مقاله پژوهشی

مجله انفورماتیک سلامت و زیست پزشکی
مرکز تحقیقات انفورماتیک پزشکی
دوره پنجم شماره دوم. ص۳۰۳-۲۹۳، ۱۳۹۷

تشخیص بیماری دیابت نوع ۲ با استفاده از درخت تصمیم C4.5

حامد صباغ گل[1]*

• **دریافت مقاله:** ۹۶/۱۱/۲۵ • **پذیرش مقاله:** ۹۷/۴/۲۱

مقدمه: یکی از شایع‌ترین بیماری‌ها در دنیای امروز بیماری دیابت است و سالانه شیوع دیابت در سطح جهان حدود درصد افزایش می‌یابد. استفاده از تکنیک‌های داده‌کاوی برای ایجاد مدل‌های پیشگویی کننده، جهت شناسایی افراد در معرض خطر برای کاهش عوارض ناشی از بیماری بسیار کمک‌کننده است. در این پژوهش با استفاده از درخت تصمیم C4.5 به روش‌های پیشگیری و تشخیص این بیماری پرداخته شد.

روش: در این پژوهش کاربردی- توصیفی از داده‌های استاندارد UCI و مجموعه داده pima-indians-diabetes استفاده شد. این پایگاه داده شامل ۷۶۸ رکورد با ۸ فیلد می‌باشد. تجزیه و تحلیل به کمک نرم‌افزار Weka 3.6 با به‌کارگیری روش CRISP3 انجام شد. در بخش مدل‌سازی درخت تصمیم C4.5 با به‌کارگیری متغیرهای ورودی و تعیین متغیر هدف ایجاد شد. همچنین جهت ارزیابی مدل از شاخص‌های حساسیت، ویژگی، دقت، ارزش اخباری مثبت و منفی استفاده شد.

نتایج: با توجه به مدل استفاده شده مشخص شد که به ترتیب متغیرهای میزان بالای قند خون دوساعته، تعداد دفعات بالای حاملگی، سن بالا، فشارخون دیاستولیک بالا، سابقه خانوادگی و شاخص توده بدنی (BMI) بالا، بیشترین تأثیر را در ابتلا به بیماری دیابت نوع ۲ دارا هستند. نرخ دسته‌بندی برابر با ۷۳/۸٪ و دقت الگوریتم C4.5 برابر با ۷۹٪ به‌دست آمد.

نتیجه‌گیری: در مقایسه با نتایج مطالعات انجام شده در حوزه داده‌کاوی بیماری دیابت، دقت به‌دست‌آمده الگوریتم پیشنهادی قابل قبول است. بیشترین عوامل تأثیرگذار بر بیماری دیابت شناسایی شدند. همچنین قوانینی استخراج شد که می‌تواند به عنوان الگویی در جهت پیشگویی احتمال ابتلا افراد به بیماری دیابت استفاده شود.

کلید واژه‌ها: داده‌کاوی، بیماری دیابت نوع ۲، درخت تصمیم C4.5

• **ارجاع:** صباغ گل حامد. تشخیص بیماری دیابت نوع ۲ با استفاده از درخت تصمیم C4.5. مجله انفورماتیک سلامت و زیست پزشکی ۱۳۹۷؛ ۵(۲): ۳۰۳-۲۹۳.

۱. مربی، کارشناسی ارشد مهندسی کامپیوتر، عضو هیات علمی گروه کامپیوتر، دانشگاه پیام نور، بیرجند، ایران

* **نویسنده مسئول:** بیرجند، انتهای بلوار شهید آوینی، دانشگاه پیام نور خراسان جنوبی

• **شماره تماس:** ۰۵۶۳۲۲۰۲۰۲۵

• **Email:** Sabbagh.h@pnu.ac.ir

مقدمه

دیابت بیماری است که در آن سطح قند خون بیشتر از حد طبیعی است. غذا در بدن به نوعی قند که گلوکز نام دارد، تبدیل می‌شود تا به مصرف سلول‌ها برسد. سلول‌ها برای جذب گلوکز نیاز به هورمونی به نام انسولین دارند که در لوزالمعده ساخته می‌شود [۱].

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 مطالعاتی 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 [۱]. با توجه به این‌که بیماری دیابت، به عنوان یک بیماری بسیار مزمن شناخته شده است و آسیب‌های جبران‌ناپذیری به اندام‌ها و اعضاء حیاتی بدن وارد می‌کند، استفاده از ابزارهای هوشمند داده‌کاوی می‌تواند برای بهبود روش‌های شناسایی و کنترل بیماری به پزشکان کمک بزرگی باشد [۱].

دو نوع دیابت وجود دارد، نوع ۱ که وابسته به انسولین نیز نامیده می‌شود و دیابت نوع ۲ که کمبود نسبی انسولین است. دیابت نوع ۲ (دیابت بزرگسالان یا دیابت غیروابسته به انسولین)، یکی از شایع‌ترین انواع دیابت بوده و حدود ۹۰ درصد بیماران دیابتی را تشکیل می‌دهد. برخلاف دیابت نوع ۱، بدن در زمان ابتلا به دیابت نوع ۲ انسولین تولید می‌کند؛ اما یا میزان انسولین تولید شده توسط پانکراس کافی نبوده و یا بدن نمی‌تواند از انسولین تولید شده، استفاده کند. زمانی که انسولین کافی وجود نداشته باشد و یا بدن از انسولین استفاده نکند، گلوکز (قند) موجود در بدن، نمی‌تواند وارد سلول‌های بدن شده و باعث جمع شدن گلوکز در بدن شود و بدن را دچار مشکل و نارسایی نماید [۲].

امروزه حجم داده‌های پزشکی به شدت در حال افزایش است و پزشکان معمولاً اطلاعات ارزشمندی را در خصوص بیماری‌ها و ارتباط آن‌ها با دیگر عوامل ایجاد کننده بیماری‌ها به دست می‌آورند؛ اما این مجموعه داده‌های خام به خودی خود ارزشی ندارند، برای معنی بخشیدن به این داده‌ها باید آن‌ها را تحلیل و تبدیل به اطلاعات یا بهتر از آن‌ها دانش کرد [۳]. با توجه به شیوع بیماری دیابت نوع ۲ در سراسر جهان، استفاده از روش‌های جدید در تحقیقات زیست پزشکی بسیار مورد توجه قرار گرفته است. داده‌کاوی، ابزاری است که برای حصول به چنین دانشی ما را یاری می‌کند. یکی از زمینه‌های پرکاربرد داده‌کاوی در علم پزشکی است؛ استفاده از تکنیک‌های داده‌کاوی برای ایجاد مدل های پیش‌گویی کننده، جهت شناسایی افراد در معرض خطر برای کاهش عوارض ناشی از بیماری بسیار کمک کننده است [۲].

محیط مراقبت سلامت غنی از اطلاعات و ضعیف از دانش است. داده‌کاوی، پتانسیل خوبی برای ایجاد یک محیط غنی از دانش دارد که می‌توانند کمک قابل توجهی به کیفیت تصمیمات بالینی نمایند [۴].

داده‌کاوی پزشکی دارای پتانسیل زیادی برای کشف الگوهای پنهان موجود در داده‌ها را دارا است که این الگوها می‌تواند برای تشخیص‌های بالینی مورد استفاده قرار گیرد [۵]. امروزه استفاده از روش‌های متنوع داده‌کاوی برای شناسایی الگوها و ارتباطات میان متغیرهای مختلف در تولید مدل‌های پیش‌بینی‌کننده در علوم پزشکی بسیار مورد توجه قرار گرفته است [۶]. کاربرد روش‌های داده‌کاوی در حوزه‌های مختلف پزشکی مانند تشخیص، پیش‌گویی و حتی درمان به اثبات رسیده است [۷].

یکی از عملکردهای پیش‌گویانه در داده‌کاوی، دسته‌بندی است. از مهم‌ترین روش‌های رایج دسته‌بندی درخت تصمیم می‌باشد و از میان الگوریتم‌های مورد استفاده در ساخت درخت تصمیم، مهم‌ترین آن‌ها الگوریتم C4.5 است [۸،۹].

Breault و همکاران به طبقه‌بندی با استفاده از روش CART(Classification and Regression Trees) پرداختند و وابستگی بین یک سری از ویژگی‌های بیماران را نشان دادند [۱۰]. Fang خوشه‌بندی و رگرسیون را مورد مطالعه قرار داد و به خوشه‌بندی بیماران براساس مبتلا بودن به دیابت پرداخت [۱۱]. Patil و همکاران الگوریتم Apriori را مورد استفاده قرار دادند و در مطالعه خود قوانین تلازمی برای پیداکردن ارتباطات پنهان بین ویژگی‌ها را نشان دادند [۱۲]. Aljumah و همکاران [۱۳] از روش رگرسیون به پیش‌بینی درمان دیابت در دو دسته گروه سنی جوان و پیر بر اساس نوع درمان استفاده نمودند. همچنین Antonelli و همکاران روش خوشه بندی چندسطحی را مورد استفاده قرار دادند که نتایج حاصل کمک شایانی به شناسایی مسیر درمان بیماران دیابتی کرد [۱۴]. Anbananthen و همکاران [۱۵] با استفاده از درخت تصمیم و شبکه عصبی مصنوعی به پیش‌بینی وجود دیابت در بیماران پرداختند. علاوه بر این Gandhi و همکاران [۱۶] به تشخیص بیماری دیابت با استفاده از روش F-score و SVM پرداختند. Aslam و همکاران [۱۷] به تشخیص بیماری با استفاده از

برنامه نویسی ژنتیک پرداختند. Han و همکاران [۱۸] به ارائه یک مدل درخت تصمیم ID3 با استفاده از Rapidminer جهت تشخیص بیماری دیابت پرداختند. همچنین با نگاهی کلی به مطالعات انجام شده در این حوزه، می‌توان کاربرد داده‌کاوی در دیابت را به چهار دسته کلی تقسیم کرد:

الف) ارتباط ویژگی‌های فردی و شاخص‌های خونی

ب) تشخیص و پیش‌بینی دیابت

ج) تعیین میزان دارو و نوع درمان

د) پیش بینی بروز عوارض

هدف اصلی در این پژوهش استفاده از درخت تصمیم C4.5 بر روی مجموعه داده استاندارد بیماران دیابتی pima-indians-diabetes [۱۹] برای پیش‌بینی و تشخیص بیماری دیابت نوع ۲ بر اساس داده‌های مؤسسه ملی دیابت و بیماری‌های گوارشی و کلیه و ارائه یک مدل به منظور انجام غربالگری پیشگیرانه جهت کاهش عوارض ناشی از بیماری است.

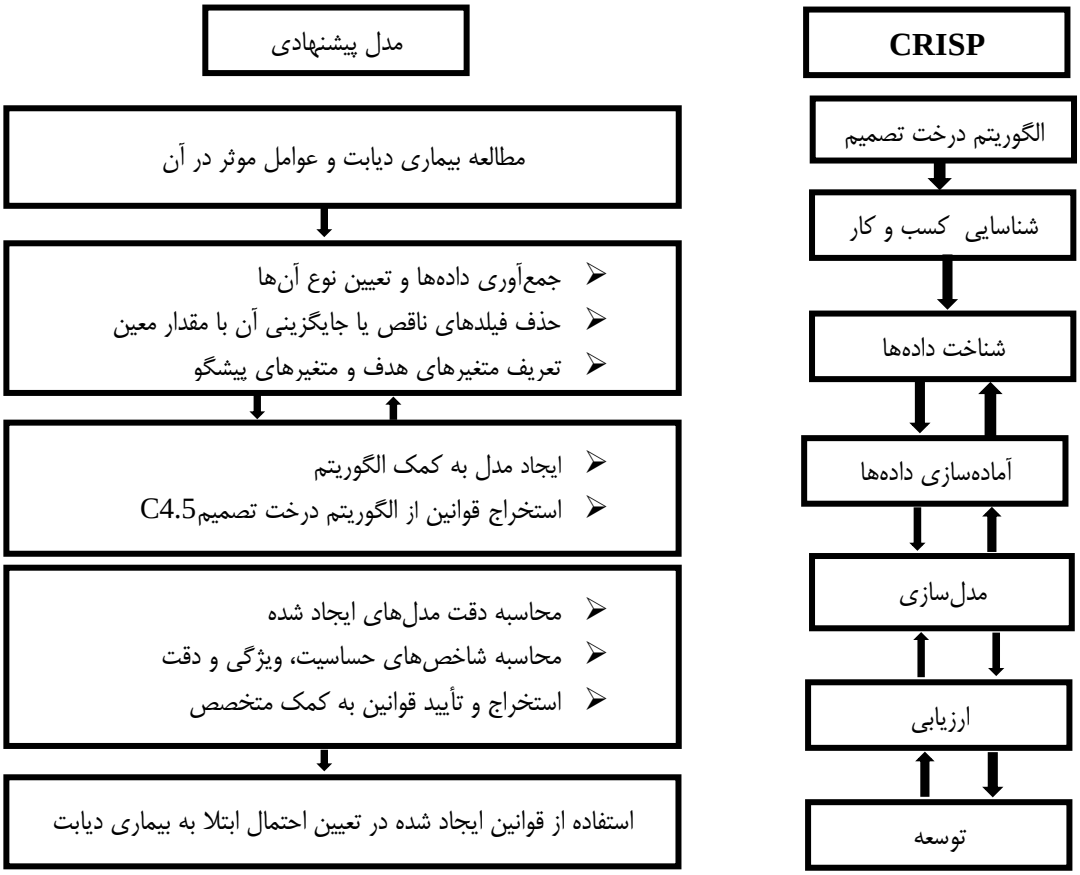
می‌توان با استفاده از نتایج به دست آمده از این پژوهش، پیشنهاداتی به متخصص بالینی جهت تشخیص سریع‌تر و در صورت لزوم کاهش آسیب‌های ناشی از روش‌های تشخیص بیماری‌های دیابتی ارائه نمود.

درخت تصمیم

درخت تصمیم یکی از روش‌های قوی و متداول برای دسته‌بندی و پیش‌بینی است. در واقع درخت‌های تصمیم بالا به پایین رایج‌ترین تکنیک دسته‌بندی هستند و از مهم‌ترین دلایل رایج بودن‌شان می‌توان شفاف بودن، قابل فهم، انعطاف‌پذیری و پردازش نسبتاً سریع ساختار آن‌ها را نام برد. پیش‌بینی به‌دست آمده از درخت در قالب یک سری قواعد توضیح داده می‌شود. در این درخت هر گره داخلی شامل سؤالی بر مبنای یک متغیر مشخص و یک فرزند برای هر پاسخ ممکن بوده و هر برگ با یکی از کلاس‌های ممکن برچسب‌گذاری می‌شود [۲۰]. درخت تصمیم جهت دسته‌بندی یک نمونه با شروع از ریشه مسیری را بر اساس سؤالات مطرح شده در گره‌های داخلی و پاسخ‌های آن دنبال می‌کند تا زمانی که به یک برگ برسد، در نهایت برچسب مربوطه کلاس نمونه موردنظر خواهد بود. اغلب الگوریتم‌های یادگیری درخت تصمیم بر پایه یک عمل جستجوی بالا به پایین عمل می‌کنند [۶].

روش

روش‌های متعددی برای اجرای پروژه‌های داده‌کاوی وجود دارد که یکی از روش‌های قدرتمند در این زمینه استفاده از متدولوژی کریسپ (Cross-industry standard process for data mining)CRISP-DM می‌باشد [۲۱]. این پژوهش نیز بر اساس این روش تنظیم شد. شکل ۱ در ادامه به بررسی هریک از این مراحل در جهت رسیدن به مدلی برای پیش‌گویی احتمال ابتلا به بیماری دیابت می‌پردازد.



شکل ۱: گام‌های متدولوژی کریسپ و چارچوب استفاده شده در این مطالعه

الف) شناخت سیستم

به کارگیری موفق داده‌کاوی مستلزم شناخت حوزه‌ای است که قرار است داده‌کاوی در آن به کار برده شود و علاوه بر آن شناخت کافی از روش‌ها و ابزارهای داده‌کاوی نیز لازم است. به طور کلی تیم داده‌کاوی بایستی دانش کافی در حوزه‌ای که قرار است بررسی شود، داشته باشند. در گام اول پژوهش با مشورت پزشک متخصص و نیز با مطالعه بر روی بیماری دیابت نوع ۲ و تعیین فاکتورهای مؤثر در ابتلا و همچنین روش‌های تشخیصی و درمانی و روش‌های پیشگیری از ابتلا به بیماری، سعی در شناخت کافی حوزه مورد بررسی می‌باشد.

ب) آماده‌سازی داده‌ها

در این پژوهش از مجموعه داده دیابت pima-indians-diabetes [۱۹] استفاده شده است. علائم زیادی از بیماری دیابت وجود دارد، یافتن الگوهایی از داده بیماری دیابت در تشخیص دلایل آینده این بیماری کمک می‌کند. پایگاه داده بیماری دیابت توسط مرکز پزشکی National Institute of Diabetes and Digestive and Kidney Diseases در سال ۱۹۹۰ ایجاد و در سال ۲۰۱۱ به روزرسانی شده است. تمامی بیماران زنان با سن حداقل ۲۱ می‌باشند [۱۹].

پایگاه داده موردنظر شامل ۷۶۸ رکورد با ۸ فیلد است. این پایگاه داده ۸ متغیر خام دارد و تمامی آزمایش‌ها بر روی این ۸ متغیر انجام شده است؛ بنابراین، این پایگاه داده شامل ۸ علامت بیماری و یک متغیر تشخیص است که فیلد هدف به وجود بیماری دیابت بر اساس علائم موجود در بیمار اشاره دارد که یک مقدار عددی ۰ (عدم ابتلا به بیماری دیابت) یا ۱ (ابتلا به بیماری دیابت) است که در ادامه مفهوم هر کدام از علائم بیان می‌شود. در این مرحله داده‌هایی که در حال حاضر در دسترس هستند و داده‌هایی که برای ساخت مدل نیاز بود، تعیین شدند. برای شروع پژوهش از طریق مطالعات کتابخانه‌ای و مشاوره بالینی و براساس داده‌های مجموعه داده pima-indians-diabetes، متغیرهایی که بیشترین تأثیر را در ابتلای افراد به بیماری دارا بودند، تعیین شد که در جدول ۱ آورده شد.

متغیرهای تعیین شده برای ایجاد مدل به دو دسته متغیرهای هدف و متغیرهای پیشگو دسته‌بندی شدند که متغیر هدف ابتلا یا عدم ابتلا و سایر متغیرها به عنوان متغیر پیشگو مورد استفاده قرار گرفتند.

جدول ۱: متغیرهای اطلاعاتی مورد استفاده

ردیف	نام متغیر	توضیحات	مجموعه مقادیر	نوع داده ای
۱	Preg	تعداد دفعات حاملگی	۰=<Preg<=۱۷	عددی صحیح
۲	Plas	قند خون دو ساعت بعد صبحانه	۰=<Plas<=۱۹۹	عددی صحیح
۳	Pres	فشار خون دیاستولیک	۰=<Pres<=۱۲۲	عددی صحیح
۴	Skin	ضخامت عضله سر بازو	۰=<Skin<=۹۹	عددی صحیح
۵	Insu	انسولین سرم	۰=<Insu<=۸۴۶	عددی صحیح
۶	Mass	شاخص توده بدنی	۰=<Mass<=۶۷/۱	عددی اعشاری
۷	Pedi	سابقه خانوادگی	۲/۴۲=<Pedi<=۰/۰۷۸	عددی اعشاری
۸	Age	سن (سال)	۸۱=<Age<=۲۱	عددی صحیح
۹	Class	کلاس	۱ دیابت دارد ۰ دیابت ندارد	عددی دودویی

ج) مدل‌سازی

روش‌های داده‌کاوی متنوعی برای مدل‌سازی وجود دارد. در این مرحله با استفاده از الگوریتم درخت تصمیم C4.5 به ارائه مدل پیش‌گویانه پرداخته شد. مدل‌سازی با استفاده از نرم‌افزار Weka 3.6 انجام شد. در این پروژه از ۷۶۸ رکورد معتبر آن استفاده شد. در ادامه الگوریتم درخت تصمیم C4.5 با به‌کارگیری متغیرهای ورودی و تعیین متغیر هدف ایجاد شد. برای ساخت مدل درخت تصمیم متغیرهای پیشگو متغیر هدف تعیین گردید. در مرحله بعد داده‌ها به دو دسته آموزش (۸۰ درصد) و آزمون (۲۰ درصد) تقسیم شدند. داده‌های بخش آموزش مدل را می‌سازند و داده های بخش آزمون مدل ایجاد شده را مورد ارزیابی قرار می‌دهند. یک درخت تصمیم ترکیب تعدادی استلزام منطقی (قانون اگر- آنگاه) است. معمولاً مجموعه قوانین استخراج شده از درخت تصمیم، مهم‌ترین اطلاعاتی است که از آن‌ها به دست می‌آید. در مدل ایجاد شده در این نرم‌افزار به منظور تقسیم شاخص‌ها از شاخص جینی استفاده شده است [۲۲]. دلیل انتخاب این مدل نیز به این جهت بود که با محاسبه شاخص‌های موردنظر دارای بالاترین دقت در بین مدل‌های اجرا شده بود. نحوه محاسبه شاخص‌ها در بخش تجزیه‌وتحلیل درخت تصمیم ارائه شد.

د) ارزیابی مدل

در این مرحله پس از ایجاد مدل به ارزیابی مدل ایجاد شده، پرداخته شد. جهت ارزیابی مدل‌ها می‌توان از شاخص های حساسیت، ویژگی، دقت، ارزش اخباری مثبت و ارزش اخباری منفی استفاده کرد. جهت محاسبه این شاخص‌ها از ماتریس تداخلی ایجاد شده در محیط نرم‌افزار استفاده شد. در ادامه پژوهش؛ دقت، حساسیت، ویژگی، مقدار پیش‌بینی مثبت و مقدار پیش‌بینی منفی بررسی و محاسبه شد. که هرکدام به صورت زیر محاسبه می‌شوند:

$$\text{Classification rate}=\frac{TP+TN}{TP+FN+FP+TN}$$

$$\text{Precision}=\frac{TP}{TP+FP}$$

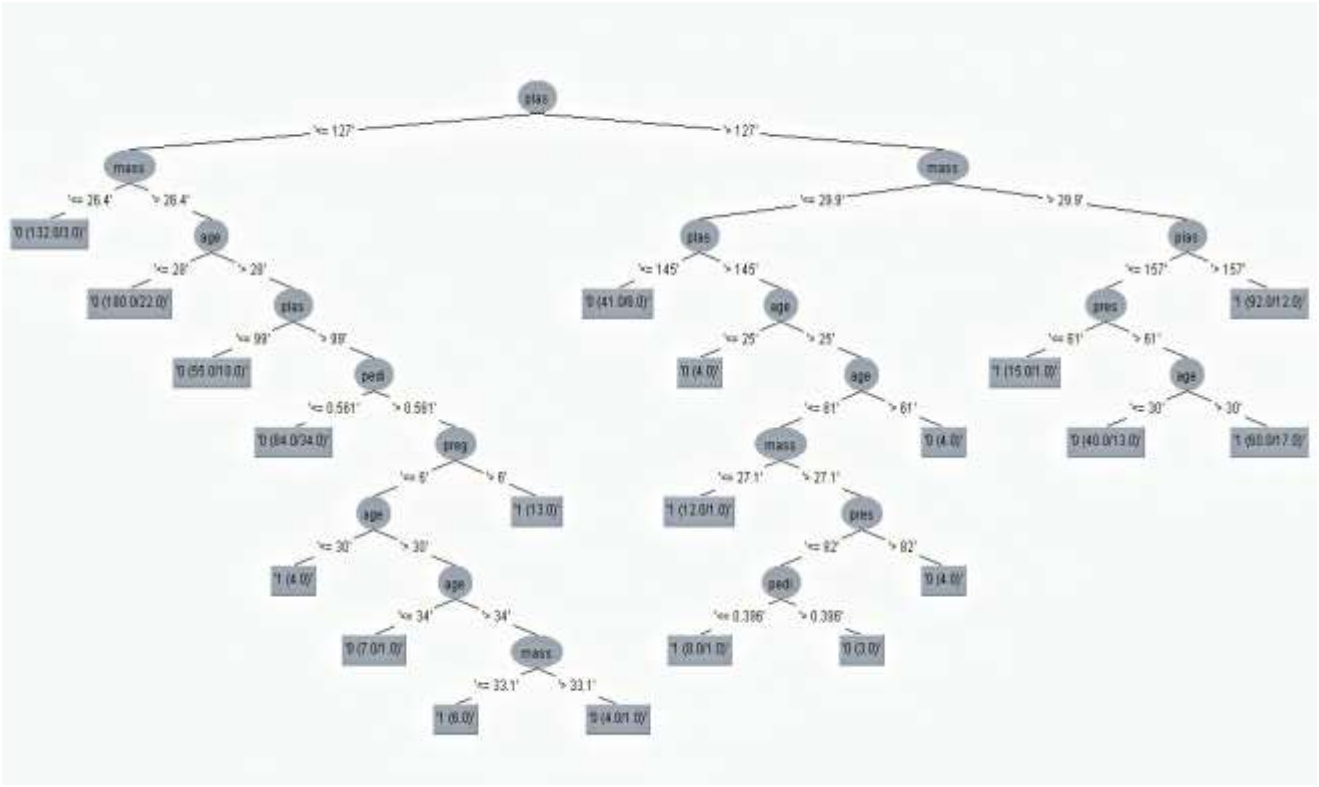
$$\text{Recall}=\frac{TP}{TP+FN}$$

$$\text{TPR}=\frac{TP}{TP+FP}$$

$$\text{FPR}=\frac{FP}{FP+TN}$$

نتایج

با توجه به مدل‌های ا ستفاده شده م شخص شد که به ترتیب میزان بالای قند خون دوساعت بعد صبحانه، تعداد دفعات بالای حاملگی، سن بالا، ف شارخون دیا ستولیک بی شتر از ۸۲، سابقه خانوادگی و شاخص توده بدنی (BMI(Body Mass Index بالا، بی شترین تأثیر را در ابتلا به بیماری دیابت دارا ه ستند. به کمک درخت ت صمیم ایجاد شده، قوانینی ا ستخراج شده ا ست که می‌تواند به عنوان الگویی در جهت پیشگویی احتمال ابتلا افراد به بیماری دیابت استفاده شود. درخت تصمیم حاصل شده در شکل ۲ نشان داده شد.



شکل ۲: بخشی از درخت تصمیم ایجاد شده

همچنین در جدول ۲ تعدادی از قوانین ایجاد شده براساس مشاوره بالینی و نتایج حاصله توسط مدل C4.5 بیان شد.

جدول ۲: تعدادی از قوانین ایجاد شده توسط الگوریتم C4.5

ردیف	قوانین
۱	اگر۱۲۷=> قندخون دوساعت بعد صبحانه و ۲۶/۴=> شاخص توده‌بدنی باشد آنگاه فرد دچار دیابت نیست.
۲	اگر۱۲۷=> قندخون دوساعت بعد صبحانه و شاخص توده‌بدنی>۲۶/۴ و۲۸=> سن باشد آنگاه فرد دچار دیابت نیست.
۳	اگر قندخون دوساعت بعد صبحانه > ۹۹ و شاخص توده‌بدنی>۲۶/۴ و سن> ۲۸ و۰،۵۶۱=> سابقه خانوادگی باشد آنگاه فرد دچار دیابت نیست.
۴	اگر قندخون دوساعت بعد صبحانه > ۹۹ و شاخص توده‌بدنی>۲۶/۴ وسن> ۲۸ و سابقه خانوادگی > ۰،۵۶۱ و تعداد دفعات حاملگی >۶ باشد آنگاه فرد دچار دیابت است.
۵	۱۴۵=> اگر قند خون دوساعت بعد صبحانه > ۱۲۷ و ۲۹،۹=> شاخص توده بدنی باشد فرد دیابت ندارد.
۶	اگر قند خون دوساعت بعد صبحانه > ۱۴۵ و ۲۹/۹=> شاخص توده بدنی و ۲۵=> سن باشد فرد دیابت ندارد.
۷	اگر قند خون دوساعت بعد صبحانه > ۱۴۵ و ۲۹/۹=> شاخص توده بدنی و سن > ۶۱ باشد فرد دیابت ندارد.
۸	اگر قند خون دوساعت بعد صبحانه > ۱۴۵ و ۲۷/۱=> شاخص توده بدنی و۶۱=> سن > ۲۵ باشد فرد دیابت دارد.
۹	اگر قند خون دوساعت بعد صبحانه > ۱۵۷ و شاخص توده بدنی > ۲۹/۹ باشد فرد دارای دیابت است.
۱۰	اگر ۱۵۷=> قند خون دوساعت بعد صبحانه > ۱۲۷ و شاخص توده بدنی > ۲۹/۹ و۶۱=> فشار خون دیاستولیک باشد فرد دارای دیابت است.
۱۱	اگر ۱۵۷=> قند خون دوساعت بعد صبحانه > ۱۲۷ و شاخص توده بدنی > ۲۹/۹ و فشار خون دیاستولیک > ۶۱ و۳۰ => سن باشد فرد دارای دیابت نیست .
۱۲	اگر ۱۵۷=> قند خون دوساعت بعد صبحانه > ۱۲۷ و شاخص توده بدنی > ۲۹/۹ و فشار خون دیاستولیک > ۶۱ وسن > ۳۰ باشد فرد دارای دیابت است.
۱۳	اگر قندخون دوساعت بعد صبحانه > ۹۹ و شاخص توده‌بدنی>۲۶،۴ و۳۰=> سن > ۲۸ و سابقه خانوادگی > ۰/۵۶۱ و۶=> تعداد دفعات حاملگی باشد آنگاه فرد دچار دیابت است.
۱۴	اگر قندخون دوساعت بعد صبحانه > ۹۹ و ۳۳/۱=> شاخص توده‌بدنی>۲۶،۴ وسن> ۳۴ و سابقه خانوادگی > ۰،۵۶۱ و تعداد دفعات حاملگی >۶ باشد آنگاه فرد دچار دیابت است.
۱۵	اگر قندخون دوساعت بعد صبحانه > ۹۹ و شاخص توده‌بدنی> ۳۳،۱ وسن> ۳۴ و سابقه خانوادگی > ۰/۵۶۱ و تعداد دفعات حاملگی >۶ باشد آنگاه فرد دارای دیابت است.

در ادامه به تجزیه و تحلیل درخت تصمیم و بحث در مورد قسمت‌های مختلف آن پرداخته شد.

الف) براساس ساختار درخت تصمیم براساس قند خون دوساعت بعد صبحانه، همان‌طور که در شکل ۲ مشاهده شد:

- اگر فرد قندخون دوساعت بعد از صبحانه آن بالای ۱۵۷ باشد به بیماری دیابت مبتلا است.
- افرادی که قند خون دو ساعت بعد صبحانه آن‌ها کمتر از ۱۵۷باشد و فشارخون دیاستولیکشان کمتر از ۶۱ باشد از سایر افراد بیشتر در معرض بیماری دیابت قرار دارند.
- افرادی که فشارخون دیاستولیک بالای ۶۱ دارند و سن آن‌ها کمتر از ۳۰ سال باشد ، بیماری دیابت ندارند.
- و افرادی که فشار خون دیاستولیک بالای ۶۱ دارند و سن آن‌ها بیشتر از ۳۰ سال باشد، بیماری دیابت دارند.

ب) همچنین براساس ساختار درخت تصمیم براساس تعداد دفعات حاملگی، همان‌طور که در شکل ۲ مشاهده شد:

- خانم‌هایی که تعداد دفعات حاملگی آن‌ها بیشتر از ۶ بار است. در اولویت بیماری دیابت هستند.
- اگر تعداد حاملگی، کمتر از ۶ بار باشد و سن آن‌ها کمتر از ۳۰ سال باشد نسبت به سایرین بیشتر در معرض بیماری دیابت قرار دارند.

ج) براساس ساختار درخت تصمیم براساس سن همان‌طور که در شکل ۲ مشاهده شد:

- افرادی که بزرگ‌تر از ۳۴ سال باشند و شاخص توده بدنی‌شان کمتر از ۳۳/۱ باشد با احتمال بالاتری مبتلا به بیماری دیابت خواهند شد و افرادی که شاخص توده بدنی آن‌ها بیشتر از ۳۳/۱ باشد، دیابت ندارند.

د) همچنین براساس ساختار درخت تصمیم براساس سابقه خانوادگی، مشاهده می‌گردد:

- افرادی که فشارخون دیاستولیک آن‌ها بیشتر از ۸۲ می‌باشد بیماری دیابت ندارند.
- و افرادی که فشارخون دیاستولیک آن‌ها کمتر از ۸۲ می‌باشد و سابقه خانوادگی دارند باشد با احتمال بالاتری مبتلا به بیماری دیابت خواهند شد.
- و افرادی که فشارخون دیاستولیک آن‌ها کمتر از ۸۲ باشد و سابقه خانوادگی ندارند، با احتمال کمتری مبتلا به بیماری دیابت خواهند شد.

با توجه به ارزیابی انجام شده نتایج زیر حاصل گردید:

TP=۴۰۷, TN=۱۶۰, FN= 93, FP=۱۰۸

$$\text{Classification rate} = \frac{407+160}{407+93+108+160} = 0.738$$

$$\text{Precision} = \frac{407}{407+108} = 0.79$$

$$\text{Recall} = \frac{407}{407+93} = 0.814$$

$$\text{TP Rate} = \frac{407}{93+407} = 0.814$$

$$\text{FP Rate} = \frac{108}{108+160} = 0.403$$

همان‌طور که مشاهده شد الگوریتم مورد استفاده در این مطالعه، الگوریتم C4.5 بود که دارای میزان دقت قابل قبول ۷۹٪ می‌باشد. در مرحله ارزیابی نظر متخصص موردنظر نیز در مورد قوانین ایجاد شده، اعمال می‌گردد. به این ترتیب که قوانین به دست آمده به متخصص موردنظر ارائه شده و قوانینی که از نظر بالینی معتبر باشند به عنوان قوانین نهایی ارائه گردیدند. همچنین ریسک فاکتورهای: میزان بالای قند خون دوساعت بعد صبحانه، تعداد دفعات بالای حاملگی، سن بالا، فشارخون دیاستولیک بیشتر از ۸۲، سابقه خانوادگی و شاخص توده بدنی (BMI) بالا به ترتیب بیشترین تأثیر در ابتلا به بیماری دیابت را دارا هستند؛ این در حالی است که براساس مقایسه‌های انجام شده براساس اولویت‌بندی متغیرها توسط الگوریتم‌های موردبررسی نیز این متغیرها جزء فاکتورهای اول قرار گرفته‌اند که نشان از اهمیت این متغیرها دارد. برخی از مهم‌ترین موارد این قوانین به شرح ذیل می‌باشد:

۱- در ۸۱/۲٪ از بیماران، قند خون ۲ ساعت بعد صبحانه بیشتر از۱۵۷ واحد و ابتلا به بیماری دیابت با هم مشاهده گردید.

۲- در ۵۶٪ از افراد با حاملگی بیشتر از ۶ بار ابتلا به بیماری دیابت مشاهده گردید.

۳- در ۴۶/۶٪ از بیماران با فشارخون دیاستولیک بالای ۸۲ ابتلا به بیماری دیابت مشاهده گردید.

۴- همچنین در ۴۲/۷٪ از بیماران با شاخص توده بدنی بیشتر از ۲۷ واحد ابتلا به بیماری دیابت مشاهده گردید و فقط در ۱۰٪/۲ از بیماران با شاخص توده بدنی کمتر از ۲۷ ابتلا به دیابت مشاهده گردید.

۵- در ۴۰/۲٪ از بیماران با انسولین سرم بالای ۱۶۶ ابتلا به بیماری دیابت مشاهده گردید.

۶- در ۴۹/۶٪ از بیماران با سن بالای ۳۴ سال ابتلا به بیماری دیابت مشاهده گردید.

۷- در ۴۵/۸٪ از بیماران با سابقه خانوادگی با احتمال بیشتر از ۰/۵۶۱ بیماری دیابت مشاهده گردید.
۸- در ۲۹/۸٪ از افراد با دارا بودن ضخامت عضله سربازو کمتر از ۲۹ به بیماری دیابت مشاهده گردید.

بحث و نتیجه‌گیری

در این پژوهش، با استفاده از الگوریتم درخت تصمیم C4.5 به ارائه مدل و استخراج قوانین آن در راستای پیشگویی احتمال ابتلا به بیماری دیابت پرداخته شد. از درخت تصمیم C4.5 نتایج قابل قبولی به دست آمد که دقت آن ۷۹٪ بود و حساسیت آن یعنی تعداد نمونه‌هایی که به درستی عدم وجود بیماری دیابت را نشان داده‌اند، نسبت به کل نمونه‌هایی که واقعاً بیماری دیابت ندارند، ۸۱/۴٪ می‌باشد.

همچنین با توجه به محاسبات انجام شده نرخ دسته‌بندی برابر با ۶/۷۲٪ و مقدار پیش بینی مثبت، ۸۱/۴٪ و مقدار پیش‌بینی منفی، ۴۰/۳٪ به دست آمد که در مقایسه با نتایج مطالعات انجام شده [۱۹] و جدول ۳ در حوزه داده‌کاوی بیماری دیابت با الگوریتم درخت تصمیم، دقت به دست آمده الگوریتم پیشنهادی، قابل قبول است.

همچنین با بررسی دقیق‌تر یافته های حاصل از مدل موردنظر مان بیشترین عوامل تأثیر گذار در ابتلا به‌ترتیب متغیرهای: میزان بالای قند خون دوساعت بعد صبحانه، تعداد دفعات بالای حاملگی، سن بالا، فشارخون دیاستولیک بیشتر از ۸۲، سابقه خانوادگی و شاخص توده بدنی(BMI) بالا می‌باشد. با استفاده از قوانین به دست‌آمده برای یک فرد جدید با داشتن متغیرهای مشخص، می‌توان تعیین کرد که احتمال ابتلای وی به بیماری دیابت چقدر خواهد بود. در جدول ۳ به مقایسه نتایج پژوهش مشابه با پژوهش حاضر آورده شد.

جدول ۳: مقایسه نتایج مطالعات انجام شده در حوزه داده‌کاوی در بیماری دیابت

نویسندگان و سال ارائه تحقیق	روش مورداستفاده	دقت مدل	حساسیت مدل	تأثیرگذارترین ویژگی ها
Miyaki و همکاران (۲۰۰۲) [۲۳]	CART	محاسبه نشده	محاسبه نشده	سن، BMI و فشارخون سیستولیک
Chan و همکاران (۲۰۰۸) [۲۴]	دسته بندی ماشین بردار و رگرسیون	٪۹۱	٪۸۱	فشارخون و شاخص BMI بالا، میکرو آلبومین و گلبول سفید
Cho و همکاران (۲۰۰۸) [۲۵]	درخت تصمیم و شبکه عصبی مصنوعی	محاسبه نشده	۶۷٪/۱	کراتینین، سن، سابقه خانوادگی، چربی مضر
Kumari و همکاران (۲۰۱۴) [۲۷]	شبکه بیزین	٪۷۸	۸۲٪	تعداد دفعات حاملگی، سن بالا، فشار خون، BMI و ضخامت عضله سربازو
Rakshit و همکاران (۲۰۱۷) [۲۶]	شبکه عصبی مصنوعی	۷۳٪	٪۷۴/۳	میزان بالای قند خون دوساعته، تعداد دفعات حاملگی، سن بالا، فشارخون دیاستولیک
مدل پیشنهادی	درخت تصمیم C4.5	۷۹٪	٪۸۱/۴	میزان بالای قند خون دوساعت بعد صبحانه، تعداد دفعات بالای حاملگی، سن بالا، فشارخون دیاستولیک، BMI بالا

مطابق مطالعات گذشته، عملکرد مدل‌های طبقه‌بندی کننده ممکن است بر روی پایگاه‌های داده مختلف نتایج متفاوتی داشته باشد. برای مثال Miyaki و همکاران مهم‌ترین ویژگی‌های تأثیرگذار را سن، BMI و فشارخون سیستولیک معرفی کردند و از روش CART استفاده نمودند [۲۳]. Chan و همکاران در پژوهش خود از دسته بندی ماشین بردار پشتیبان و رگرسیون استفاده نمود و تأثیرگذارترین ویژگی‌ها: فشارخون، شاخص BMI بالا، میکروالبومین و گلبول سفید می‌باشد [۲۴] که دارای ویژگی‌های مشترکی با مدل مطالعه حاضر نیز می‌باشد و یافته-های آن تا حدودی مشابه با قوانین استخراج شده از الگوریتم درخت تصمیم در مطالعه حاضر بوده، ضمن این که درخت تصمیم ارائه شده در آن مطالعه از دقت بالاتری برخوردار بوده است. در پژوهش انجام شده Cho و همکاران برای شناسایی عوامل مؤثر، ویژگی‌های کراتینین، سن، سابقه خانوادگی، و چربی مضر بررسی شدند. وی از درخت تصمیم و شبکه عصبی مصنوعی در مطالعه خود استفاده نمود [۲۵]. دقت مدل پیشنهادی این مطالعه بالاتر می‌باشد.

در پژوهش انجام شده توسط Rakshit و همکاران [۲۶] جهت پیش‌بینی احتمال ابتلا به بیماری دیابت نوع ۲ از شبکه عصبی مصنوعی استفاده نمود که دارای دقت ۷۳٪ بوده است. در این

مطالعه نیز تأثیر گذارترین ویژگی ها میزان بالای قند خون دوساعته، تعداد دفعات حاملگی، سن بالا و فشارخون دیاستولیک بالا به دست آمد. در پژوهش Kumari و همکاران با استفاده از شبکه بیزین به تشخیص دیابت پرداخته شده و این مدل دارای دقت ۷۸٪ به عنوان مدل مناسبی جهت پیشگویی معرفی شده است [۲۷].

درنتیجه همان طور که در جدول ۳ مشاهده شد دقت ارائه شده در این پژوهش در مقایسه با سایر پژوهش‌های انجام شده بالا و قابل قبول بوده و می‌تواند در طراحی مدل‌های مناسب جهت پیشگویی امکان ابتلای افراد به بیماری‌های دیابت استفاده شود. همچنین می‌تواند در برنامه‌های غربالگری جهت شناسایی افراد در معرض خطر استفاده شود. ضمناً با توجه به این که نتایج تحقیق بر روی داده‌های استاندارد صورت گرفت، این نتایج می‌تواند به عنوان مبنای ارزیابی برای پژوهش‌های آینده قرار گیرد. همچنین باتوجه به محدودیت دسترسی به مجموعه داده های واقعی بیماران، عدم دسترسی به تعداد نمونه‌های کافی و ثبت نبودن اطلاعات کامل مربوط به بیماران واقعی پیشنهاد می شود که این مدل با مجموعه داده‌های واقعی و در بازه زمانی طولانی اجرا شده و پس از رسیدن به سطح دقت مطلوب در برنامه‌های غربالگری مورد استفاده قرار گیرد. آنگاه پس از ایجاد تغییرات ضروری به عنوان مدل مناسب جهت پیشگویی بیماری دیابت مورد استفاده قرار گیرد. علاوه بر این می‌توان از سایر ریسک فاکتورها مثل: مصرف دخانیات وکشیدن سیگار، سبک زندگی، رژیم غذایی و... نیز استفاده نمود.

تضاد منافع

در این پژوهش تضاد منافع وجود ندارد.

References

1. Mohamed EI, Linder R, Perriello G, Di Daniele N, Poppl SJ, De Lorenzo A. Predicting type 2 diabetes using an electronic nose-based artificial neural network analysis. Diabetes Nutr Metab 2002; 15(4):215-21.

2. Pickup JC. Williams G. Textbook of Diabetes. 3th ed. Oxford: Wiley-Blackwell; 2003.

3. Al Jarullah AA. Decision tree discovery for the diagnosis of type II diabetes. International Conference on Innovations in Information Technology; 2011 Apr 25-27; Abu Dhabi, United Arab Emirates: IEEE; 2011. p. 303-7.

4. Khajehei M. Etemady F. Data Mining and Medical Research Studies. Second International Conference on Computational Intelligence, Modelling and Simulation; 2010 Sep; 28-30; Tuban, Indonesia: IEEE; 2010. p. 119-22.

5. Jayalakshmi T. Santhakumaran A. A novel classification method for diagnosis of diabetes mellitus using artificial neural networks. International Conference on Data Storage and Data Engineering; 2010 Feb 9-10; Bangalore, India: IEEE; p. 159-63.

6. Ghazanfari M, Alizadeh S, Teimorpour B. Data Mining and Knowledge Discovery. 2th ed. Tehran: Publication of Iran University of Science and Technology; 2011. Persian

7. Fayyad UM, Piatetsky-Shapiro G, Smyth P, Uthurusamy R. Advances in Knowledge Discovery and Data Mining. USA: American Association for Artificial Intelligence; 1996.

8. Lavrac N. Selected techniques for data mining in medicine. Artif Intell Med 1999; 16(1):3-23.

9. Han J, Kamber M, Pei J. Data Mining: Concepts and Techniques. 2th ed. USA: Morgan Kaufmann; 2006.

10. Breault JL, Goodall CR, Fos PJ. Data mining a diabetic data warehouse. Artif Intell Med 2002; 26(1-2):37-54.

11. Fang X. Are you becoming a diabetic? Are you becoming a diabetic? A data mining approach. 2009 Aug 14-16; Tianjin, China: IEEE; 2009. p. 18-22.

12. Patil BM, Joshi RC, Toshniwal D. Association rule for classification of type-2 diabetic patients. Second International Conference on Machine Learning and Computing; 2010 Feb 9-11; Bangalore, India: IEEE; 2010. p. 330-4.

13. Aljumah AA, Ahamad MG, Siddiqui MK. Application of data mining: Diabetes health care in young and old patients. Journal of King Saud University - Computer and Information Sciences 2013; 25(2):127-36.

14. Antonelli D, Baralis E, Bruno G, Cerquitelli T, Chiusano S, Mahoto N. Analysis of diabetic patients through their examination history. Expert Systems with Applications 2013; 40(11):4672-8.

15. Anbananthen SK, Sainarayanan G, Chekima A, Teo J. Data Mining using Pruned Artificial Neural Network Tree (ANNT). 2nd International Conference on Information & Communication Technologies; 2006 Apr 24-28; Damascus, Syria: IEEE; 2006. p. 1350-16.

16. Gandhi KK, Prajapati NB. Diabetes prediction using feature selection and classification. International Journal of Advance Engineering and Research Development 2014; 1(5): 1-7.

17. Aslam MW, Nandi AK. Detection of diabetes using genetic programming. 18th European Signal Processing Conference; 2010 Aug 23-27; Aalborg, Denmark: IEEE; 2010. p. 1184-8.

18. Han J, Rodriguez JC, Beheshti M. Diabetes data analysis and prediction model discovery using Rapidminer. Second International Conference on Future

Generation Communication and Networking; 2008 Dec 13-15; Hainan Island, China: IEEE; 2008. p. 96-9.
19. Sigillito VI. Pima-indians-diabetes. Phoenix, AZ: National Institute of Diabetes and Digestive and Kidney Diseases; 1990. Available from: https://gist.github.com/ktisha/c21e73a1bd1700294ef790c56c8aec1f
20. Roiger RJ. Data Mining: A Tutorial-Based Primer. 2th ed. U.S. Florida: CRC Press; 2017.
21. Wirth R, Hipp J. CRISP-DM: Towards a standard process model for data mining. Proceedings of the Fourth International Conference on the Practical Application of Knowledge Discovery and Data Mining; 2000 Apr 11-12; UK: Practical Application Company; 2000. p. 29-39.
22. Fayyad UM, Irani KB. The attribute selection problem in decision tree generation. Proceedings of the Tenth National Conference on Artificial Intelligence; 1992 Jul 12 -16; San Jose, California: AAAI Press; 1992. p. 104-10.
23. Miyaki K, Takei I, Watanabe K, Nakashima H, Watanabe K, Omae K. Novel statistical classification model of type 2 diabetes mellitus patients for tailor-made prevention using data mining algorithm. J Epidemiol 2002;12(3):243-8.
24. Chan CL, Liu YC, Luo SH. Investigation of diabetic microvascular complications using data mining techniques. IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence); 2008 Jun 1-8; Hong Kong, China: IEEE; 2008. p. 830-4.
25. 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. Artif Intell Med 2008; 42(1):37-53.
26. Rakshit S, Manna S, Biswas S, Kundu R, Gupta P, Maitra S, Barman S. Prediction of Diabetes Type-II Using a Two-Class Neural Network. First International Conference, Computational Intelligence, Communications, and Business Analytics; 2017 Mar 24–25; Kolkata, India: 2017. p. 65-71.
27. Kumari M, Vohra R, Arora A. Prediction of diabetes using bayesian network. International Journal of Computer Science and Information Technologies 2014; 5(4): 5174-8.

Original Article

Journal of Health and Biomedical Informatics
Medical Informatics Research Center
2018; 5(2): 293-303

A Detection of Type2 Diabetes using C4.5 Decision Tree

Sabbagh Gol Hamed[1]*

• **Received:** 14 Feb, 2018 • **Accepted:** 12 Jul, 2018

Introduction: One of the most common diseases in the world is diabetes and the global prevalence of diabetes increases by about six percent annually. The use of data mining techniques to create predictive models is very helpful in identifying people at risk and reducing the complications of the disease. In this study, through using decision tree C4.5, methods of prevention and treatment of diabetes were investigated.

Methods: In this applied and descriptive study, we used the standard UCI data and the pima-Indians-diabetes data set. This database contains 768 records with 8 fields. The analysis was done using Weka software using the CRISP3 methodology. In modeling decision tree, C4.5 was created using input variables and determining target variables. Also, the sensitivity, specificity, accuracy, as well as positive and negative predictive values were used to evaluate the model.

Results: According to the model, high blood sugar levels, high gravidity, high age, high diastolic blood pressure, familial history and high BMI have respectively the highest effects on type 2 diabetes mellitus. The ranking rate was 73.8% and the accuracy of the C4.5 algorithm was 79%.

Conclusion: Compared to the results of studies in the field of data mining for diabetes, the accuracy of the proposed algorithm is acceptable. The most effective factors on diabetes were identified. Also, rules were developed that can be used as a model to predict the risk of diabetes in people.

Keywords: Data mining, Type2 diabetes, C4.5 Decision tree

• **Citation:** Sabbagh Gol H. Detection of Type2 Diabetes using C4.5 Decision Tree. Journal of Health and Biomedical Informatics 2018; 5(2): 293-303.

1. M.Sc in Computer Engineering, Faculty of Computer, Computer Engineering Dep., Payame Noor University (PNU), Birjand, Iran

***Correspondence**: Shahid Avini Blvd, South Khorasan Payame Noor University, Birjand Branch. Birjand
• **Tel:** 05632202025 • **Email:** sabbagh.h@pnu.ac.ir