکته‌ی قلبی دارد و کلسترول بالا ندارد برابر ۲۸/۹ درصد است.

۳- شاخه‌ی نهایی ۲۱: احتمال رخداد عارضه‌ی دیابت برای بیمار دیابتی که سن زیر ۵۳ سال و بیماری قلبی‌وعروقی ندارد و هموگلوبین زیر ۷/۲۸ دارد و تجویز استاتین نداشته است و فشارخون بالا هم ندارد برابر ۱/۴ درصد است.

۴- شاخه‌ی نهایی ۲۲: احتمال رخداد عارضه‌ی دیابت برای بیمار دیابتی که سن زیر ۵۳ سال و بیماری قلبی‌وعروقی ندارد و هموگلوبین زیر ۷/۲۸ دارد و تجویز استاتین نداشته است و فشارخون بالا هم ندارد برابر ۳/۳ درصد است.

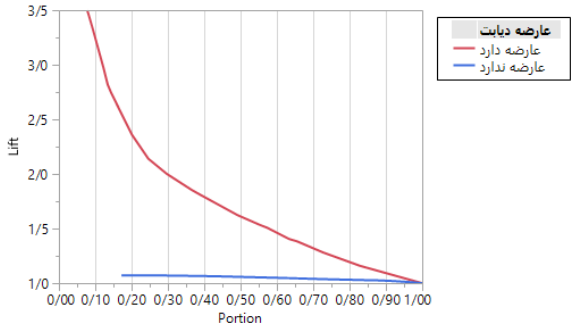
۵- شاخه‌ی نهایی ۴۱: احتمال رخداد عارضه‌ی دیابت برای بیمار دیابتی که سن بالای ۵۳ سال و سابقه‌ی سکته‌ی قلبی ندارد و هموگلوبین کمتر از ۶/۸۶ دارد و و بیماری قلبی‌وعروقی دارد و روغن مصرفی آن جامد است و شهر نشین است و کلسترول بالا هم دارد و کلسترول آن بیشتر از ۱۸۴ است برابر ۱۴/۵ درصد است.

۶- شاخه‌ی نهایی ۴۲: احتمال رخداد عارضه‌ی دیابت برای بیمار دیابتی که سن بالای ۵۳ سال و سابقه‌ی سکته‌ی قلبی ندارد و هموگلوبین کمتر از ۶/۸۶ دارد و و بیماری قلبی عروقی دارد و روغن مصرفی آن جامد است و شهر نشین است و کلسترول بالا هم دارد و کلسترول آن کمتر از ۱۸۴ است برابر ۲۹/۳ درصد است.



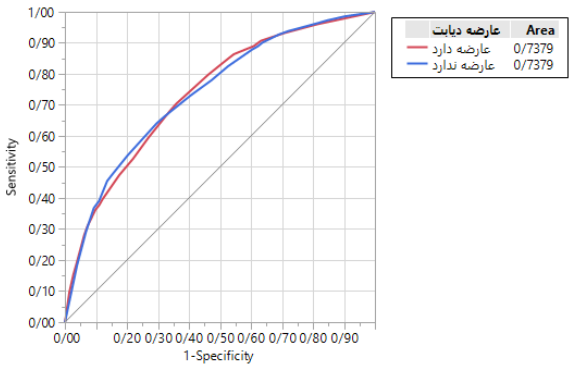
شکل ۱- مدل درختی برازش داده شده به بیماران دیابتی

با توجه به نمودار Lift در شکل ۲ با مدل درختی برازش داده شده در شکل ۱، قدرت تصمیم‌گیری رخ دادن عارضه‌ی دیابت برای فردی که مراجعه می‌کند ۳/۵ برابر افزایش می‌یابد.



شکل ۱- نمودار منحنی قدرت تصمیم گیری مدل

در بررسی کارایی مدل با استفاده از مساحت زیر منحنی مشخصه‌ی محرکه‌ی گیرنده (ROC Curve) برای تعیین میزان صحت مدل رده‌بندی درختی، مقدار مساحت ۷۳/۸ درصد تعیین شد که نشان دهنده‌ی توان نسبتاً بالای مدل رده‌بندی درختی در تعیین عوامل مؤثر بر فراوانی رخداد عوارض بیماری دیابت است.



شکل ۵- نمودار منحنی ROC

ضریب تعیین مدل برابر ۰/۱۱۳ برای داده‌های آموزشی و ۰/۱۰۴ برای داده‌های اعتبارسنجی است. این ضریب عدد کوچکی است و بیانگر این موضوع می‌تواند باشد که عوامل خطر مؤثر با اهمیت دیگری بر متغیر عارضه‌ی دیابت وجود دارد که در این مجموعه داده‌ها لحاظ نشده‌اند.

مدل برازش داده شده در این مقاله، با توجه به ضریب تعیین پایینی که دارد مدلی مناسب برای پیش‌بینی نیست ولی مدل خیلی با اهمیت برای کنترل جامعه است. یعنی می‌تواند برای تصمیم‌گیری مدیران اجرایی که برای برنامه‌ریزی دقیق‌تر در کنترل عارضه‌ی دیابت در بین بیماران دیابتی مفید واقع شود.

به‌عنوان مثال با توجه به مدل درختی شکل ۱ این مدل کمک می‌کند متوجه شویم که برای بیماران بالای سن ۵۳ سال چه نوع مراقبت‌ها و راهکارهایی می‌تواند مفیدتر باشد که عارضه‌ی دیابت کمتر پدیدار شود، همچنین برای افراد بالای سن ۵۳ سال که بیماری قلبی‌وعروقی دارند چه مراقبت‌هایی را باید در نظر گرفت. به همین ترتیب برای کل شاخه‌های درخت که طبقه‌ای از افراد بیماران دیابتی است می‌توان مراقبت‌های کنترلی را اعمال کرد. در مورد شاخه‌های انتهایی درخت هم می‌توان برنامه ریزی‌ها را طوری در جامعه انجام داد که تا حد

امکان بیماران دیابتی به شاخه‌های سمت چپ درخت که فراوانی رخداد عارضه کمتری دارند بروند.

نتیجه‌گیری

با توجه به نوع مجموعه داده‌ها و تعداد زیاد داده‌های تکمیل نشده (گمشده) و حجم بالای داده، روش درخت رگرسیون و طبقه‌بندی روشی مناسب و قابل تفسیر برای این تحقیق است. با مدلی که از این روش به داده‌ها برازش شد می‌توان کنترل جامعه را با توجه به عوامل بررسی شده در این تحقیق به خوبی انجام داد. با توجه به ضریب تعیین پایین مدل، این مدل پیش‌بینی خوبی برای آینده نخواهد داشت و در تحقیقات بعدی باید عوامل دیگری هم برای پیش‌بینی بهتر در نظر گرفته شود.

نتایج مدل رگرسیون و طبقه‌بندی درختی نشان داد که از متغیرهای کمی به‌ترتیب نزولی سن، Risk Score، گلوکز، هموگلوبین، مجموع زمان فعالیت، کمیت کلسترول، FBS و HDL و از متغیرهای کیفی نیز بیماری قلبی‌وعروقی، سابقه‌ی سکته، فشارخون، کلسترول، تجویز استاتین، شغل با فعالیت فیزیکی سخت، منطقه‌ی زندگی، روغن مصرفی، پیاده‌روی، مصرف سبزی‌ها و جنسیت در فراوانی رخداد عارضه‌ی دیابت مؤثرتر از عوامل دیگر هستند، اما عوامل دیگری هم برای پیش‌بینی دقیق‌تر آینده باید در نظر گرفته شود.

در آینده پیشنهاد می‌شود روش‌های دیگر داده‌کاوی را برای تحلیل و بررسی بر روی این مجموعه داده و تأثیراتی که در این تحقیق بررسی نشده است کار شود. یکی از روش‌های مدرن در این زمینه شبکه‌های عصبی است که می‌توان روی این داده‌ها بررسی کرد و همچنین با روشی که در این تحقیق کار شد مقایسه شود. پیشنهاد می‌شود این تحقیقات برای دیگر بخش‌های داده‌های استخرج شده از سامانه‌ی سینا در دانشگاه علوم پزشکی مشهد و همچنین در سطح وسیع‌تر از سامانه‌های نظام سلامت کشور انجام شود تا باعث برنامه‌ریزی دقیق‌تر مدیران و پزشکان و پیراپزشکان از حجم وسیع داده‌های ثبت شده شود.

سپاسگزاری

از گروه آمار زیستی و گروه اپیدیمیولوژی در دانشکده بهداشت دانشگاه علوم پزشکی مشهد بابت همکاری علمی و همچنین معاونت بهداشتی دانشگاه علوم پزشکی مشهد بابت همکاری در اختیار گذاشتن داده های سامانه سینا تشکر و قدردانی به عمل می آید.

مآخذ

1. World Health Organization. *Prevalence of diabetes*. Retrieved December 11, 2008.
2. Wild S, Roiglic G, Grren A, Sicree R, King H. Global Prevalence of Diabetes. *Diabetes Care*, 2009, 27:1047–1053.
3. Recommendations for Health care system and self-management Education Interventions to reduce morbidity and mortality from diabetes, American *Journal of Preventive Medicine*, 2002; 22:10-14.
4. Zimmet PZ. Diabetes epidemiology as a tool to trigger diabetes research and care. *Diabetologia*, 1999; 42:499-518.
5. Garcia AA. Clinical and life quality differences between Mexican American diabetic patients at a free clinic and hospitals affiliated clinic in Texas. *Public Health Nursing*, 2008; 25:1496-158.
6. Welschen L. Disease management for patients with type 2 diabetes: towards patient empowerment. *International Journal of Integrated Care*, 2008; 8:992.
7. Coelho R, Amorim I, Prata J. Coping Styles and quality of life in patients with non-insulin dependent diabetes mellitus. *Psychosomatics*, 2003;44:312-318.
8. Dunn SM, Welch GW, Butow PN, Coates AS. Refining the measurement of psychological adjustment in cancer. *Australian Journal of psychology*, 1997; 49:144-151.
9. Harrison TR, Braunwald E, Longo D, Jameson JL. *Harrison's principals of internal medicine, Endocrinilogy.*16th edition. McGraw-Hill, 2004.

10. Henry D, Bernard S. *Life manner with diabetes. Trans*. By Panahi A. Tehran; Javid: 1989 [in Persian].
11. Amini M, Khadivi R, Haghighi S. Costs of type 2 diabetes in Isfahan, Iran in 1998. *Iranian Journal of Endocrinology and Metabolism*, 2002;4(14): 104-97.
12. Ghanbari A. Determination of model of effective factors on quality of life domains among type 2 diabetic patients. *Journal of medicine faculty*, 2001; 10(37-8):82-9. [In Persian].
13. Snoek FJ, Pouwer F, Welch GW, Polowsky WH. Diabetes-related emotional distress in Dutch and US. Diabetic patients, *Diabetes Care*, 2000; 23(9):1305–9.
14. Wagner EH, Austin BT, Von Korff M. Organizing care for patients with chronic illness. *Milbank* 1996; 74 (4) :511–44
15. Tapak L, Mahjub H, Hamidi O, PoorolajalJ. Realdata comparison of data mining methods in prediction of diabetes in iran. *Healthcare informatics research*, 2013; 19(3):177-85.
16. Jayalakshmi T, Santhakumaran A, et al. A Novel Classification Method for Diagnosis of Diabetes Mellitus Using Artificial Neural Networks. 2010 *International Conference on Data Storage and Data Engineering*; 2010 9-10 Feb. 2010.
17. Choi SB, Kim WJ, Yoo TK, Park JS, et al. Screening for Prediabetes Using Machine Learning Models. *Computational and Mathematical Methods in Medicine*, 2014; 2014:8.
18. Boskabadi M, Doostparast M. Modeling and data mining of global data on patients with COVID-19. *Iranian Journal of Emergency Medicine*, 2020; 7(1):1 [e40]-6.
19. Boskabadi M, Afzalaghaee M, Talkhi N, Jamalian Z, Musa Farkhani E, & Esmaily H. Modeling the Impact of some Variables the COVID-19 Severe with CART Algorithm in Mashhad University of Medical Sciences. *Iranian Journal of Emergency Medicine*, 2022; 9(1): e36.
20. Breiman L, Friedman JH, Olshen RA, Stone CJ. *Classification and Regression Trees*. CRC Press; New York, 1984.
21. Boskabadi M, Doostparast M, Sarmad M. Survival analyses with dependent covariates: A regression tree-base approach. *Journal of Algorithms and Computation*, 2020;52(1):105-29.

Modeling and Identification of Factors Affecting the Existence of Complications Caused by Diabetes with Data Mining Methods

Mostafa Boskabadi[1], Najmeh Mohajeri[2], Ali Taghipour[3], Habibollah Esmaily[4], Seyed Javad Hoseini[2], Ehsan Musa Farkhani[3*]

1. School of Health, Mashhad University of Medical Sciences, Mashhad, Iran
2. Health Assistance, Mashhad University of Medical Sciences, Mashhad, Iran
3. Department of Epidemiology, School of Health, Mashhad University of Medical Sciences, Mashhad, Iran
4. Department of Biostatistics, School of Health, Mashhad University of Medical Sciences, Mashhad, Iran

ABSTRACT

Background: In Iran, with the advancement of technology and the development of registration statistics, the need to use data mining methods has attracted more attention from researchers. Regression and classification tree is one of the important methods in Big data modeling, which has attracted the attention of many researchers for community control and prediction. The purpose of this study is to determine the influencing variables on the occurrence of complications caused by diabetes.
Methods: This paper is a cross sectional-analytical study. In this research, all diabetic patients covered by Mashhad University of Medical Sciences in 2017 were extracted from the SINA system. The number of diabetics with complications was 5016 and diabetics without complications were 53613. The method of fitting the regression tree model and classification and measurement criteria of the model is the coefficient of determination and the area of the Rock curve and the Lift diagram.
Results: The rock curve for the fitted tree model is 73.8%, which shows the relatively high power of the model. Based on the Lift chart, the decision-making power of diabetes complications increases 3.5 times for the person who comes to visit.
Conclusion: The results of the regression model and tree classification showed that, in descending order, age, risk assessment factor, FBS, HbA1C, total activity time, cholesterol, FBS and HDL, cardiovascular disease, history of stroke, blood pressure, cholesterol Statin prescription, job with hard physical activity, living area, consumed oil, walking, consumption of vegetables and gender are more effective than other factors in the occurrence of diabetes complications.

Keywords: Regression and Classification Tree, Complications, Diabetes, Rock curve

* P.O.Box: 9137673119, Mashhad University of Medical Sciences, Daneshgah St, Mashhad, Iran, Tel: +985138514548, Email: farkhanie@mums.ac.ir