استخراج دانش از داده‌های بیماران دیابتی با استفاده از روش درخت تصمیم C5.0

حکیمه عامری[1]/ سمیه علیزاده[2]/ اکبر برزگری[3]

چکیده

مقدمه: بروز دیابت در ده سال اخیر در سطح جهان دو برابر شده است . حدود ۲۰۰ میلیون نفر به این بیماری مبتلا هستند و سالانه شیوع دیابت در جهان حدود شش درصد افزایش می‌یابد. بیش از دو میلیون نفر در ایران به این بیماری مبتلا هستند. در این تحقیق به بررسی ارتباط بین عوارض مشاهده شده در بیماران دیابتی نوع دو و برخی ویژگی‌های آن‌ها از قبیل میزان قند خون، فشار خون، سن و سابقه خانوادگی بیماران می‌پردازیم. هدف این مطالعه، پیش بینی عوارض بیماران بر اساس علائم مشاهده شده در آن‌ها است.

روش کار: داده‌های مورد نیاز برای این تحقیق از پرونده‌های سال ۱۳۸۸ مرکز دیابت استان گلستان جمع آوری شده است. تعداد پرونده‌های اولیه بیماران ۸۵٦ رکورد بود. در این مقاله مدل جدیدی بر اساس متدولوژی استاندارد CRISP ارائه شده است. در بخش مدل سازی از دو روش شناخته شده در داده کاوی به نام‌های درخت تصمیم C5.0 و شبکه عصبی استفاده شده است. برای تحلیل داده‌ها از نرم افزار Celementine 12.0 استفاده شده است.

یافته‌ها: در این تحقیق برای اولین بار احتمال بروز عوارض میکروواسکولار، ماکروواسکولار و یا هر دو نوع عارضه در بیماران و ویژگی‌های تأثیرگذار بر آن‌ها مورد بررسی قرار گرفته است. با استفاده از داده کاوی و روش‌های آن تعیین شده است که متغیرهای فشار خون بالا، سن و سابقه خانوادگی در عوارض مشاهده شده بیشترین تأثیر را داشته اند. به کمک درخت تصمیم ایجاد شده، قوانینی استخراج شده اند که می‌تواند به عنوان الگویی برای پیش بینی وضعیت بیماران و احتمال بروز عوارض در آن‌ها استفاده شود. صحت مدل ایجاد شده بر روی داده‌های مورد استفاده در درخت تصمیم C5.0، ۸۹.۷٤ درصد و در شبکه عصبی مصنوعی ۵۱.۲۸ درصد می‌باشد.

نتیجه گیری: با توجه به روش‌های استفاده شده، بالاترین دقت با استفاده از الگوریتم C5.0 به دست آمده است. بیشترین عوامل تأثیرگذار بر بروز عوارض شناسایی شدند. با توجه به قوانین ایجاد شده برای یک نمونه جدید با ویژگی‌های مشخص، می‌توان پیش بینی کرد بیمار احتمالاً دچار چه نوع عارضه ای خواهد شد.

کلید واژه‌ها: دیابت نوع دو، عوارض دیابت، داده کاوی، الگوریتم C 5.0، شبکه عصبی مصنوعی

•وصول مقاله: ۹۲/۳/۱۹ •اصلاح نهایی:۹۲/٦/۳ •پذیرش نهایی:۹۲/٦/۲۷

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

۲. استادیار گروه مهندسی فناوری اطلاعات، دانشکده مهندسی صنایع، دانشگاه خواجه نصیرالدین طوسی، تهران، ایران؛ نویسنده مسئول (s_alizadeh@kntu.ac.ir)

۳. پزشک، دانشگاه علوم پزشکی استان گلستان

مقدمه

به گزارش مرکز تحقیقات دیابت 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 شش درصد افزایش می‌یابد. در مطالعه ای که در ایران انجام شده است، گزارش شده که ۷.۷ درصد بالغین ۲۵ تا ۶۴ ساله که حدود دو میلیون نفر هستند، مبتلا به دیابت بوده و ۱۶.۸ درصد بالغین معادل با چهار میلیون نفر در وضعیت عدم تحمل گلوکز قرار دارند که تعداد زیادی از این بیماران در آینده به دیابت مبتلا خواهند شد. با توجه به اینکه بیماری دیابت، به عنوان یک بیماری بسیار مزمن شناخته شده است و آسیب های جبران ناپذیری به اندام‌ها و اعضاء حیاتی بدن وارد می کند، استفاده از ابزارهای هوشمند داده کاوی می‌تواند برای بهبود روش های شناسایی و کنترل بیماری به پزشکان کمک بزرگی باشد. به گزارش مرکز دیابت تحقیقات نشان داده است که میتوان با شناسایی‌های اولیه افراد در معرض خطر از ۸۰ درصد عوارض مزمن دیابت نوع دو جلوگیری کرد یا آن‌ها را به تعویق انداخت [۱].

دو نوع دیابت وجود دارد، نوع یک که وابسته به انسولین نیز نامیده می‌شود و دیابت نوع دو که کمبود نسبی انسولین است [۲]. عوارض مزمن دیابت بطور عمده به دو دسته تقسیم می‌شود: عوارض عروقی و عوارض غیرعروقی. عوارض عروقی دیابت خود دو دسته اند: میکرواسکولار (Micro vascular) شامل رتینوپاتی (نابینایی)، نفروپاتی (آسیب‌های گلومرول و دفع آلبومین) و نوروپاتی (کاهش یا ازدست دادن حس درد) و ماکرواسکولار (Macro vascular) بیماری کرونر، درگیری عروق محیطی و درگیری عروق مغزی را شامل می‌شود. از عوارض غیرعروقی می‌توان به گاستروپارزی، عفونت‌ها و تغییرات پوستی اشاره کرد [۳].

با توجه به شیوع دیابت نوع دو در سراسر جهان، استفاده از روش‌های جدید در تحقیقات پزشکی زیستی بسیار مورد توجه قرار گرفته است. داده کاوی می‌تواند ارتباطات و وابستگی‌های جدید و بدیعی را کشف کند که برای پزشکان مفید هستند.

داده کاوی نشان دهنده یک پیشرفت قابل توجه در انواع ابزار تحلیلی در دسترس است و به عنوان یک روش معتبر، حساس و قابل اعتماد برای کشف الگوها و روابط بین آن‌ها در نظر گرفته می‌شود [۴]. امروزه، ابزار داده کاوی به طور گسترده ای برای درک الگوهای بازاریابی، رفتار مشتری، بررسی داده‌های بیماران و شناسایی تقلب استفاده می‌شوند [۵]. داده کاوی به عنوان تکنیکی برای شناسایی و تشخیص بیماری ها و دسته بندی بیماران در مدیریت بیماری و پیدا کردن الگوهایی برای تشخیص سریعتر بیماران و جلوگیری از بروز عوارض در آن‌ها می‌تواند کمک بسیار بزرگی باشد. افزایش دقت تشخیص، کاهش هزینه‌ها و کاهش منابع انسانی به عنوان مزایای معرفی داده کاوی در تجزیه و تحلیل پزشکی توسط خواجوی و جایالاکشی ثابت شده است [۶،۷].

روش‌های اصلی داده کاوی دو دسته می‌باشند: توصیفی و پیش بینانه. وظایف توصیفی خواص عمومی داده‌ها را مشخص می‌کند و هدف آن پیدا کردن الگوهای قابل تفسیر توسط انسان برای داده‌ها است. وظایف پیش بینانه، پیش بینی رفتار آینده آن‌ها است و منظور از آن به کارگیری چند متغیر یا فیلد در پایگاه داده برای پیش بینی مقادیر آینده یا ناشناخته دیگر متغیرها است. یکی از عملکردهای پیش بینی، دسته بندی است. دسته بندی فرایند یافتن مدلی است که با تشخیص دسته‌ها یا مفاهیم داده می‌تواند دسته ناشناخته اشیاء دیگر را پیش بینی کند. در واقع دسته بندی یک تابع یادگیری است که یک قلم داده را به یکی از دسته‌های از قبل تعریف شده نگاشت می‌کند [۸]. یکی از روش‌های متداول دسته بندی، درخت تصمیم است. درخت تصمیم یک توصیف صریح از شاخه زنی با استفاده از الگوریتم است. این درخت ساختاری شبیه به فلوچارت دارد که بالاترین گره، ریشه درخت است، هر شاخه بیانگر خروجی‌های آزمایش و گره‌های برگ دسته

یا توزیع دسته‌ها را نمایش می‌دهد [۹،۸]. قوانین ایجاد شده توسط درخت تصمیم به صورت «اگر» و «آنگاه» بیان می‌شوند.

از الگوریتم‌های پرکاربرد، درخت تصمیم الگوریتم C5.0 است. C5.0 یک الگوریتم برای ساخت درخت‌های تصمیم گیری است که توسعه یافته الگوریتم ID3 است [۱۰]. این الگوریتم می‌تواند برای بیان دسته بندی به صورت درخت تصمیم و یا مجموعه قوانین به کار برده شود. در بسیاری از برنامه‌های کاربردی، مجموعه قوانین ترجیح داده می‌شوند زیرا درک آن‌ها نسبت به درخت‌های تصمیم گیری، ساده تر است.

الگوریتم‌های شبکه عصبی مصنوعی تلاش می‌کنند ساختار شبکه عصبی انسان را شبیه سازی کنند. این الگوریتم‌ها از مجموعه ای از گره‌ها به نام نرون ساخته شده اند که هر گره ورودی و خروجی‌هایی دارد و هر کدام وزنی خاص دارند. هر گره بر اساس تابعی خاص، محاسبه ساده ای انجام می‌دهد. بین گره‌ها اتصالاتی وجود دارد که بر اساس معماری شبکه مشخص می‌شوند. خروجی الگوریتم شبکه عصبی مصنوعی به شکل جعبه سیاه (Black Box) است. از این شبکه‌ها می‌توان به عنوان روش مناسب در ایجاد مدل‌های تحلیلی و تخمینی و برخورد با داده‌های متفاوت استفاده کرد [۸].

بریولت و همکاران [۱۱] با استفاده از سیستم CART طبقه بندی و تجزیه و تحلیل رگرسیون را در سال ۲۰۰۲ انجام داده و وابستگی بین یک سری از ویژگی‌های آن را استنباط کرده اند. میزان دقت طبقه بندی ۵۹.۹ درصد بوده است. همچنین میاکی و همکاران [۱۲] از روش کارت برای قضاوت عوامل مؤثر بر بروز عوارض دیابت در سال ۲۰۰۲ استفاده کرده اند. رولفینگ و همکاران [۱۳] از روش تجزیه و تحلیل رگرسیون خطی برای بررسی ارتباط بین قند خون در دیابت نوع یک و HbA_1c در سال ۲۰۰۲ استفاده کرده اند. سیلورستین و همکاران [۱۴] آزمایش‌هایی را بر روی سه پایگاه داده پزشکی انجام داده اند و قوانینی تولید کرده اند و سپس این قوانین را با قوانین از پیش تعیین شده مقایسه کرده اند.

کوانتین تروونتل و همکاران [۱۵] از روش قواعد انجمنی و درخت تصمیم برای استخراج دانش از پایگاه داده پزشکی استفاده کرده اند. گیائو جان و همکاران [۱۰] در سال ۲۰۰۷ با استفاده از ترکیب الگوریتم‌های C4.5 و EM (حداکثر انتظار) سیستم پردازش داده‌های دیابت نوع دو را ایجاد کرده اند. هوانگ و همکاران [۱۶] در سال ۲۰۰۷ تحقیقی بر روی شناسایی عوامل عمده تأثیرگذار بر کنترل دیابت، با به کار بستن انتخاب ویژگی‌ها (Feature Selection) در سیستم مدیریت بیمار انجام دادند. جیانکوهان و همکاران [۱۷] در سال ۲۰۰۸ با استفاده از نرم افزار Rapid Miner و با به کاربردن الگوریتم درخت تصمیم ID3 وجود دیابت را در پایگاه داده بیماران پیش بینی کرده اند. آنبانانتن و همکاران [۱۸] در سال ۲۰۰۵ شبکه عصبی مصنوعی و درخت تصمیم ساخته شده از الگوریتم C4.5 را به کار بستند تا وجود دیابت در افراد را بر اساس ویژگی‌هایی مثل سن و فشار خون تشخیص دهند. بیک هوان چو و همکاران [۱۹] در سال ۲۰۰۷ با استفاده دسته بندی ماشین بردار پشتیبانی SVM (Support Vector Machine) و با روش انتخاب ویژگی و تجسم سازی، وجود نروپاتی در بیماران دیابتی را پیش بینی کرده اند. فانگ [۵] در سال ۲۰۰۹ با استفاده از تکنیک‌های مختلف داده کاوی بیماران را بر اساس مبتلا بودن به دیابت، خوشه بندی کرده است. ویژگی‌هایی که در این مدل‌ها مهم شناخته شدند عبارتند از سن، سابقه خانوادگی و وزن. دقت مدل ایجاد شده با استفاده از خوشه بندی ۸۰ درصد است. چین [۲۰] در سال ۲۰۰۸ عوارض میکرووواسکولار دیابت را بررسی کرده است. وی برای این کار الگوریتم C5.0 و شبکه عصبی را با هم مقایسه کرده است. عوامل مختلفی برای هر کدام از این عوارض را شناسایی شده و میزان تأثیرگذاری آن‌ها بر هر کدام از این عوارض مورد بررسی قرار گرفته است. پتیل و همکاران [۲۱] در سال ۲۰۱۰ با استفاده از الگوریتم Apriori قوانین تلازمی را برای یافتن رابطه‌های پنهان بین متغیرها ایجاد

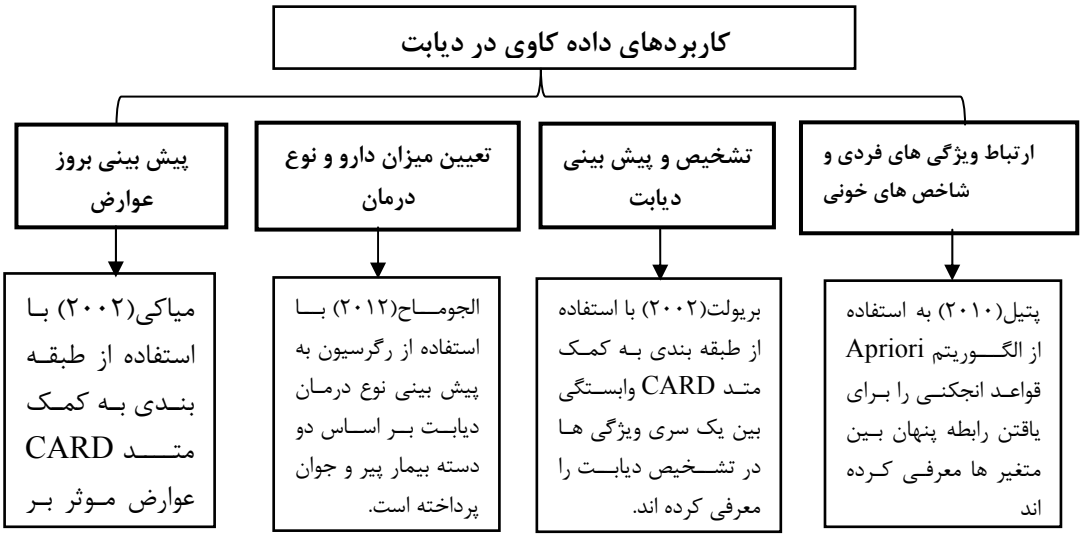
کرده اند. اسما الجراح [۴] از درخت تصمیم برای تشخیص بیماری دیابت نوع دو استفاده کرده است. با به کار بردن طبقه بندی الگوریتم درخت تصمیم J48 بروی داده‌ها در نرم افزار Weka، درخت تصمیم تولید شده است. الجوماج و همکاران [۲۲] با استفاده از روش رگرسیون به تجزیه و تحلیل پیش بینی درمان دیابت در دو دسته گروه سنی جوان و پیر بر اساس درمان دارویی و درمان‌های جانبی پرداخته اند. آنتونلی و همکاران [۲۳] در سال ۲۰۱۳ یک چارچوب تجزیه و تحلیل مبتنی بر خوشه بندی چندسطحی برای شناسایی مسیرهای درمان و معاینه بیماران برای بیماری‌های خاص ارائه کرده اند. روش پیشنهادی در شناسایی گروه‌های بیماران با تاریخچه بیماری مشابه و افزایش شدت عوارض آن‌ها به خوبی عمل کرده است.

در جدول شماره یک به طور اجمالی تاریخچه استفاده از تکنیک‌های داده کاوی بر روی تجزیه تحلیل و پیش بینی عوامل تاثیرگذار بر دیابت از سال ۲۰۰۲ تا به امروز مورد بررسی قرار گرفته است.

جدول ۱: تاریخچه استفاده از داده کاوی در بیماری دیابت

سال تحقیق	نویسندگان مقاله	روش مورد استفاده	نتایج بدست آمده
۲۰۰۲	Breault و همکاران	طبقه بندی با استفاده از روش CARD	وابستگی بین یک سری از ویژگی‌های بیماران
۲۰۰۲	Miyaki و همکاران	طبقه بندی با استفاده از روش CARD	قضاوت عوامل موثر بر بروز عوارض دیابت
۲۰۰۲	C.L.Rohlfingو همکاران	تجزیه و تحلیل رگرسیون خطی	ارتباط بین قند خون در دیابت نوع یک وHIbAlc
۲۰۰۵	Anbananthenو همکاران	شبکه عصبی مصنوعی و الگوریتم C4.5	پیش بینی وجود دیابت در بیماران
۲۰۰۷	Baek Hwan Cho و همکاران	دسته بندی ماشین بردار پشتیبانی (SVM)	پیش بینی بروز نروپاتی در بیماران دیابتی
۲۰۰۷	jaun.Gaoو همکاران	ترکیب الگوریتم‌های C4.5 و EM	سیستم پردازش داده‌های دیابت نوع ۲
۲۰۰۷	Yue Huangو همکاران	انتخاب ویژگی‌ها(Feature Selection)	عوامل عمده تاثیرگذار بر کنترل دیابت مشخص شده‌اند
۲۰۰۸	Jianchao Han و همکاران	الگوریتم‌های ID3 و Decision Tree	پیش بینی وجود دیابت در بیماران
۲۰۰۸	Chain	الگوریتم C5.0 و شبکه عصبی مصنوعی	بررسی عوارض میکروواسکولار دیابت
۲۰۰۹	Xiao Fang	خوشه بندی و رگرسیون	خوشه بندی بیماران بر اساس مبتلا بودن به دیابت
۲۰۱۰	Patil و همکاران	الگوریتم Apriori	قوانین تلازمی برای پیدا کردن ارتباطات پنهان بین ویژگی‌ها
۲۰۱۲	A. AlJarullah .Asma	الگوریتم درخت تصمیم J48	تشخیص دیابت نوع ۲
۲۰۱۲	Aljumah و همکاران	رگرسیون	پیش بینی درمان دیابت در دو دسته گروه سنی جوان و پیر بر اساس نوع درمان
۲۰۱۳	Antonelli و همکاران	خوشه بندی چند سطحی	شناسایی مسیر درمان بیماری

با نگاهی کلی به مطالعات انجام شده در این حوزه، میتوان کاربرد داده کاوی در دیابت را به چهار دسته کلی تقسیم کرد. شکل شماره یک این دسته بندی‌ها را نمایش می‌دهد. مطالعات معدودی در زمینه پیش بینی بروز عوارض در بیماران دیابتی انجام شده است.



شکل ۱: دسته بندی کاربرد داده کاوی در دیابت

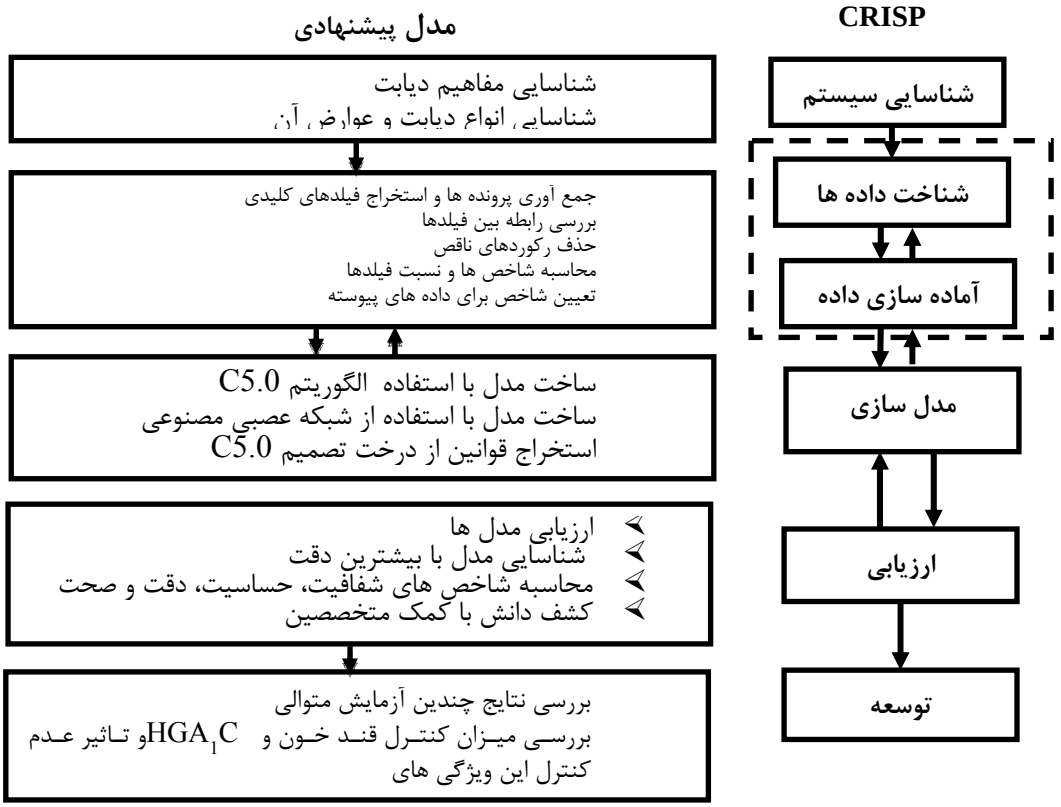
هدف این مقاله دسته بندی بیماران دیابتی نوع دو و پیدا کردن الگویی بین نشانه‌ها و علائم آزمایشگاهی بیماران و سوابق خانوادگی و شاخص BMI (Body Mass Index) آن‌ها با عوارض مشاهده شده دیابت در بیماران می‌باشد. با شناسایی این عوامل می‌توان به بیماران و پزشکان در جهت کنترل آن‌ها کمک بسیاری کرد. همچنین می‌توان امیدوار بود با کنترل عوارض دیابت، هزینه‌های درمانی مورد نیاز را کاهش داد.

روش کار

روش‌های مختلفی برای پیاده سازی و اجرای پروژه‌های داده کاوی وجود دارد.یکی از روش‌های بسیار قوی متدولوژی CRISP (Cross Industry Process For Data Mining) است[۸]. در این مقاله مدل پیشنهادی بر اساس CRISP ارائه شده است که شامل پنج فاز است. هر یک از این فازها خود شامل زیر بخشهایی می‌شوند .حرکت رو به جلو و عقب بین فازهای مختلف نیاز است زیرا ورودی هر فاز به خروجی فاز مرحله قبل وابسته است. هر یک از این پنج فاز در شکل ۲ نشان داده شده اند. در ادامه به بررسی هر یک از فازهای این مدل پیشنهادی متناسب با مطالعه موردی بیماری دیابت نوع دو می‌پردازیم.

الف) شناخت سیستم

در این فاز به شناخت سیستم مورد نظر پرداخته می‌شود و سپس اهداف مورد نظر و عوامل موفقیت کلیدی سیستم تعیین و دوباره بازنگری می‌شود. با توجه به رشد روز افزون بیماری دیابت و عدم وجود درمان قطعی برای این بیماری و تأثیرات و عوارض شدیدی که روی اعضاء حیاتی بدن در درازمدت می‌گذارد، بررسی داده‌های جمع آوری شده در رابطه با این بیماری برای تشخیص شباهت بیماران دیابتی جدید می‌تواند مفید باشد. بیماران جدید هم می‌توانند با توجه به دسته ای که در آن قرار می‌گیرند تا حد ممکن از توصیه‌های پزشکی تجویز شده برای بیماران قبلی بهره برند، از رژیم غذایی و نوع برنامه غذایی تجویز شده برای آن بیماران استفاده کنند و حرکات ورزشی متناسب با نوع غذای مصرفی برای کاهش میزان قند خون داشته باشند و از پیدایش عوارض دیابت جلوگیری کنند.



شکل ۲: گام‌های روش شناختی CRISP و مدل پیشنهادی

ب) شناخت داده‌ها و آماده سازی آن‌ها

در این فاز به جمع آوری داده‌های اولیه، توصیف داده‌ها، بازرسی و بررسی داده‌ها و اعتبار سنجی کیفیت داده‌ها می‌پردازیم. اطلاعات این تحقیق از یک مرکز درمان دیابت در شمال ایران جمع آوری شده است. اطلاعات پرونده‌ها مربوط به سال ۱۳۸۸ می‌باشد. ۸۵۶ رکورد اولیه از بیماران وجود داشت که پس از پالایش و حذف رکوردهایی که اطلاعات اصلی آن‌ها وجود نداشت به ۲۵۴ رکورد نهایی رسیدیم. 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 (آماده سازی داده‌ها یا پیش پردازش داده‌ها) به بررسی پرونده بیماران پرداختیم. در جهان واقعی، داده همیشه کامل نیست و در مورد اطلاعات پزشکی، این موضوع همیشه درست است. برای حذف تعدادی از تناقض‌ها و داده‌های ناقص در ارتباط با داده‌ها از پردازش داده استفاده کردیم. بسیاری از تکنیک‌های پردازش داده توسط چین و همکاران و هن و همکاران در [۲۴،۹] ارائه شده اند. در این تحقیق ما مواردی را که ارزش صفر برای ویژگی‌های فشار خون و FBS (Fasting blood glucose)، 2HDD-BS (2 Hour Post Prandial Blood Sugar)، TG (Triglycerides) و BMI داشتند را حذف کردیم. وایتن و همکاران در [۲۵] ثابت کردند که حذف عاقلانه، یک روش کارآمد به جای جایگزین کردن ارزش‌ها با تکنیک‌های مانند میانگین، انتساب تصادفی، انتساب رگرسیون و مدل‌های بیزی است.

بعضی از اطلاعات موجود در پرونده مانند نام و نام خانوادگی، شماره پرونده بیمار و آدرس حذف شدند. در مرحله بعدی پرونده بیمارانی را که فقط یک بار مراجعه داشته اند کنار گذاشتیم زیرا اطلاعات کاملی از آزمایشات و عوارض آن‌ها در دسترس نبود. بعضی از فیلدها به تنهایی اهمیتی نداشتند

(مثل BUN (Blood urea) و CR (Chromium)) اگر این نسبت بین ده تا ۲۰ باشد، نرمال و بیشتر از ۲۰ معنی خونریزی گوارشی یا انسداد دستگاه ادرار است. نسبت این فیلدها نشان دهنده احتمال وجود عارضه کلیوی است و یا قد و وزن که به تنهایی اهمیتی ندارند، بلکه شاخص BMI آن‌ها تأثیرگذار است. درنتیجه این فیلدها حذف شدند و به جای آن‌ها از شاخص‌های مرتبط استفاده شد. شاخص BMI به این صورت محاسبه میشود:

$$BMI = Weight(kg) / (Height(cm))^2 \quad (1)$$

در نتیجه پس از پالایش داده‌ها به رکوردهایی با مشخصات جدول شماره دو رسیدیم.

جدول ۲: داده‌ها و نوع آن‌ها پس از پاکسازی

مشخصه	توضیحات	نوع
Age	سن	عددی
Sex	جنسیت	اسمی
BMI	شاخص نسبت وزن به قد	عددی
Family History	سابقه خانوادگی	رتبه ای
SBP(Systolic blood pressure)	فشار خون سیستولیک	عددی
FBS	قند ناشتا	عددی
2HPP-BS	قند ۲ ساعت بعد از غذا	عددی
LDL(Low-Density Lipoprotein)	چربی مضر	عددی
HDL (High-Density Lipoprotein)	چربی مفید	عددی
TG	تری گلیسرید	عددی
BUN/CR	احتمال عوارض کلیوی (اوره به کروم)	عددی
Complications	نوع عارضه	اسمی

ج) مدل سازی

روش‌های داده کاوی بسیاری برای مدل سازی وجود دارد. در این فاز با استفاده از تکنیک‌های مختلف داده کاوی به پیدا کردن مدل و الگوی بهینه میپردازیم. مدل سازی با استفاده از نرم افزار SPSS Celementine 12.0 انجام شده است. اصل روش کار، در اینجا داده کاوی پیش بینانه می‌باشد. از روش الگوریتم درخت تصمیم گیری استفاده شده تا بهترین نسبت بین فیلدهای مختلف به دست آید. در این مرحله درخت‌های تصمیم C5.0 و شبکه عصبی با ورودی‌های مختلف مورد آزمایش قرار گرفته اند. درخت تصمیم یک توصیف صریح از شاخه زنی با استفاده از الگوریتم است. هر گره پایانی یا برگ زیر مجموعه ای از داده‌های آموزشی را توصیف می‌کند و هر نمونه در بخش آموزش دقیقاً به یک گره پایانی در درخت تعلق دارد.

برای آموزش درخت تصمیم مدل C5.0 یک متغییر طبقه ای باید فیلد خروجی باشد و یک یا تعداد بیشتری فیلد ورودی وجود داشته باشد. فیلدهای ورودی مقادیر به دست آمده از آزمایشات بیماران و خروجی عارضه‌های مشاهده شده است. گزینه آزمایشی انتخاب شده، اعتبارسنجی متقاطع با ده تکرار (10-Fold Cross-Validation) بود. اعتبارسنجی متقاطع ده به این دلیل انتخاب شده است که آزمایش‌ها نشان داده اند بهترین انتخاب برای به دست آوردن دقیق ترین تخمین است [۴،۲۶].

معمولاً مجموعه‌های قوانین، مهمترین اطلاعات مرتبط با درخت تصمیم را در بردارند. در گزارش مدل تولید شده توسط نرم افزار، معیار تقسیم شاخه‌ها بر اساس شاخص جینی (Gini Index) انتخاب و درخت به زیرشاخه‌هایی شکسته می‌شود. این روند تناوبی ادامه

می‌یابد تا در نهایت داده‌های هر گره در یک دسته قرار داشته باشند. برای بیان قوانین استخراج شده، مسیر ریشه تا برگ پیمایش می‌شود و قوانین به صورت شرطی بیان می‌شود. معیار اطمینان (Confidence) برای قوانین ایجاد شده ۷۰ درصد تعیین شده است. برچسب دسته (نوع عارضه) در مدل ایجاد شده در جدول شماره سه توضیح داده شده است.

جدول ۳: برچسب دسته مدل

برچسب دسته	توضیحات
False	بیمارانی که عارضه ای در آن‌ها مشاهده نشده است
Micro vascular	عوارض میکروواسکولار (عارضه چشمی، فشارخون، دیالیز، زخم پا)
Macro vascular	عوارض ماکروواسکولار (سکته قلبی، سکته مغزی)
Micro-Macro	عوارض میکروواسکولار و ماکروواسکولار

جدول شماره پنج مجموعه قوانین ایجاد شده توسط درخت C5.0 را نمایش میدهد. قوانین درخت C5.0 به این دلیل انتخاب شده اند که صحت و دقت بالاتری نسبت به الگوریتم‌های دیگر دارد. در بخش ارزیابی این شاخص‌ها و نحوه محاسبه آن‌ها توضیح داده شده است. باید توجه داشت که همه بیماران دیابتی مستعد عوارض هستند و تلاش ما بررسی گزینه‌هایی برای به تعویق انداختن و در صورت امکان عدم بروز عوارض می‌باشد.

د) ارزیابی

در این فاز پس از مدلسازی باید به ارزیابی نتایج حاصل از مدلسازی پرداخت. نتایج ارزیابی باعث بهبود مدل می‌شود و مدل را قابل استفاده می‌نماید. برای بررسی صحت مدل، 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 تست و برچسب مربوط به رکوردهای مذکور را تعیین می‌نمایند. داده‌های بخش اعتبارسنجی نیز صحت مدل تولید شده را بررسی می‌کنند. شاخص‌های مختلفی برای ارزیابی صحت روش‌های دسته بندی وجود دارد که میتوان حساسیت (Sensitivity)، شفافیت (Specificity)، دقت (Precision) و صحت (Accuracy) را نام برد. میزان صحت یک روش دسته بندی بر روی مجموعه داده‌های آموزشی، درصد مشاهداتی از مجموعه آموزش است که به درستی توسط روش مورد استفاده دسته بندی شده است. برای محاسبه این شاخص داده‌های آزمون استفاده می‌شوند. همچنین میتوان نرخ خطا (Error Rate) یا دسته بندی نادرست (Misclassification rate) را بر اساس شاخص صحت محاسبه کرد [۸،۹].

صحت - ۱= نرخ خطا (۲)

برای محاسبه میزان صحت مدل می‌توان از ماتریس اغتشاش (Confusion Matrix) استفاده کرد. این ماتریس ابزاری مفید برای تحلیل چگونگی عملکرد روش دسته بندی در تشخیص داده‌ها یا مشاهدات دسته‌های مختلف است. اگر داده‌ها در M دسته قرار گرفته باشند، یک ماتریس دسته بندی جدولی با حداقل اندازه M*M است. حالت ایده آل این است که بیشتر داده‌های مرتبط به مشاهدات روی قطر اصلی ماتریس

قرار گرفته باشند و مابقی مقادیر ماتریس صفر یا نزدیک به صفر باشند [۸،۹].

$$\text{حساسیت} = \frac{\text{تعداد داده های برچسب مثبتی که درست دسته بندی شده اند}}{\text{کل تعداد داده های مثبت}} \quad (۳)$$

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

$$\text{دقت} = \frac{\text{تعداد داده های برچسب منفی که درست دسته بندی شده اند}}{\text{تعداد داده های برچسب مثبتی که درست دسته بندی شده اند + تعداد داده های برچسب منفی که به نادرست مثبت دسته بندی شده اند}} \quad (۵)$$

$$\text{صحت} = \text{حساسیت} \frac{\text{تعداد داده ها مثبت}}{\text{تعداد کل داده ها}} + \text{شفافیت} \frac{\text{تعداد داده ها منفی}}{\text{تعداد کل داده ها}} \quad (۶)$$

در این فرمول‌ها منظور از برچسب مثبت، یکی از برچسب دسته‌های (False,Micro,Macro,Micro-Macro) است و بر چسب منفی، کل مجموعه داده‌ها بجز برچسب دسته مثبت می‌باشد.

بهترین نتایج با استفاده از گره درخت C5.0 به دست آمده است. در این مرحله از درخت تصمیم C5.0 برای تعیین وجود یا عدم وجود عارضه ای خاص در بیمار با توجه به نتایج به دست آمده از آزمایش‌های بیمار، جنسیت، شاخص BMI و سابقه خانوادگی بیمار استفاده شده است. میزان صحت مدل C5.0 برای داده‌های آموزش ۸۲.۱۸ درصد، داده‌های آزمایش ۷۸.۰۵ درصد و قسمت اعتبارسنجی ۸۹.۷۴ درصد محاسبه شده است. میزان صحت مدل برای شبکه عصبی مصنوعی در قسمت داده‌های آموزش ۴۵.۴ درصد، داده‌های آزمایش ۵۳.۶۶ درصد و قسمت اعتبارسنجی ۵۱.۲۸ درصد محاسبه شده است. با توجه به جدول‌های فوق به وضوح مشخص است که میزان صحت مدل C5.0 بسیار بیشتر از شبکه عصبی مصنوعی است.

در جدول شماره چهار شاخص‌ها را که برای هر کدام از برچسب دسته‌ها برای الگوریتم C5.0 با استفاده از ماتریس اغتشاش به صورت جداگانه محاسبه کرده ایم را نمایش می‌دهیم [۹].

جدول ٤: میزان شاخص‌ها برای الگوریتم C5.0

	شفافیت (درصد)	حساسیت (درصد)	دقت (درصد)	صحت (درصد)
FALSE	۸۳.۶۵	۷۸.۴۷	۸۸.۵۷	۸۱.۵۶
MACRO	۱۰۰	۶۶.۶۶	۱۰۰	۹۷.۸۲
MICRO-MACRO	۹۵.۷۴	۷۸.۷۹	۸۱.۲۵	۹۲.۵۰
MICRO	۹۲.۶۶	۷۸.۴۶	۸۶.۴۴	۸۷.۳۶

میانگین صحت مدل با استفاده از ماتریس اغتشاش ۸۹.۷۴ درصد است. میزان خطای مدل ۱۰.۲۶ درصد است. درنتیجه می‌توان ادعا کرد مدل از دقت و صحت نسبتاً خوبی برخوردار است.

و) توسعه

ساخت مدل، پایان یک پروژه نیست و هدف از پروژه‌های داده کاوی کشف دانش و استفاده از دانش کشف شده در آینده می‌باشد. دانش کشف شده باید سازماندهی شود و به شکل قابل استفاده برای دیگران نیز در آید. ما در این فاز علاوه بر تهیه گزارش تلاش کردیم تا نشان دهیم میزان تأثیرگذاری پارامترهای مختلف به روی عوارض بیماری چه میزان است. می‌توان از این دانش برای پیش بینی وضعیت بیماران جدید استفاده کرد و در جهت کنترل دیابت بیماران همگام با دانش کشف شده از داده‌های قبلی گام برداشت.

یافته‌ها

کشف دانش از پایگاه داده‌های پزشکی به منظور تشخیص مؤثر پزشکی بسیار مهم است. هدف از داده کاوی استخراج دانش از اطلاعات ذخیره شده در پایگاه داده و ایجاد شرح روشن و قابل فهم از الگوها است. از بین الگوریتم‌های مورد استفاده، بهترین نتایج از الگوریتم درخت C5.O به دست آمد که دقت مدل آن برابر ۸۹.۰۶ درصد و صحت مدل ۸۹.۷۶ درصد است. در جدول شماره پنج نمونه قوانین ایجاد شده توسط درخت تصمیم C5.0 ارائه شده است.

فاز ارزیابی توسط افراد متخصص نیز انجام می‌شود. با توجه به قواعد تولید شده، پارامترهایی که بیشترین تأثیر را بر روی عوارض دارند، شناسایی و تایید شده اند. طبق نظر کارشناسان می‌توان از مدل ایجاد شده نتایج زیر را گرفت.

از آنجایی که فشار سیستولیک نرمال در بیماران دیابتی بین ۱۲۵ تا ۱۳۵ می‌باشد و مهمترین فاکتور ریسک در این بیماران فشارخون بیشتر از ۱۳۵ است، اطلاعات ذیل را از درخت تصمیم ایجاد شده استخراج کرده ایم.

i. اگر فشار سیستولیک بیشتر از ۱۳۵ باشد و سن بیشتر از ۵۶ سال و نسبت احتمال بروز عارضه کلیوی بزرگتر از ۳۰ باشد آنگاه احتمال گرفتاری عوارض میکروواسکولار بالا است.

ii. کسانی که سابقه خانوادگی مثبت دارند حتی در فشار بین ۱۰۰تا ۱۳۵ نیز گرفتار عوارض میکروواسکولار می‌شوند.

iii. اگر فشار سیستولیک بیشتر از ۱۴۵ و سن بیش از ۵۶ و نسبت احتمال بروز عارضه کلیوی بزرگتر از ۳۰ و تری گلیسرید بزرگتر از ۱۳۸ باشد آنگاه احتمال گرفتاری عوارض ماکروواسکولار بالا است.

iv. بیمارانی که سابقه خانوادگی مثبت دارند و چربی مضر بیشتر از ۱۰۰ دارند، حتی با داشتن فشار نرمال ۱۳۵ دچار عوارض میکروواسکولار شده اند.

v. در فشارخون پایین و سن کمتر از ۵۰ سال و میزان چربی مفید کمتر از ۷۵ حتی با وجود سابقه خانوادگی، با کنترل نسبت احتمال بروز عارضه کلیوی باشد بیماران کمتر دچار عارضه شده اند.

vi. در فشار خون بالا و سن کمتر از ۶۰ سال در صورت بالا بودن نسبت احتمال بروز عارضه کلیوی، بیماران با احتمال بیشتری به عوارض میکروواسکولار دچار می‌شوند.

به صورت کلی می‌توان نتیجه گرفت که اگر فشار بین ۱۲۵تا ۱۳۵ باشد و چربی مضر کمتر از ۱۰۰ باشد، کمتر دچار عوارض ماکروواسکولار می‌شوند، در نتیجه با کنترل این پارامترها می‌توان از بروز این عوارض تا حدی اجتناب کرد.

جدول ۵: نمونه ای از قواعد ایجاد شده توسط درخت تصمیم C5.0

ردیف	قوانین
۱	اگر ۱۰۰=> فشارخون سیستولیک ، ۵۶=>سن>۳۶ و سابقه خانوادگی =مثبت و قند ناشتا>۱۰۰ و شاخص BMI<=۲۵.۵۵۹ و ۷۸=>چربی مضر باشد آنگاه برچسب دسته " **Micro vascular** " است.
۲	اگر ۱۲۵=> فشارخون سیستولیک>۱۰۰ ، ۵۶=>سن>۳۶ ،سابقه خانوادگی =مثبت و ۷۵=>چربی مفید ،جنسیت مرد و ۱۹.۲۲۳=> احتمال عوارض کلیوی> ۱۶.۶۶۷ باشد آنگاه برچسب دسته " **Micro vascular** " است.
۳	اگر ۱۴۵=> فشارخون سیستولیک>۱۰۰ ، ۵۲=>سن>۳۶ و سابقه خانوادگی =مثبت و قند ناشتا>۱۰۰،چربی مفید<=۳۲ و جنسیت زن آنگاه برچسب دسته " **Micro vascular**" است.
۴	اگر۱۳۰=< فشارخون سیستولیک>۱۰۰ ، ۵۲=>سن>۳۶ و سابقه خانوادگی =مثبت و ۱۳۰=> قند ناشتا، جنسیت زن و۷۵=>چربی مفید>۳۲ باشد آنگاه برچسب دسته " **Micro vascular**" است.
۵	اگر۱۳۰=< فشارخون سیستولیک>۱۰۰ ، ۵۲=>سن>۳۶ و سابقه خانوادگی =مثبت و قند ناشتا>۱۰۰ ، جنسیت زن ، ۳۱.۲۵۰<احتمال عوارض کلیوی، قند ۲ ساعت بعد از غذا<=۴۰۰و ۷۵=>چربی مفید>۳۲ باشد آنگاه برچسب دسته " **Micro vascular**" است.
۶	اگر ۱۴۵=> فشارخون سیستولیک ، ۵۶=>سن>۳۶ و سابقه خانوادگی =منفی ،۲۱۲=>تری گلیسرید، ۳۱<چربی مفید باشد آنگاه برچسب دسته " **Micro vascular**" است.
۷	اگر ۱۴۵=> فشارخون سیستولیک>۴۵، ۵۶< سن ، ۳۰=>احتمال عوارض کلیوی ، ۹۲=>چربی مضر ، ۱۶۸=>تری گلیسرید باشد آنگاه برچسب دسته " **Micro vascular**" است.
۸	اگر ۱۴۵=> فشارخون سیستولیک>۴۵، ۵۶< سن ، ۳۰=>احتمال عوارض کلیوی ، ۹۲چربی مضر< ،۲۸=>چربی مفید و جنسیت=مرد باشد آنگاه برچسب دسته " **Micro vascular**" است.
۹	اگر ۱۴۵=> فشارخون سیستولیک>۱۲۵، ۵۶< سن ، ۳۰=>احتمال عوارض کلیوی ، ۹۲چربی مضر< ،۲۱۱=>قند ۲ ساعت بعد از غذا و جنسیت=مرد باشد آنگاه برچسب دسته " **Micro vascular**" است.
۱۰	اگر ۱۴۵=> فشارخون سیستولیک>۱۲۵، ۵۶< سن ، ۳۰=>احتمال عوارض کلیوی ، ۱۱۶=>چربی مضر>۹۲ ،۲۱۱<قند ۲ ساعت بعد از غذا ،سابقه خانوادگی=مثبت و جنسیت=مرد باشد آنگاه برچسب دسته " **Micro vascular**" است.
۱۱	اگر ۱۴۵=> فشارخون سیستولیک، ۵۶< سن ، جنسیت زن، ۱۱۶=>چربی مضر>۳۲.۳۰<احتمال عوارض کلیوی و۱۳۸=>تری گلیسرید باشد آنگاه برچسب دسته " **Micro vascular**" است.
۱۲	اگر ۱۰۵=> فشارخون سیستولیک، ۵۶=>سن>۳۶، سابقه خانوادگی =منفی و ۲۱۲<تری گلیسرید باشد آنگاه برچسب دسته " **Macro vascular** " است.
۱۳	اگر ۱۴۵=> فشارخون سیستولیک، ۵۶<سن، ۳۰<احتمال عوارض کلیوی، ۱۳۸< تری گلیسرید و BMI => ۲۲.۹۱۳ باشد آنگاه برچسب دسته "**Macro vascular** " است.
۱۴	اگر ۱۴۵=> فشارخون سیستولیک، ۳۶=> سن آنگاه برچسب دسته "**False** " است.
۱۵	اگر ۱۰۰=> فشارخون سیستولیک، ۵۶=> سن>۲۵۷، ۳۶<قند ناشتا ، BMI => ۲۵.۵۵۹ و ۷۸>چربی مضر باشد آنگاه برچسب دسته "**False** " است.
۱۶	اگر ۱۰۰=> فشارخون سیستولیک، ۵۶=> سن>۳۶، سابقه خانوادگی=مثبت، ۱۰۰<قند ناشتا و BMI< ۲۵.۵۵۹ باشد آنگاه برچسب دسته "**False** " است.
۱۷	اگر ۱۲۵=> فشارخون سیستولیک>۱۰۰ ، ۵۶=> سن>۳۶، سابقه خانوادگی=مثبت،۷۵=>چربی مفید، جنسیت=مرد و ۱۶.۶۶۷=>احتمال عوارض کلیوی باشد آنگاه برچسب دسته " **False**" است.
۱۸	اگر ۱۲۵=> فشارخون سیستولیک>۱۰۰ ، ۵۶=> سن>۳۶، سابقه خانوادگی=مثبت،۷۵=>چربی مفید، جنسیت=مرد و ۱۹.۲۳۱<احتمال عوارض کلیوی باشد آنگاه برچسب دسته " **False**" است.
۱۹	اگر ۱۴۵=> فشارخون سیستولیک>۱۰۰ ، ۵۶=> سن>۵۲، سابقه خانوادگی=مثبت،۷۵=>چربی مفید، جنسیت=زن و ۳۱.۲۵۰=>احتمال عوارض کلیوی> ۱۶.۶۶۷و ۴۰۰=>قند ۲ ساعت بعد از غذا باشد آنگاه برچسب دسته "**False** " است.
۲۰	اگر ۱۴۵=> فشارخون سیستولیک>۱۰۰ ، ۵۶=> سن>۳۶، سابقه خانوادگی=مثبت و ۷۵>چربی مفید باشد آنگاه برچسب دسته "**False** " است.
۲۱	اگر ۱۲۵=> فشارخون سیستولیک>۴۵ ، ۵۹=> سن>۵۶،۷۵=>چربی مفید، جنسیت=زن،۹۲<چربی مضر ۳۰> احتمال عوارض کلیویو ۲۱۱=>قند ۲ ساعت بعد از غذا باشد آنگاه برچسب دسته "**False** " است
۲۲	اگر ۱۴۵=> فشارخون سیستولیک، ۵۶=> سن>۳۶، سابقه خانوادگی=مثبت و ۱۰۰=>قند ناشتا باشد آنگاه برچسب دسته "**Micro-Macro**" است.
۲۳	اگر ۱۰۰=> فشارخون سیستولیک، ۵۶=> سن>۳۶، سابقه خانوادگی=مثبت و ۲۵۷=>قند ناشتا>۱۰۰، BMI<=۲۵.۵۵۹، ۷۸> چربی مضر باشد آنگاه برچسب دسته "**Micro-Macro**" است.
۲۴	اگر ۱۴۵=> فشارخون سیستولیک، ۵۶=> سن>۳۶، سابقه خانوادگی=منفی و ۲۱۲=>تری گلیسریدو ۳۱=>چربی مفید باشد آنگاه برچسب دسته "**Micro-Macro**" است.
۲۵	اگر۴۵=> فشارخون سیستولیک، ۵۶< سن، و ۳۰=>احتمال عوارض کلیوی باشد آنگاه برچسب دسته "**Micro-Macro**" است.
۲۶	اگر۱۴۵=>فشارخون سیستولیک> ۴۵، ۵۶< سن، و ۳۰=>احتمال عوارض کلیوی ، ۹۲>چربی مفید و ۷۰< چربی مفید باشد آنگاه برچسب دسته "**Micro-Macro**" است.
۲۷	اگر۱۲۵=>فشارخون سیستولیک> ۴۵، ۵۹< سن، و ۳۰=>احتمال عوارض کلیوی ، جنسیت =زن ،۹۲<چربی مضر و ۲۱۱=>قند ۲ ساعت بعد از غذا باشد آنگاه برچسب دسته "**Micro-Macro**" است.
۲۸	اگر۱۴۵=>فشارخون سیستولیک> ۴۵، ۵۶< سن، و ۳۰=>احتمال عوارض کلیوی، ، جنسیت =زن ،۹۲<چربی مضر ، سابقه خانوادگی=مثبت ،۴۰=>چربی مفید و ۲۱۱<قند ۲ ساعت بعد از غذا باشد آنگاه برچسب دسته "**Micro-Macro**" است.
۲۹	اگر۱۴۵=>فشارخون سیستولیک> ۴۵، ۵۶< سن، و ۳۰=>احتمال عوارض کلیوی ، سابقه خانوادگی=مثبت ، جنسیت =زن ،۱۵۴=>چربی مضر>۱۰۱ ،۴۰<چربی مفید ،۱۸۵=>قند ناشتا و ۲۱۱<قند ۲ ساعت بعد از غذا باشد آنگاه برچسب دسته "**Micro-Macro**" است.
۳۰	اگر۱۴۵=>فشارخون سیستولیک> ۴۵، ۵۶< سن، و ۳۰=>احتمال عوارض کلیوی ، سابقه خانوادگی=منفی ، جنسیت =زن ،۱۵۴=>چربی مضر>۱۱۶ باشد آنگاه برچسب دسته "**Micro-Macro**" است.
۳۱	اگر ۱۴۵<فشارخون سیستولیک و ۵۱< چربی مفید باشد آنگاه برچسب دسته "**Micro-Macro**" است.

بحث و نتیجه گیری

در این تحقیق با استفاده از الگوریتم‌های داده کاوی به دسته بندی بیماران دیابتی بر اساس عارضه‌های مشاهده شده در آن‌ها پرداختیم. عوارض این بیماری را بر اساس دو دسته میکروواسکولار و ماکروواسکولار دسته بندی کردیم. از بین الگوریتم‌های مورد استفاده، بهترین نتایج از الگوریتم درخت C5.O به دست آمد که دقت مدل آن برابر ۸۹.۰۶ درصد و صحت مدل ۸۹.۷۴ درصد است.

بیشترین پارامترهای تأثیرگذار بر روی عوارض بیماری میزان فشار سیستولیک، سن، سابقه خانوادگی و چربی مضر شناخته شده اند. با استفاده از قوانین ایجاد شده، برای یک نمونه جدید با ویژگی‌های مشخص، می‌تواند پیش بینی کرد که این فرد احتمالاً دچار چه نوع عارضه ای خواهد شد. با کنترل عوامل تأثیرگذار بر بروز عارضه در هر بیمار، می‌توان امیدوار بود از بروز عارضه تا حدی اجتناب کرد و یا آن را به تعویق انداخت. در ادامه نتایج کار خود را با کارهای مشابه در این حوزه در جدول شش مقایسه کرده ایم. تحقیقی مشابه با کار پژوهشی انجام شده توسط ما که بروز عوارض میکروواسکولار، ماکروواسکولارو بروز هردو نوع عارضه را بررسی کرده باشد، یافت نشده است.

جدول٦: مقایسه نتایج کار با کارهای انجام شده قبلی

نویسندگان و سال ارائه تحقیق	روش انتخابی	عارضه مورد بررسی				تأثیر گذار ترین ویژگی‌ها	صحت مدل	حساسیت مدل انتخابی
		میکرو واسکولار	ماکرو واسکولار	هر دو نوع عارضه	نداشتن عارضه			
Miyaki و همکاران (۲۰۰۲)	CARD	✓	✓	☒	☒	سن، BMI و فشار خون سیستولیک	محاسبه نشده است	محاسبه نشده است
Beak hwan cho(۲۰۰۷)	دسته بندی ماشین بردار و رگرسیون	✓	☒	☒	☒	فشارخون و شاخص BMI بالا ، میکرو آلبومین و گلبول سفید	SVM: ٪۹۲ رگرسیون: ٪۹۱	SVM: ٪۳۶ رگرسیون: ٪۸۱
Chain(۲۰۰۸)	C5.0 و شبکه عصبی مصنوعی	✓	☒	☒	☒	کراتینین ، سن، سابقه خانوادگی، چربی مضر، HbA_1C	محاسبه نشده است	شبکه عصبی:٪۶۷.۱۸۳ C5.0: ٪۶۴.۲۵
مدل انتخابی ما	C5.0 و شبکه عصبی مصنوعی	✓	✓	✓	✓	فشار سیستولیک، سن، سابقه خانوادگی و چربی مضر	شبکه عصبی:٪۵۱.۲۴ C5.0: ٪۸۹.۷۴	C 5.0: ٪۷۵.۵۹

میاکی و همکاران مهم ترین ویژگی‌های تأثیر گذار بر عوارض میکرو واسکولار و ماکرو واسکولار را سن، شاخص BMI و فشارخون سیستولیک معرفی کردند، که مطابق با یافته‌های الگوریتم پیش نهادی ما می‌باشد. در مطالعه انجام شده توسط چین و همکاران برای شناسایی عوامل مؤثر بر عوارض میکروواسکولار، ویژگی‌های کراتینین، سن، سابقه خانوادگی، چربی مضرو HbA_1C بررسی شده اند. ویژگی مشترک مورد مطالعه های کراتینین، سن، سابقه خانوادگی، چربی مضر است. دقت مدل پیشنهادی ما بالاتر است. بریک در پژوهشی برای شناسایی عوارض میکروواسکولار از پارامترهای فشارخون و شاخص BMI بالا، میکرو آلبومین و گلبول سفید استفاده کرده است که ویژگی‌های مشترکی با مدل انتخابی ما دارد ولی برخلاف مدل ما، الگوریتم ارائه شده توسط بریک تنها عوارض میکروواسکولار را مورد بررسی قرار داده است.

برای ساخت مدل‌ها و دانش بهتر، استفاده از نتایج چندین آزمایش متوالی، بررسی میزان کنترل قندخون و HgA_1C و تأثیر عدم کنترل این ویژگی‌ها بروی بروز عوارض مختلف و طول مدت بیماری مفید می‌باشد.

References

1. Mohamed E. I, Linder R. Perriello G, Di Daniele N, Pöppl S. J, De Lorenzo A. Predicting type 2 diabetes using an electronic nose-based artificial neural network analysis. Diabetes, nutrition & metabolism. 2002. 15(4), 215-221.

2. Pickup J.C. Williams G. (Eds.). Textbook of diabetes, Blackwell Science, Oxford. 2003.

3. Ahmadi K. Guideline &book review. The internal (endocrine and lung). Ahmadi Cultural Institute. 2009 edition [Persian].

4. Al Jarullah, Asma A. Decision tree discovery for the diagnosis of type II diabetes. In Innovations in Information Technology (IIT), 2011 International Conference on, 2011; pp. 303-307. IEEE.

5. Fang X. Are you becoming a diabetic? A data mining approach. In Fuzzy Systems and Knowledge Discovery, 2009. FSKD'09. Sixth International Conference on, vol. 5, 2009; pp. 18-22. IEEE.

6. Khajehei M. Etemady F. Data Mining and Medical Research Studies. In Computational Intelligence, Modelling and Simulation (CIMSiM), 2010 Second International Conference on, 2010; pp. 119-122. IEEE.

7. Jayalakshmi T. Santhakumaran A. A Novel Classification Method for Diagnosis of Diabetes Mellitus Using Artificial Neural Networks. In Data Storage and Data Engineering (DSDE), 2010 International Conference on, 2010; pp. 159-163. IEEE.

8. Alizadeh S, Ghazanfari M, Teimorpour B .Data Mining and Knowledge Discovery, Publication of Iran University of Science and Technology .2nd ed. 2011 [Persian].

9. Han J. Kamber M. chapter 1: introduction :Data Mining: Concepts and Techniques , Morgan Kaufman Publisher. 2nd ed. 2006.

10. Juan G. Luo S. Jia H. Zhang T. and Han Y. Type 2 diabetes data processing with EM and C4. 5 algorithm. In Complex Medical Engineering, 2007. CME 2007. IEEE/ICME International Conference on, 2007; pp. 371-377. IEEE.

11. Breault, Joseph L. Colin R. Goodall, and Peter J. Fos. Data mining a diabetic data warehouse. Artificial Intelligence in Medicine 26, no. 1 ,2002; pp. 37-54.

12. Miyaki K. Takei I. Watanabe K. Nakashima H. & Omae K. Novel statistical classification model of type 2 diabetes mellitus patients for tailor-made prevention using data mining algorithm. Journal of epidemiology/Japan Epidemiological Association, 2002; 12(3), 243.

13. Rohlfing C. L. Wiedmeyer H. M. Little R. R. England J. D. Tennill A. & Goldstein D. E. Defining the relationship between plasma glucose and HbA1c analysis of glucose profiles and HbA1c in the Diabetes Control and Complications Trial. Diabetes care, 2002; 25(2), 275-278.

14. Silverstein C. Brin S. Motwani R. & Ullman J. Scalable techniques for mining causal structures. Data Mining and Knowledge Discovery, 2000; 4(2-3), 163-192.

15. Quentin-Trautvetter J. Devos P. Duhamel A. & Beuscart R. Assessing association rules and decision trees on analysis of diabetes data from the DiabCare program in France. Studies in health technology and informatics. 2002; 90, 557.

16. Huang Y. McCullagh P. Black N. & Harper R. Feature selection and classification model construction on type 2 diabetic patients' data. Artificial intelligence in medicine, 41(3). 2007; 251-262.

17. Han J. Rodriguez J. C. & Beheshti M. Diabetes data analysis and prediction model discovery using

rapidminer. In Future Generation Communication and Networking, 2008. FGCN'08. Second International Conference on. vol. 3, 2008; pp. 96-99. IEEE.

18. Anbananthen K. S. M. Sainarayanan G. Chekima A, & Teo J. Artificial Neural Network Tree Approach in Data Mining. Malaysian Journal of Computer Science, 20 no. 1 ,2007; 51.

19. Cho B. H. Yu H. Kim K. W. Kim T. H. Kim I. Y. & Kim S. I. Application of irregular and unbalanced data to predict diabetic nephropathy using visualization and feature selection methods. Artificial Intelligence in Medicine, 42 no. 1, 2008; 37-53.

20. Chan C. L. Liu Y. C. & Luo S. H. Investigation of diabetic microvascular complications using data mining techniques. In Neural Networks, 2008. IJCNN 2008.(IEEE World Congress on Computational Intelligence). IEEE International Joint Conference on, 2008; pp. 830-834. IEEE.

21. Patil B. M. R. C. J. and Durga T. Association rule for classification of type-2 diabetic patients. In Machine Learning and Computing (ICMLC), 2010 Second International Conference on, 2010; 330-334. IEEE.

22. Aljumah A. A. Ahamad M. G. & Siddiqui M. K. Application of Data Mining: Diabetes Health Care in Young and Old Patients. Journal of King Saud University-Computer and Information Sciences. vol.25, 2012; 127-136.

23. Antonelli D. Baralis E. Bruno G.Cerquitelli T. Chiusano S. & Mahoto N. Analysis of diabetic patients through their examination history. Expert Systems with Applications. Vol.40,2013; 4672–4678.

24. Newman D. J.Hettich S. Blake C. L. & Merz C. J. UCI Repository of machine learning databases, Irvine, CA: University of California, Department of Information and Computer Science.1998, last accessed: 1/10/2009.

25. Chen G. Åstebro T. How to deal with missing categorical data: test of a simple Bayesian method. Organizational Research Methods, 6 no.3, 2003; 309-327.

26. Witten I. H. Frank E. Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann. 2005.

Knowledge Extraction of Diabetics' Data by Decision Tree Method

Ameri H[1]/ Alizade S[2]/ Barzegari A[3]

Abstract

Introduction: In the last 10 years The incidence of diabetes has doubled worldwide with annual increasing rate of about 6%. More than 2 million people in Iran are now affected by this disease. The present research deals with the relation between the observed complications of type 2 diabetic patients and some related features like Blood Glucose Level, Blood Pressure, Age, and Family History. The main purpose was to predict the patients' complications based on the observed signs.
Methods: The research data were gathered from 856 patient records related to the 2009's cases in the Diabetes Center of Golestan province. A new model based on the standard methodology CRISP was developed. In the modeling section, two well-known data mining techniques called C5.0 decision tree and Neural Network were used. Celementine 12.0 software was implemented For data analysis.
Results: The results of data mining showed that the variables of high blood pressure, age, and family history had the most impact on the observed complications. Based on the created decision tree, some rules have been extracted which can be used as a pattern to predict the probability of occurring these complications in the patients. The accuracy of the C5.0 model on the data was shown to be 89.74% and on the Artificial Neural Network was 51.28%.
Conclusion: As the highest accuracy was shown to be achieved using C5.0 algorithm, according to the created rules, it can be predicted which complication(s) any diabetic patient with new specified features may probably suffer from.

Keywords: type 2 diabetic, diabetic complications, Data mining, C5.0 Algorithm, Artificial neural network

● **Received:** 4/June/2013 ● **Modified:** 25/August/2013 ● **Accepted:** 18/Sep/2013

1.MSc in E-Commerce, Information Technology Department, Faculty of Industrial Engineering, KN Toosi University of Technology, Tehran, Iran
2.Assistant Professor of Information Technology Department, Faculty of Industrial Engineering, KN Toosi University of Technology, Tehran, Iran; Corresponding Author (s_alizadeh@kntu.ac.ir)
3.General Physician, Golestan University of Medical Science, Gorgan, Iran