استفاده از تکنیک های داده کاوی جهت تشخیص دیابت با استفاده از چربی خون

رضا رافع*[۱]، محمد اربابی[۲]

۱) گروه مهندسی کامپیوتر، دانشگاه اراک، اراک، ایران

۲) گروه مهندسی کامپیوتر، دانشگاه آزاد اسلامی، واحد ملایر، ملایر، ایران

تاریخ دریافت: ۹۳/۱۰/۱۳ **تاریخ پذیرش:۹۴/۱/۲۳**

چکیده

مقدمه: بیماری دیابت یکی از شایع ترین، خطرناک ترین و پرهزینه ترین بیماری های حال حاضر دنیا است که با نرخ هشدار دهنده ای در حال افزایش است. استفاده از روش های داده کاوی می تواند به تشخیص زودهنگام دیابت کمک کند که باعث جلوگیری از پیشرفت این بیماری و خیلی از عوارض آن مانند بیماری قلب و عروق، مشکلات بینایی و بیماری های کلیوی می شود.

مواد و روش ها: در این تحقیق از نرم افزار داده کاوی رپیدماینر برای مدل سازی به منظور دسته بندی بیماران به دیابتی و غیر دیابتی استفاده شده است. داده های مورد نیاز این تحقیق از پایگاه داده یکی از آزمایشگاه های شهرستان نهاوند استخراج شده است که شامل داده های ۵۷۰۶ بیمار در بازه سال های ۱۳۸۷ تا ۱۳۹۲ است. این داده ها شامل متغیرهای عمومی سن و جنسیت و هم چنین متغیرهای انواع چربی خون و میزان قندخون ناشتا است.

یافته های پژوهش: پس از مدل سازی با استفاده از تکنیک های مختلف دسته بندی بهترین دقت مدل مربوط به مدل درخت تصمیم C4.5 بوده که برابر ۰۲/۹۰ درصد می باشد.

بحث و نتیجه گیری: به منظور تشخیص به موقع دیابت تکنیک های مختلفی با روش ها و متغیرهای گوناگونی ارائه گردیده است. در تحقیق پیش رو نیز با استفاده از رابطه هم افزایی انواع چربی خون با قندخون ناشتا و با استفاده از تکنیک های داده کاوی روشی برای تشخیص دیابت ارائه شده است.

واژه های کلیدی: داده کاوی، دیابت، تکنیک های دسته بندی، درخت تصمیم C4.5

* نویسنده مسئول: گروه مهندسی کامپیوتر، دانشگاه اراک، اراک، ایران

Email: r-rafeh@araku.ac.ir

مقدمه

دیابت چهارمین علت مرگ و میر در بیشتر کشورهای توسعه یافته است(۱). بر اساس آمار فدراسیون جهانی دیابت در سال ۲۰۱۲ بیش از ۳۷۱ میلیون نفر از مردم جهان مبتلا به دیابت بوده که هر سال نیز به آن افزوده می شود به گونه ای که بیش از نیمی از مبتلایان به دیابت از بیماری خود بی خبر هستند. برای بیماران دیابتی بیش از ۴۷۰ میلیارد دلار هزینه می شود. این بیماری عامل مرگ ۵۰ میلیون نفر است. هم چنین بر اساس سرشماری آماری که در سال ۲۰۱۳ صورت گرفته بیش از ۶ میلیون نفر ایرانی مبتلا به دیابت هستند(۲).

داده کاوی راهی است برای تحلیل اتوماتیک داده ها و شناسایی الگوهای پنهان که انجام این امر به صورت دستی ممکن نمی باشد(۳). داده کاوی می تواند در پیش بینی و تشخیص سریع و کم هزینه بیماری ها به طور موثری استفاده شود(۴). اهمیت پیش بینی دیابت از این لحاظ است که بیمار پس از آگاهی از آن می تواند با اصلاح رژیم غذایی و ورزش از اثرات مخربش جلوگیری کند.

در سال ۲۰۰۶ آقای سو و همکاران توانستند بر پایه روش های شبکه عصبی مصنوعی، درخت تصمیم، رگرسیون و قواعد وابستگی بر پایه عکس های سه بعدی و دو بعدی بدن با دقت ۸۹ درصد پیش بینی کنند که آیا فرد مورد نظر به بیماری دیابت نوع دوم مبتلا است یا خیر(۵). هم چنین در سال ۲۰۰۹ پورنامی و همکاران با استفاده از روش ماشین بردار پشتیبان توانستند با دقت ۹۳ درصد بیماری دیابت نوع دوم را در افراد تشخیص دهند. داده های آن ها مربوط به ۷۶۸ بیمار می شد و از هشت متغیر از جمله فشار خون افراد و میزان انسولین تزریقی برای پیش بینی خود استفاده نمودند(۶). با استفاده از همین روش یعنی ماشین بردار پشتیبان، براکات و همکاران نیز یک سال بعد یعنی در سال ۲۰۱۰ توانستند دقت تشخیص بیماری را بهبود دهند و با دقت ۹۴ درصد بیماری دیابت نوع دوم را در افراد تشخیص دهند. آن ها داده های تحقیق خود را از داده های مربوط به ۴۶۸۲ نفر مراجعه کننده استخراج کردند و متغیرهای آن ها شامل: جنسیت، شاخص نسبت وزن به قد(BMI)، فشارخون، کلسترول و قندخون بوده است(۷).

در سال ۲۰۱۰ عده ای از محققان تایلندی با استفاده از درخت تصمیم توانستند با دقت بیش از ۹۰ درصد سندرم متابولیک را در افراد تشخیص دهند. آن ها ازداده های مربوط به ۵۶۳۸ نفر استفاده کردند(۸).

در آزمایشگاه های مختلف پزشکی حجم زیادی از داده ها و اطلاعات وجود دارد و گاهی ارزش این اطلاعات را می توان به اندازه نجات انسانی از مرگ برشمرد. این داده ها و اطلاعات می تواند در تشخیص و پیشگیری از خیلی بیماری ها مفید باشد. آزمایش های تعیین میزان انواع چربی و قندخون از مهم ترین و متداول ترین آزمایش ها می باشد که در تشخیص و پیشگیری خیلی از بیماری ها از جمله بیماری های قلبی، عروقی، مغزی و دیابت می تواند استفاده شود. با توجه به این که تحقیق جاری بر پایه نتایج برخی از فاکتورهای آزمایش های مذکور است در بخش سوم مقاله به تشریح آن ها پرداخته خواهد شد.

اگر میزان قندخون ناشتا(FBS) در آزمایش بزرگ تر یا مساوی ۱۲۶ میلی گرم در هر ۱۰۰ سی سی خون باشد، شخص دیابتی تلقی می شود. با توجه به این تعریف که تحقیق پیش رو بر پایه آن صورت گرفته شخص مبتلا به هر نوع دیابتی که باشد قابل تشخیص و پیش بینی است در حالی که تحقیقاتی که تاکنون در زمینه تشخیص دیابت با استفاده از تکنیک های داده کاوی صورت گرفته تنها قادر به تشخیص یک نوع دیابت بودند و علت آن در روش جمع آوری داده های به کار برده شده در تحقیق بود به گونه ای که داده ها و متغیرهای تحقیق را از پرونده بیماران مبتلا به یک نوع خاص از دیابتی ها و یک سری افراد سالم تهیه می کردند.

هم چنین انواع چربی های خون در رابطه با میزان قندخون اثر هم افزایی دارند(۹). البته متغیرهای عمومی سن و جنسیت نیز در ابتلاء به انواع دیابت نقش اساسی دارند. با توجه به این مطالب در تحقیق جاری از فاکتورهای سن، جنسیت و انواع مقادیر چربی خون افراد برای پیش بینی میزان قندخون و تشخیص نوع دیابت استفاده شده است.

مواد و روش ها

نوع داده ها و متغیرهای مورد استفاده در این تحقیق منحصر به فرد است و از لحاظ تعداد افراد مورد بررسی یعنی ۵۷۰۶ نفر نیز قابل توجه است.

در این مقاله ابتدا انواع تکنیک های دسته بندی در داده کاوی پرداخته شده است و در ادامه مجموعه داده های مورد استفاده و متغیرهای تحقیق تشریح و تحلیل خواهد شد و سپس مدل سازی و ارزیابی انجام خواهد گرفت و در پایان نیز نتیجه گیری آمده است.

انواع تکنیک های دسته بندی داده کاوی: تکنیک دسته بندی یکی از متداول ترین روش های یادگیری مدل به منظور پیش بینی در داده کاوی می باشد. در روش های پیش بینی از مقادیر بعضی از ویژگی ها برای پیش بینی کردن مقدار یک ویژگی مشخص استفاده می شود. دسته بندی فرآیندی برای پیدا کردن مدلی به منظور مشخص کردن کلاس اشیاء با توجه به ویژگی های آن ها می باشد(۱۰).

در الگوریتم های دسته بندی، مجموعه داده اولیه به دو مجموعه داده های آموزشی و مجموعه داده های آزمایشی تقسیم می شود. با استفاده از مجموعه داده های آموزشی مدل ساخته می شود و از مجموعه داده های آزمایشی برای اعتبارسنجی و محاسبه دقت مدل استفاده می شود.

برخی از روش ها و الگوریتم های دسته بندی عبارتند از: ۱-روش های مبتنی بر درخت تصمیم، ۲-قوانین استنتاج، ۳-ماشین های بردار پشتیبان، ۴-روش های مبتنی بر نظریه بیز، ۵-شبکه های عصبی.

روش های مبتنی بر درخت تصمیم: یکی از مشهورترین و پرکاربردترین روش های دسته بندی، استفاده از درخت تصمیم است. در یک درخت تصمیم گره های داخلی شامل شروطی بر روی ویژگی ها می باشد و برگ ها هم برچسب کلاس ها می باشد. برای پیش بینی یک شئ با توجه به مقادیر ویژگی ها و نحوه ارضای شروط گره های داخلی مسیری از ریشه به یک برگ طی می شود که برچسب آن برگ کلاس شئ مورد نظر را تعیین می کند(۱۱).

قوانین استنتاج: قوانین استنتاج شامل پیش شرط هایی است که در سمت چپ هر قاعده قرار می گیرد و برچسب کلاس هم در سمت راست قاعده قرار دارد. برای پیش بینی کلاس یک شئ با توجه به مقادیر ویژگی های آن پیش شرط های قوانین بررسی می شود و هر قانونی که قابل اعمال باشد از برچسب سمت راست آن برای پیش بینی استفاده می شود(۱۲).

ماشین های بردار پشتیبان: مبنای کاری دسته بندی کننده ماشین بردار پشتیبان دسته بندی خطی داده ها است که سعی می کند خطی را انتخاب کند که حاشیه اطمینان بیشتری داشته باشد. حل معادله پیدا کردن خط بهینه برای داده ها به وسیله روش های برنامه ریزی غیر خطی که روش های شناخته شده ای در حل مسائل دارای محدودیت هستند صورت می گیرد(۱۲).

روش های مبتنی بر نظریه بیز: روش بیز بر اساس قانون بیز به دسته بندی اشیاء می پردازد:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

احتمال تعلق شئ با ویژگی های B به کلاس A بر اساس رابطه بالا محاسبه می شود. برای این شئ هر کلاسی که بیشترین مقدار احتمال را داشته باشد در نظر گرفته می شود(۱۲).

شبکه عصبی مصنوعی: شبکه عصبی مصنوعی یک سامانه پردازشی داده ها است که از مغز انسان ایده گرفته و پردازش داده ها را به عهده پردازنده های کوچک و بسیار زیادی سپرده که به صورت شبکه ای به هم پیوسته و موازی با یکدیگر رفتار می کنند تا یک مسئله را حل نمایند. در این شبکه ها به کمک دانش برنامه نویسی، ساختار داده ای طراحی می شود که می تواند همانند نرون عمل کند. بعد با ایجاد شبکه ای بین این نرون ها و اعمال یک الگوریتم آموزشی به آن، شبکه را برای ایجاد رابطه ای بین ویژگی ها و برچسب کلاس آموزش می دهند(۱۳).

تشریح داده های مورد استفاده و متغیرهای تحقیق: قبل از تشریح داده ها لازم است تعاریف پزشکی مربوط به انواع چربی خون و قندخون بیان شود.

چربی خون به دو نوع کلی کلسترول وتری گلیسرید تقسیم می شود. کلسترول خود به دو نوع عمده چربی مضر(LDL) و چربی مفید(HDL) تقسیم

می شود. چربی مضر یا لیپوپروتئین با چگالی کم دارای مقدار زیادی کلسترول و مقدار کمی پروتئین می باشد. وظیفه چربی مضر حمل کلسترول و دیگر چربی ها در خون است. افزایش چربی مضر در خون می تواند باعث باریک و سخت شدن رگ های تغذیه کننده قلب و مغز و به دنبال آن بروز بیماری های قلبی و مغزی شود. چربی مفید دارای مقدار زیادی پروتئین و مقدار کمی کلسترول می باشد که کلسترول را از خون بر می دارد یعنی از رگ ها خارج کرده و به کبد می برد. چربی مفید از قلب محافظت می کند و مقدار کم آن در خون می تواند از عوامل ایجاد بیماری های قلبی باشد. تری گلیسرید، همان چربی است که در غذا وجود دارد. اگر کالری زیادی وارد بدن شود، بدن مقدار اضافی کالری را به تری گلیسرید تبدیل کرده و در سلول های چربی ذخیره می کند. افزایش مقدار تری گلیسرید در خون باعث مسدود شدن سرخرگ ها آسیب رسیدن به لوزالمعده می شود. بنا بر این انسولین به اندازه کافی توسط لوزالمعده تولید نمی شود. چون انسولین سبب کاهش قندخون می شود، فقدان آن باعث افزایش قندخون و در نتیجه بروز دیابت می شود(۱۴).

FBS میزان قند گلوکز خون ناشتا می باشد. همه افراد دارای مقداری قند در خون خود هستند که میزان آن به طور طبیعی در حالت ناشتا بین ۷۰ تا ۱۱۰ میلی گرم در هر ۱۰۰ سی سی خون می باشد.

داده های مورد استفاده در این تحقیق مربوط به اطلاعات یک آزمایشگاه در شهرستان نهاوند می باشد به تفکیک هر سال کار آزمایشگاه از سال ۱۳۸۷ تا ۱۳۹۲ در دسترس ما قرار گرفته است. داده های ۵۷۰۶ بیمار به صورت یک رکورد برای هر بیمار شامل متغیرهای عمومی سن و جنسیت و هم چنین متغیرهای انواع چربی خون و میزان قندخون ناشتا از پایگاه داده مذکور استخراج شدند. جدول شماره ۱ مشخصات متغیرهای تحقیق را نشان می دهد.

جدول شماره ۱. متغیرهای استفاده شده

نام متغیر در مجموعه داده	تشریح فارسی متغیر	نوع متغیر
SEX	جنسیت بیمار	دومقداری
AGE	سن بیمار	عددی
CHOLOSTROL	میزان کلسترول	عددی
LDL	میزان چربی مضر	عددی
HDL	میزان چربی مفید	عددی
TG	میزان تری گلیسیرید	عددی
BFBS	ابتلا به دیابت	دو مقداری

جنسیت بیمار: این متغیر نمایانگر جنسیت بیمار می باشد. در این پایگاه داده ۳۸۱۴ نفر زن و نفر ۱۸۹۲ مرد موجود می باشد.

سن بیمار: این متغیر نمایانگر سن بیمار می باشد. در این پایگاه داده سن بیماران از ۳ تا ۹۲ سال موجود می باشد.

میزان کلسترول: این متغیر نمایانگر میزان کلسترول بیمار است. در این پایگاه داده محدوده آن از ۷۲ تا ۴۸۸ میلی گرم در دسی لیتر است. مقدار کمتر از ۲۰۰ میلی گرم در دسی لیتر نرمال و بالاتر از آن غیر نرمال است.

میزان چربی مضر: این متغیر نمایانگر میزان چربی مضر بیمار است. در این پایگاه داده محدوده آن از ۶ تا ۴۰۵ میلی گرم در دسی لیتر است. مقدار کمتر از ۱۳۰ میلی گرم در دسی لیتر نرمال و بالاتر از آن غیر نرمال است.

میزان چربی مفید: این متغیر نمایانگر میزان چربی مفید بیمار است. در این پایگاه داده محدوده آن از ۱۰ تا ۱۰۷ میلی گرم در دسی لیتر است. مقدار بزرگ تر یا مساوی ۳۵ میلی گرم در دسی لیتر نرمال و کمتر از آن غیر نرمال است.

میزان تری گلیسرید: این متغیر نمایانگر میزان تری گلیسرید بیمار است. در این پایگاه داده محدوده آن از ۲۶ تا ۱۲۸۵ میلی گرم در دسی لیتر است. مقدار کمتر از ۲۰۰ میلی گرم در دسی لیتر نرمال و بیشتر از آن غیر نرمال است.

میزان قندخون ناشتا : در این پایگاه داده محدوده میزان قندخون ناشتا از ۲۹ تا ۵۱۰ میلی گرم در دسی لیتر است. این متغیر ابتلاء یا عدم ابتلاء فرد به دیابت را تعیین می کند. افرادی که مقدار قندخون ناشتا آن ها کمتر از ۱۲۶ میلی گرم دردسی لیتر است فرد غیر دیابتی و در غیر این صورت فرد دیابتی محسوب می شود.

تحلیل داده های مورد استفاده و متغیرهای تحقیق: در این مرحله با تحلیل آماری و مصور سازی داده های تحقیق به یک دانش مقدماتی از داده ها می رسیم که این دانش مقدماتی در مراحل بعدی به تشریح و تفسیر نتایج مدل سازی کمک شایانی خواهد نمود. جدول شماره ۲ برخی از پارامترهای آماری متغیرهای تحقیق را نشان می دهد.

جدول شماره ۲. برخی از پارامترهای آماری تحقیق

تعداد		SEX		AGE		CHOLOSTROL		LDL		HDL		TG		FBS	
		زن	مرد	۴۰-۳۰	۹۲-۴۱	نرمال	غیرنرمال	نرمال	غیرنرمال	نرمال	غیرنرمال	نرمال	غیرنرمال	نرمال	غیرنرمال
کل بیماران	۵۷۰۶	۳۸۱۴	۱۸۹۲	۱۹۵۹	۳۷۴۷	۳۳۴۲	۲۳۶۴	۳۳۴۴	۲۳۶۲	۳۳۳۳	۲۳۷۳	۴۳۷۶	۱۳۳۰	۵۰۷۵	۶۳۱
برحسب درصد	۱۰۰	۸۴/۶۶	۱۶/۳۳	۳۴/۳۴	۶۶/۶۵	۵۶/۵۸	۴۴/۴۱	۶۰/۵۸	۴۰/۴۱	۴۱/۵۸	۵۹/۴۱	۶۹/۷۶	۳۱/۲۳	۹۴/۸۸	۰۶/۱۱
دیابتی	۶۳۱	۴۶۱	۱۷۰	۱۲۴	۵۰۷	۳۲۸	۳۰۳	۳۸۰	۲۵۱	۳۱۷	۳۱۴	۳۷۹	۲۵۲	۰	۶۳۱
برحسب درصد	۱۰۰	۰۵/۷۳	۹۵/۲۶	۶۵/۱۹	۳۵/۵۰	۹۹/۵۱	۰۱/۴۸	۲۳/۶۰	۷۷/۳۹	۲۴/۵۰	۷۶/۴۹	۰۷/۶۰	۹۳/۳۹	۰	۱۰۰

از ۵۷۰۶ بیمار ۶۳۱ بیمار یعنی ۰۶/۱۱ درصد بیماران مبتلا به دیابت بودند.

از ۶۳۱ بیمار دیابتی ۴۶۱ نفر زن و ۱۷۰ نفر مرد بودند یعنی درصد ابتلاء دیابت در زنان بیشتر از مردان است که یکی از دلایل آن دیابت دوران بارداری زنان است.

از ۶۳۱ بیمار دیابتی ۶۲۳ نفر یعنی ۷۳/۹۸ درصد یکی از فاکتورهای چربی غیر نرمال بودند و یا سن بالای ۴۰ سال بودند که به احتمال زیاد دیابت نوع ۲ داشته اند این نتیجه خود گویای صحت روش تحقیق پیش رو می باشد.

از ۴۶۱ نفر زن بیمار دیابتی ۴۵۸ نفر یعنی ۳۴/۹۹ درصد دارای یکی از فاکتورهای چربی غیر نرمال و یا سن بالای ۴۰ سال بودند.

از ۶۳۱ نفر بیمار دیابتی تنها ۸ نفر یعنی ۲۶/۱ درصد آن ها دارای فاکتورهای چربی نرمال و یا سن زیر ۴۰ سال بودند که به احتمال زیاد دیابت نوع ۱ داشته اند.

از ۶۳۱ نفر بیمار دیابتی ۵۷۶ نفر یعنی ۲۸/۹۱ درصد آن ها دارای سن بالای ۴۰ سال بودند که این نتیجه تاثیر سن روی دیابت را مشخص می کند.

از کل بیماران یعنی ۵۷۰۶ نفر بیمار ۱۳۸۴ نفر یعنی ۲۵/۲۴ درصد به عبارتی تقریباً یک چهارم بیماران سالم بودند یعنی دارای انواع چربی خون نرمال و میزان قندخون ناشتا نرمال بودند.

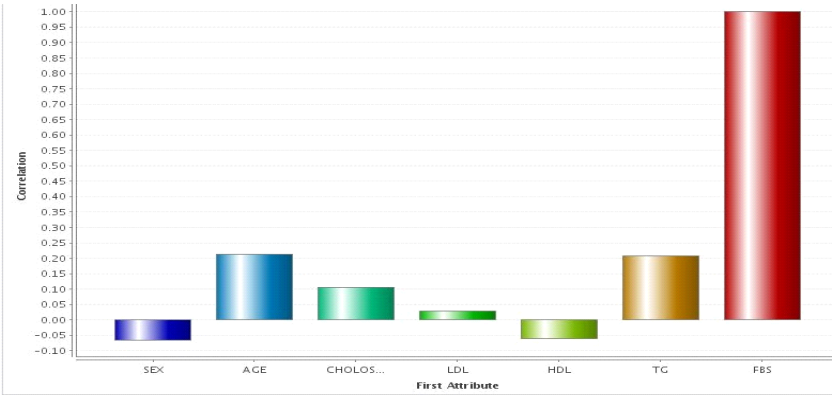
تحلیل های آماری بالا تاثیر انواع چربی و سن و جنسیت روی میزان قندخون و دیابت را نشان می دهد که با توجه به آن ها می توان یک مدل برای دسته بندی و پیش بینی بیماران به دیابتی و غیر دیابتی ارائه داد.

با استفاده از نرم افزار رپید ماینر می توانیم ماتریس و نمودار همبستگی متغیرهای تحقیق را به دست آوریم. ضریب همبستگی ابزاری آماری برای تعیین نوع و درجه رابطه یک متغیر کمی با متغیر کمی دیگر است. ضریب همبستگی شدت رابطه و هم چنین نوع رابطه(مستقیم یا معکوس) بین دو متغیر را نشان می‌دهد

که در بازه(۱،۱-) است. مقدار منفی آن وجود رابطه خطی معکوس، مقدار مثبت آن وجود رابطه خطی مستقیم و مقدار صفر آن عدم وجود رابطه بین دو متغیر را نشان می دهد که نتایج آن در جدول شماره ۳ و نمودار شماره ۱ نشان داده شده است(۱۰).

جدول شماره ۳. ماتریس ضریب همبستگی بین متغیرها

Attributes	SEX	AGE	CHOLOST...	LDL	HDL	TG	FBS
SEX	1	-0.011	-0.126	-0.120	-0.173	0.039	-0.066
AGE	-0.011	1	0.185	0.138	0.060	0.108	0.213
CHOLOSTR	-0.126	0.185	1	0.895	0.224	0.359	0.106
LDL	-0.120	0.138	0.895	1	0.152	-0.033	0.028
HDL	-0.173	0.060	0.224	0.152	1	-0.300	-0.060
TG	0.039	0.108	0.359	-0.033	-0.300	1	0.207
FBS	-0.066	0.213	0.106	0.028	-0.060	0.207	1



نمودار شماره ۱. ضریب همبستگی سایر متغیرها با میزان قندخون ناشتا

مدل سازی و ارزیابی: از کل داده های مورد استفاده در این پژوهش ۹۰ درصد آن ها را برای آموزش مدل ها و ۱۰ درصد بقیه را برای ارزیابی مدل های به دست آمده استفاده کردیم. جهت ارزیابی مدل از ماتریس اغتشاش استفاده نمودیم که یک معیار رایج برای ارزیابی مدل های داده کاوی می باشد(۱۰). عناصر این ماتریس به صورت زیر است:

ماتریس اغتشاش		رکوردهای پیش بینی شده	
		دسته-	دسته+
رکوردهای واقعی	دسته-	TN منفی درست	FP مثبت اشتباه
	دسته+	FN منفی اشتباه	TP مثبت درست

TN: تعداد افرادی که واقعاً دیابت نداشته اند و مدل هم به درستی آن ها را غیردیابتی معرفی کرده است.

TP: تعداد افرادی که واقعاً دیابت داشته اند و مدل هم به درستی آن ها را دیابتی معرفی کرده است.

FP: تعداد افرادی که واقعاً دیابت نداشته اند و مدل به اشتباه آن ها را دیابتی معرفی کرده است.

FN: تعداد افرادی که واقعاً دیابت داشته اند و مدل به اشتباه آن ها را غیر دیابتی معرفی کرده است.

در حقیقت TP و TN پیش بینی های درست مدل و FP و FN پیش بینی های اشتباه مدل را نشان می دهند. دقت مدل از فرمول زیر حساب می شود:

در مورد مدل به دست آمده با روش درخت تصمیم عناصر ماتریس اغتشاش به صورت زیر است:

FN= 7 ،FP= 50 ،TP= 6 ،TN= 508

با استفاده از فرمول بالا، دقت مدل برابر است با:

کارایی مدل آزمایش وارزیابی به منظور پیش بینی مبتلا شدن افراد به دیابت با داده های ورودی به صورت گسسته با استفاده از روش های مختلف مدل سازی پیاده سازی شد که بهترین نتایج دقت مدل ها در جدول شماره ۴ به دست آمده است.

جدول شماره ٤. بهترین نتایج مدل سازی با تکنیک های مختلف دسته بندی

نام تکنیک	نام بهترین عملگر با داده به شکل گسسته در هر تکنیک	دقت مدل ارزیابی پیش بینی دیابت(درصد)
درخت تصمیم	Decision Tree	۹۰/۰۲
قوانین استنتاج	Tree to Rules	۸۹/۰۸
شبکه عصبی مصنوعی	Perceptron	۸۷/۰۳
مبتنی بر نظریه بیز	Naive Bayes	۸۷/۱۲
بردار پشتیبان	Fast Large Margin	۸۸/۷۷

البته برای رسیدن به بالاترین دقت مدل مربوط به هر تکنیک لازم بود تا تکنیک گسسته سازی روی داده ها صورت گیرد. هم چنین با تمام عملگرهای مربوط به هر تکنیک و تغییر پارامترهای هر عملگر به صورت جداگانه مدل سازی در رپید ماینر صورت گیرد.

بحث و نتیجه گیری

روش های داده کاوی در سال های اخیر در حوزه پزشکی و مراقبت های بهداشتی در زمینه تشخیص و پیشگیری بیماری و انتخاب روش درمان و پیش بینی مرگ و میر و پیش بینی هزینه های درمانی به طور گسترده ای مورد استفاده قرار گرفته است.

در این تحقیق با استفاده از تکنیک های مختلف دسته بندی در نرم افزار داده کاوی رپیدماینر مدل هایی به منظور تشخیص انواع دیابت و پیش بینی و دسته بندی بیماران به دیابتی و غیردیابتی ساخته شده است. بهترین دقت مدل ارزیابی مربوط به درخت تصمیم C4.5 معادل ۰۲/۹۰ درصد می باشد.

یکی از مواردی که می تواند به این تحقیق و تحقیقات آتی در زمینه کار روی داده های آزمایشگاهی کمک کند ایجاد یک پرونده الکترونیکی از وضعیت و سوابق پزشکی هر بیمار و وضعیت فیزیکی بیمار از جمله فشارخون و وزن و قد و میزان دور شکم و توده چربی در آزمایشگاه ها می باشد. با در اختیار داشتن این اطلاعات می توان به دقت مدل بالاتری به منظور پیش بینی میزان قندخون و تشخیص دیابت دست یافت.

این مدل می تواند در آزمایشگاه ها برای افرادی که آزمایش چربی خون انجام داده اند ولی آزمایش قندخون انجام نداده اند به منظور پیش بینی میزان قندخون و تشخیص دیابت مورد استفاده قرار گیرد.

References

1.Nazarzadeh M, Bidel Z, Sanjari Moghaddam A. Meta analysis of diabetes mellitus and risk of hip fractures small study effect. Osteoporos Int2015;26:123-9.

2.Janahmadi Z, Nekooeian AA, Mozafari M. Hydroalcoholic extract of Allium eriophyllum leaves attenuates cardiac impairment in rats with simultaneous type 2 diabetes and renal hypertension. Res Pharm Sci 2015;10:125-33.

3. Elsappagh S, Elmogy M, Riad AM. A fuzzy ontology oriented case based reasoning framework for semantic diabetes diagnosis. Artif Intell Med2015;14:92-5.

4. Baronepel O, Heymann AD, Friedman N, Kaplan G. Development of an unsupportive social interaction scale for patients with diabetes. Patient Prefer Adherence2015;23;9:1033-41.

5. Chaoton S, Chienhsin Y, Kuanghung H, Wenko C. Data mining for the diagnosis of type II diabetes from three dimensional body surface anthropometrical scanning data. Comput Matemat Appl 2006; 51:1075-92.

6. Santi WP. A new smooth support vector machine and its applications in diabetes disease diagnosis. J Comput Sci2009; 5:1003-8.

7. Barakat NB, Bradley AP, Barakat MNH. Intelligible support vector machines for diagnosis of diabetes mellitus. IEEE Trans Info Technol biomed 2010; 14:56-66.

8. Worachartcheewan A, Shoombuatong W, Pidetcha P, Nopnithipat W, Prachayasittikul V, Nantasenamat C. Predicting metabolic syndrome using the random forest method.- Scientific World J 2015; 2015:581501.

9.Esteghaamati A. Comprehensive guide diagnosis and treatment diabetes. Spe Soc Diabetes America 2004;5:16-30.

10. Hische M, Larhlimi A, Schwarz F, Fischerrosinsky A, Bobbert T, Assmann A, et al. A distinct metabolic signature predicts development of fasting plasma glucose. J Clin Bioinfo2012;2:2-3.

11. Meng XH, Huang YX, Rao DP, Zhang Q, Liu Q .Comparison of three data mining models for predicting diabetes or prediabetes by risk factors. Kaohsiung J Med Sci 2013;29:93-9.

12. Karaolis MA, Moutiris JA, Hadjipanayi D, Pattichis CS. Assessment of the risk factors of coronary heart events based on data mining with decision trees. IEEE Trans Inf Technol Biomed2010;14:559-66.

13. Toussi M, Lamy JB, Le Toumelin P, Venot A. Using data mining techniques to explore physicians therapeutic decisions when clinical guidelines do not provide recommendations methods and example for type 2 diabetes. BMC Med Info Dec Mak2009;10;9:28.

14. Asadollahi K, Delpisheh A, Asadollahi P, Abangah G. Hyperglycaemia and its related risk factors in Ilam province west of Iran a population based study. J Diabetes Metab Disord2015;14:81.

Data Mining Techniques to Diagnose Diabetes Using Blood Lipids

Rafeh R[*]*, Arbabi M*

(Received: January 3, 2015 Accepted: April 12, 2015)

Abstract

Introduction: Nowadays, diabetic disease is one of the most common, dangerous and costly diseases in the world spreading rapidly. Data mining techniques can be used for early diagnosis of this disease which results in preventing a lot of problems for patients including heart diseases, vision problems and kidney disorders.

Matherials & methods: In this research, the Rapid Miner software has been used as a modeling tool to classify each patient as either diabetic or non-diabetic. The data set of this research has been collected from the database of one lab in Nahavand which includes the information of 5706 patients in a five years period from 2009 to 2013. The data set includes such information about patients as: age, gender, the level of lipid in the blood and the amount of fasting blood sugar.

Findings: After modeling with different classification techniques, the best accuracy achieved from the decision tree c4.5 which was 90.02%.

Discussion & Conclusion: For early diagnosis of diabetes in many countries around the world many techniques have been proposed using a variety of methods and variables. In the current research, using the relationship between blood lipids and fasting blood sugar, a method based on data mining techniques for diagnosing diabetes has been proposed.

Keywords: Data Mining, Diabetes, Classification techniques, Decision tree C4.5.

1.Dept of Computer Engineering, Arak University, Arak, Iran
2. Dept of Computer Engineering, Islamic Azad University, Malayer Branch, Malayer, Iran
**Corresponding author Email: r-rafeh@araku.ac.ir*

Scientific Journal of Ilam University of Medical Sciences