مجله دیابت و متابولیسم ایران، خرداد و تیر ۱۴۰۰؛ دوره ۲۱ (شماره ۲)۱۰۰-۹۲

مقاله پژوهشی

تشخیص بیماری دیابت با استفاده از الگوریتم جنگل تصادفی

صادق مشرف زاده[1*]،بهمن روایی[1]،احسان اله کوزه گر[2]

چکیده

مقدمه: دیابت چهارمین عامل مرگ و میر در دنیا است. و از آنجایی که بسیاری از مردم جهان به این بیماری مبتلا و یا در معرض خطر آن هستند، می‌توان دیابت را بیماری قرن نامید. دیابت تأثیرات مخربی بر سلامتی افراد جامعه دارد و در صورت تشخیص دیر هنگام، می‌تواند صدمات جبران‌ناپذیری به بینایی، کلیه‌ها، قلب، شریان‌ها و غیره وارد کند. بنابراین لازم است که روش‌هایی برای تشخیص این بیماری در مراحل اولیه وجود داشته باشد. در این مقاله، از داده‌کاوی در تشخیص دیابت استفاده شده است.

روش‌ها: الگوریتم اصلی مورد استفاده در این مقاله، الگوریتم جنگل تصادفی است. برای بررسی کارایی الگوریتم پیشنهادی در تشخیص بیماری دیابت از مجموعه داده‌هایی استفاده شده است که شامل ۷٦۸ نمونه (بیمار) و دارای ۸ ویژگی بوده است. از آنجایی که الگوریتم جنگل تصادفی یک الگوریتم ترکیبی است و از چندین درخت تصمیم ایجاد شده است، باعث رسیدن به دقت بالایی در تشخیص بیماری دیابت می‌گردد.

یافته‌ها: با استفاده از این الگوریتم توانستیم میزان دقت تشخیص بیماری دیابت را به ۹۹/۸٦ ٪ افزایش دهیم.

نتیجه‌گیری: برای تشخیص دیابت از الگوریتم‌های مختلفی استفاده شده است ما سعی کردیم از الگوریتمی استفاده کنیم که نسبت به بقیه الگوریتم‌ها برای تشخیص این بیماری از میزان دقت بسیار بالایی برخوردار باشد.

واژگان کلیدی: بیماری، داده‌کاوی، دیابت، الگوریتم جنگل تصادفی

۱- گروه کامپیوتر، دانشکده‌ی فنی و مهندسی، دانشگاه یاسوج، ایران

۲- گروه کامپیوتر، دانشکده‌ی فنی و مهندسی شرق گیلان، دانشگاه گیلان، ایران

* ***نشانی:*** یاسوج، خیابان دانشجو، دانشگاه یاسوج، کد پستی: ۷٤۹۳٤-۷۵۹۱۸، تلفن: ، ۰۷٤۳۳۳۱۳۲۱۲، پست الکترونیک: **info@yu.ac.ir**

تاریخ دریافت: ۱۴۰۰/۰۲/۱۶ **تاریخ درخواست اصلاح: ۱۴۰۰/۰۳/۱۲** **تاریخ پذیرش: ۱۴۰۰/۰۴/۲۷**

مقدمه

بیماری دیابت یکی از شایع‌ترین بیماری‌های شناخته شده در دنیا است [۱]. دیابت عمدتاً از طریق آزمایش‌های قندخون تشخیص داده می‌شود [۱]. همچنین این بیماری چهارمین عامل مرگ و میر در دنیا است [۲]. از طرف دیگر این بیماری یک اختلال در سوخت‌ساز (متابولیسم) بدن است [۳]. در حالت طبیعی، غذا در معده تبدیل به گلوکز یا قندخون می‌شود. قند از معده وارد جریان خون شده، سپس لوزالمعده (پانکراس) هورمون انسولین را ترشح می‌کند و این هورمون باعث می‌شود قند از جریان خون وارد سلول‌های بدن گردد، در نتیجه قندخون در حد نرمال و متعادل باقی می‌ماند [۳]. ولی در بیماری دیابت، انسولین به میزان کافی در بدن وجود ندارد و یا انسولین موجود قادر نیست تا وظایف خود را به‌درستی انجام دهد، در نتیجه به‌علت وجود مقاومت در برابر آن قندخون نمی‌تواند به‌طور مؤثری وارد سلول‌های بدن شود و باعث بالا رفتن سطح خون می شود[۳].. دیابت تأثیرات مخربی بر سلامتی افراد دارد و در صورت تشخیص دیر هنگام، می‌تواند صدمات جبران‌ناپذیری به بینایی، کلیه‌ها، قلب، شریان‌ها و غیره وارد کند [٤]. در مطالعه‌ای که در ایران انجام شده است، گزارش شده که ۷/۷ درصد بالغین ۲۵ تا ٦٤ ساله که حدود دو میلیون نفر هستند، مبتلا به دیابت بوده و ١٦/٨ 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 [۵].براساس آمار فدراسیون جهانی دیابت در سال ۲۰۱۲ بیش از ۳۷۱ میلیون نفر از مردم جهان مبتلا به دیابت بوده‌اند که هر سال نیز به آن افزوده می‌شود به گونه‌ای که بیش از نیمی از مبتلایان به دیابت از بیماری خود بی‌خبر هستند. برای بیماران دیابتی بیش از ٤۷۰ میلیارد دلار هزینه می‌شود. هم‌چنین براساس سرشماری آماری که در سال ۲۰۱۳ صورت گرفته بیش از ٦ میلیون نفر ایرانی مبتلا به دیابت هستند [٦]. در مطالعه‌ای که سال ۱۳۸۹ منتشر شد نشان داد که شیوع دیابت در منطقه‌ی خاورمیانه به‌طور قابل توجهی تا سال ١٤٩٩ افزایش خواهد یافت و برآورد می‌شود نرخ رشد سالیانه دیابت تا سال ١٤٩٩ در ایران بعد از پاکستان به رتبه‌ی دوم منطقه برسد [۲]. در این مقاله، از تکنیک‌های داده‌کاوی برای طبقه‌بندی تشخیص بیماران دیابتی استفاده شده است. داده‌کاوی، فرآیندی است خودکار برای استخراج الگوهایی که دانش را بازنمایی می‌کنند [۷]. داده‌کاوی قادر به کشف الگوهایی هست که در حالت عادی ممکن است توسط هوش انسانی قابل کشف نباشد [۸]. استفاده از روش‌های یادگیری ماشین و داده‌کاوی در تحقیقات دیابت ضروری است تا بتوان حجم زیاد داده‌های مرتبط با دیابت را بررسی و تحلیل نمود. تأثیر شدید اجتماعی این بیماری، یکی از اولویت‌های اصلی در تحقیقات علوم پزشکی است که به ناچار مقادیر عظیمی از داده‌ها را تولید می‌کند. بدون شک، یادگیری ماشین و روش‌های داده‌کاوی در دیابت مربوط به تشخیص، مدیریت و دیگر جنبه‌های مرتبط با مدیریت بالینی هستند[۹]. برای تشخیص بیماری دیابت در مقالات قبل از الگوریتم‌های مختلف همانند شبکه‌ی عصبی مصنوعی چندلایه پرسپترون [۲]، الگوریتم درخت تصمیم C5.0 [۵]، الگوریتم بهینه‌سازی موج آب [٤]. استفاده شده است، که در بخش بعدی به صورت مفصل شرح داده شده‌اند. الگوریتم جنگل تصادفی یک الگوریتم ترکیبی است که از چندین درخت تصمیم ایجاد شده و به‌صورت جنگلی از درخت‌های تصمیم که دارای میزان دقت بالاتری به نسبت دیگر روش‌ها است. با توجه به عدم دقیق بودن روش‌های قبل، نیاز به ارائه روش جدید کاملاً احساس می‌شود. بنابراین در این مقاله، یک روش پیشنهادی براساس الگوریتم جنگل تصادفی ارائه شده است. برای بررسی کارایی روش پیشنهادی در تشخیص دیابت اقدام به پیاده‌سازی آن در ابزار $WEKA$[۲] کرده‌ایم.

روش ها

روش‌های داده‌کاوی مختلفی برای تشخیص بیماری دیابت استفاده شده‌اند که در ادامه بعضی از مهم‌ترین روش‌های موجود تشریح شده اند.

ارائه‌ی یک سیستم ترکیبی هوشمند به‌منظور تشخیص بیماری دیابت: در این تحقیق، داده‌های ۷٦۸ بیمار با ۸ ویژگی مورد ارزیابی قرار گرفت. با استفاده از فرآیندهای پردازش و هوش مصنوعی شامل ماشین بردار پشتیبان[۳] به‌منظور کاهش و مدیریت داده، الگوریتم‌های

[2] - Waikato Environment for Knowledge Analysis
[3] Support vector machines

تکاملی BPSO[1] برای انتخاب بهترین جواب، سیستم‌های دقیق فازی برای افزایش دقت، سرعت، صحت الگوریتم تکاملی، و شبکه‌های عصبی مصنوعی[2] به‌منظور برآورد، آموزش، انطباق پذیری، یادگیری ماشینی، در جهت شناسایی و تشخیص این بیماری استفاده شده است. با استفاده از ترکیب روش‌های مذکور میزان دقت تشخیص بیماری دیابت ۹۵/۸۱٪ است [۱]. معایب روش، پس از مقداردهی اولیه به ذرات، ذرات در جمعیت معمولاً در تکرار پی‌درپی فرآیند جستجو به بهینه محلی یا سراسری در فضای جستجو همگرا می‌شوند و ذرات جمعیت تحت تأثیرپذیری و تبعیت از بهترین ذره کشف شده در کل جمعیت اطراف آن ذره جمع می‌شوند در نتیجه ذرات همگرا شده از توانایی جستجوی سراسری مناسبی جهت دنبال کردن بهینه برخوردار نخواهند بود.

تشخیص بیماری دیابت با استفاده از شبکه‌ی عصبی مصنوعی و عصبی- فازی: این تحقیق، از نوع تحلیلی بوده و پایگاه داده آن مشتمل بر ۷۶۸ بیمار با ۸ ویژگی است. میزان دقت تشخیص بیماری دیابت با استفاده از روش شبکه‌ی عصبی مصنوعی چندلایه پرسپترون[3] و شبکه‌ی عصبی بردار یادگیر کوانتیزه [4] و شبکه‌های عصبی- فازی[5] به‌ترتیب ۹۸/۶٪ و ۹۸/۲٪ و ۹۹/۶٪ است [۲]. معایب روش، قواعد یا دستورات مشخصی برای طراحی شبکه جهت یک کاربرد اختیاری وجود ندارد و دقت نتایج بستگی زیادی به اندازه‌ی مجموعه آموزش دارد.

پیش‌بینی و تشخیص دیابت با استفاده از الگوریتم بهینه‌سازی موج آب: در این تحقیق، به‌منظور کشف الگوهای پنهان‌شده و تشخیص دیابت، الگوریتم بهینه‌سازی موج آبی (WWO)، به‌عنوان یک الگوریتم دگرگونی دقیق، همراه با یک شبکه‌ی عصبی برای افزایش دقت پیش‌بینی دیابت استفاده شده است. نتایج حاصل از اجرا در محیط برنامه‌نویسی متلب با استفاده از مجموعه داده مربوط به دیابت، نشان می‌دهد که میزان دقت تشخیص بیماری دیابت با استفاده از این الگوریتم ۹۵,۴۶ ٪ است [۴]. معایب روش دقت نتایج بستگی زیادی به اندازه مجموعه آموزش دارد

استخراج دانش از داده‌های بیماران دیابتی با استفاده از روش درخت تصمیم C5.0: در این تحقیق، برای اولین‌بار احتمال بروز عوارض میکرو‌واسکولار، ماکرو‌واسکولار و یا هر دو نوع عارضه در بیماران دیابتی و ویژگی‌های تاثیرگذار بر آنها مورد بررسی قرار گرفت. متغیرهای فشارخون بالا، سن و سابقه‌ی خانوادگی در عوارض مشاهده شده بیشترین تأثیر را داشته‌اند. به کمک درخت تصمیم ایجاد شده، قوانینی استخراج شده‌اند که می‌تواند به‌عنوان الگویی برای پیش‌بینی وضعیت بیماران و احتمال بروز عوارض در آنها استفاده شود. میزان دقت تشخیص بیماران دیابتی با استفاده از درخت تصمیم C5.0، ۸۹,۷۴ درصد و در شبکه‌ی عصبی مصنوعی ۵۱/۲۸ درصد است [۵]. معایب روش، هرس کردن درخت هزینه‌ی بالایی دارد و در مواردی با تعداد دسته‌های زیاد و نمونه‌های آموزشی کم احتمال خطا بالاست.

مشکلات روش‌های موجود

همان‌گونه که ذکر گردید، الگوریتم‌های مختلفی برای تشخیص بیماری دیابت وجود دارند، امّا میزان دقت به‌دست آمده با استفاده از الگوریتم‌های موجود مطلوب نیست. بنابراین در این مقاله، سعی بر این بوده که از الگوریتمی استفاده گردد که دقت تشخیص دیابت را نسبت به الگوریتم‌های موجود بهبود بخشد.

روش پیشنهادی

روش پیشنهادی تشخیص بیماری دیابت با استفاده از الگوریتم جنگل تصادفی است که در این بخش به‌طور مختصر توضیح داده خواهد شد.

الگوریتم جنگل تصادفی[6]

الگوریتم جنگل تصادفی یک الگوریتم گروهی با مجموعه‌ای از درختان تصمیم است. دقت طبقه‌بندی روش جنگل تصادفی با ساخت مجموعه‌ای از درختان و رأی‌گیری بین آنها برای به دست‌آوردن رده‌های با بیشترین تعداد رأی، پیشرفت‌های قابل توجهی داشته است.

[1] Best Practice Spotlight Organization
[2] Neural Network
[3] Multilayer perceptron
[4] Learning Vector Quantization
[5] Nero fuzzy
[6] Description of random forest algorithm

دو ویژگی مهم در ساخت جنگل‌های تصادفی، روش بگینگ و انتخاب تصادفی در هر گره است. در ادامه، ویژگی‌ها و خصوصیات جنگل تصادفی توضیح داده شده‌اند [۸].

روش بگینگ[1]

روش بگینگ توسط Leo Breiman در سال ۱۹۹۶ مطرح شد. این روش یک فرا الگوریتم بر مبنای مفاهیم خود راه انداز و ترکیب، برای بهبود یادگیری ماشین است. الگوریتم‌های گروهی در یادگیری ماشین، چند یادگیرنده ضعیف را ترکیب می‌کنند تا به یک یادگیرنده قوی دست یابند. این روش از بیش‌برازش داده‌ها جلوگیری می‌کند. در بگینگ، نتایج خوب زمانی تولید می‌شود که طبقه‌بندهای پایه جزء الگوریتم‌های یادگیری ناپایدار باشند (مانند درخت تصمیم‌گیری یا شبکه‌ی عصبی)، به‌طوری‌که تغییرات کوچک در داده‌های آموزشی منجر به تغییرات عمده‌ای در مدل ساخته شده توسط آن الگوریتم شود. فرآیند الگوریتم بگینگ بدین شرح است: یک مجموعه آموزش D به اندازه m را در نظر بگیرید. بگینگ با نمونه‌گیری یکنواخت و با جایگزینی نمونه‌ها از D، n مجموعه آموزشی جدید D_i با اندازه اولیه m تولید می‌کند. نمونه‌گیری با جایگزینی این امکان را می‌دهد که در هر D_i بعضی از نمونه‌ها امکان تکرار داشته باشند. این نوع نمونه‌گیری به‌عنوان نمونه‌گیری خود راه‌انداز شناخته می‌شود. خروجی ترکیب n مدل با میانگین‌گیری برای رگرسیون و رأی‌گیری برای طبقه‌بندی به‌دست می‌آید. بنابراین، با استفاده از نمونه‌گیری دوباره و تولید مجموعه داده‌های مختلف، تنوع مورد نیاز حاصل خواهد شد [۸].

ویژگی‌های الگوریتم جنگل تصادفی

الگوریتم جنگل تصادفی در میان الگوریتم‌های موجود از دقت بی‌نظیری برخوردار است. این الگوریتم می‌تواند به‌طور مؤثر بر روی پایگاه داده‌های بزرگ اجرا شود و می‌تواند با هزاران متغیر ورودی بدون حذف متغیر سروکار داشته باشد [۱۰].

تعریف جنگل تصادفی

جنگل تصادفی یک طبقه‌بند مجموعه‌های متشکل از طبقه‌بندهای درخت تصمیم است. هر طبقه‌بند برای هر نمونه ورودی به صورت $h(x, \theta_k)$ است، که x یک نمونه ورودی و θ_k مجموعه آموزش برای درخت kام است. θها مستقل از یکدیگر ولی با توزیع یکسان هستند. برای هر نمونه x، هر درخت یک پیش‌بینی را برای رده‌ی نمونه‌ی x ارائه می‌دهد و در نهایت رده‌های با بیشترین تعداد رأی درختان روی ورودی x به عنوان رده‌ی نمونه انتخاب می‌شود، این فرآیند را جنگل تصادفی می‌نامند [۸،۱۱]. الگوریتم جنگل تصادفی می‌تواند دقت پیش‌بینی را نسبت به درخت طبقه‌بندی فردی افزایش دهد. در درخت فردی با تغییرات کوچک در مجموعه آموزش بی‌ثباتی به‌وجود می‌آید که باعث اختلال در دقت پیش‌بینی در نمونه آزمایشی می‌شود، اما گروهی بودن الگوریتم جنگل تصادفی باعث سازگاری با تغییرات می‌شود و بی‌ثباتی را از بین می‌برد [۸].

تخمین خارج از کیسه[2]

فرض کنید هر طبقه‌بند با مجموعه آموزشی جدید با روش درخت تصمیم ساخته می‌شود. با توجه به مجموعه آموزش θ و با روش خود راه انداز مجموعه‌های آموزشی θ_k تشکیل می‌شوند. سپس طبقه‌بندهای درخت $h(x, \theta_k)$ ساخته می‌شود و از هر درخت برای پیش‌بینی رده رأی‌گیری می‌شود. نمونه‌های آموزش در مجموعه داده آموزش اصلی که در مجموعه آموزش طبقه‌بند k نیست، نمونه‌های خارج از کیسه‌ی طبقه‌بند kام نامیده می‌شود. در هر مجموعه آموزش به‌دست آمده از روش خود راه‌انداز، نمونه‌های خارج از کیسه‌ی در حدود یک سوم از نمونه‌های مجموعه آموزش اصلی است که در مجموعه آموزش قرار نمی‌گیرد. رابطه‌ی (۱) شیوه‌ی تخمین رده‌ی نمونه خارج از کیسه‌ی را روی جنگل نشان می‌دهد. برای به‌دست آوردن رده‌ی نمونه باید ابتدا پیش‌بینی درختانی که مجموعه آموزش آنها حاوی نمونه نیست، جمع‌آوری شود و سپس رده‌ای با بیشترین میانگین رأی روی پیش‌بینی‌های درختان جنگل به‌عنوان رده‌ی نمونه در نظر گرفته می‌شود [۸].

$$y(x) = \arg max_c\left(\frac{1}{k}\sum_{k=1}^{k} I\left(h_k(x) = c, x \epsilon\, OOB_K\right)\right) \quad (۱)$$

که K تعداد درختان، c نشان دهنده‌ی رده‌ی $h_k(x)$ پیش‌بینی درخت kام روی نمونه‌ی x را نشان می‌دهند و OOB_K مجموعه نمونه‌های OOB درخت kام هستند. رابطه‌ی (۲) نشان می‌دهد که مقدار تابع

[1] bagging

[2] Out of Bag

شاخص یک خواهد بود اگر x در مجموعه نمونه‌های درخت kام قرار دارد (عضو مجموعه آموزش درخت kام نیست) و همچنین درخت k ام نمونه x را به رده‌ی c طبقه‌بندی کند. در غیر اینصورت، مقدار تابع شاخص صفر می‌شود[۸].

$$I(h_k(x) = c, x\epsilon\ OOB_K) = \begin{cases} 1\ h_k(x) = c, x \in OOB_K \\ 0\ otherwise \end{cases} \quad (۲$$

برای به‌دست آوردن تخمین نمونه‌های OOB روی جنگل $error_k(OOB)$ در رابطه‌ی (۳) استفاده می‌کنیم که خطای طبقه بندی جنگل روی نمونه‌های OOB درخت kام است.

$$I(y_i,\ x_i)\epsilon\ OOB_K = \begin{cases} 1(x_i,\ y_i)\epsilon\ OOB_K \\ 0(x_i,\ y_i) \notin\ OOB_K \end{cases} \quad (۳$$

N تعداد همه نمونه‌های مجموعه آموزش اصلی، x_i نمونه iام روی مجموعه آموزش اصلی، y_i رده‌ی واقعی $x_i, y(x_i)$ رده‌ی پیش‌بینی شده برای x_i برحسب رابطه ۱ است. در رابطه‌ی (٤) مقدار تابع I یک خواهد بود، اگر نمونه (x_i, y_i) متعلق به مجموعه OOB درخت k باشد و در غیر این صورت، صفر است [۸].

$$error_k(OOB) = \frac{\sum_{i=1}^{N}(y(x_i)=y_i;(x_i,y_i)\epsilon\ OOB_K}{\sum_{i=1}^{N} I(x_i,y_i)\epsilon\ OOB_K} \quad (٤$$

قدرت و وابستگی

در جنگل تصادفی، برای خطا یک کران بالا در نظر گرفته می‌شود. دو پارامتر برای این کران بالا اندازه‌گیری می‌شود. پارامتر اول، قدرت طبقه‌بندهای فردی و پارامتر دوم، وابستگی بین درختان جنگل است [۸].

در رابطه‌ی (۵)،s قدرت طبقه‌بندهای فردی در جنگل و $\bar{\rho}$ وابستگی بین درختان جنگل را نشان می‌دهد. براساس این رابطه برای خطای $\overline{PE}^*$ هرچه مقدار s بیشتر و مقدار $\bar{\rho}$ کمتر باشد، میزان خطا نیز کمتر خواهد بود. نسبت $\frac{\rho}{s^2}$ برای جنگل تصادفی در درک عملکرد آن، یک راهنمای مفید است. این نسبت، تقسیم وابستگی بر مربع قدرت است و هر چه کوچکتر باشد، عملکرد جنگل بهتر و خطا نیز کمتر است [۸].

$$PE^* \leq \bar{\rho}(1 - s^2)/s^2 \quad (۵$$

پیش‌بینی رده‌ی نمونه توسط الگوریتم جنگل تصادفی

برای به‌دست آوردن رده‌ی نمونه‌ی آزمایشی برای جنگل تصادفی از پیش‌بینی تمام درختان جنگل استفاده می‌شود. برای تعیین رده‌ی نمونه ی آزمایشی از رابطه (٦ و٧) استفاده می‌کنیم [۸].

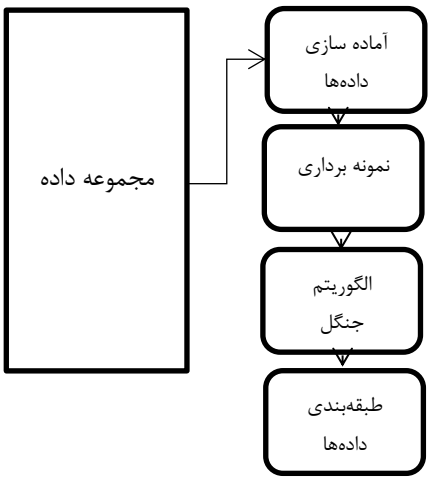
$$y(x) = \arg max_c(\frac{1}{k}\sum_{k=1}^{k} I\ (h_k(x) = c \quad (٦$$

$$(h_k(x) = c) = \begin{cases} 1\ h_k(x) = c \\ 0\ otherwise \end{cases} \quad (٧$$

یافته ها

این بخش که شامل روش پیاده‌سازی و معیارهای ارزیابی است، به‌طور مختصر توضیح داده خواهد شد.

پیاده‌سازی: برای پیاده‌سازی از چندین مرحله استفاده می‌کنیم که در شکل ۱ نشان داده شده‌اند.



شکل ۱- فلوچارت مراحل پیاده‌سازی الگوریتم جنگل تصادفی

مجموعه داده: این مرحله که شامل ۷٦۸ نمونه (بیمار) که ۵۰۰ بیمار غیر دیابتی و ۲٦۸ بیمار دیابتی است را از سایت UCI بارگذاری کرده که هر نمونه ۸ ویژگی (مطابق جدول ۱ نشان داده شده است) در آن بررسی شده است.

آماده‌سازی داده‌ها: در این مرحله مجموعه داده را به ابزار وکا اضافه می‌کنیم.

نمونه برداری داده‌ها: در این مرحله تعداد مجموعه داده‌های آموزشی و تست انتخاب می‌شود که شامل $seed = 1$ و $fold = 10$ است.

الگوریتم جنگل تصادفی: این مرحله که توضیحات آن در بخش قبلی داده شد، برای پیاده‌سازی کنترل، این الگوریتم را به ابزار اضافه می‌کنیم.

طبقه‌بندی داده‌ها: با توجه به درصد نمونه‌های تست و یا مجموعه داده‌ی آزمایشی که طبقه‌بندی می‌شوند، دقت طبقه‌بندی تخمین زده می‌شود. نتایج پیاده‌سازی در جدول ۲ نمایش داده شده است.

جدول ۱- ویژگی‌های مهم تشخیص بیماری دیابت[۱۲]

ویژگی	توضیحات
بارداری	تعداد دفعاتی که فرد باردار بوده است
گلوکز	غلظت گلوکز پلاسما خون (دو ساعت بعد از نوشیدن محلول قند)
فشارخون	فشار خون دیاستولیک در میلی متر جیوه
پوست	ضخامت سه برابر پوست در میلی متر
انسولین	انسولین سرم (دو ساعت بعد از نوشیدن محلول گلوکز)
جرم	شاخص جرم بدن (وزن / قد) * ۲)
نژاد	عملکرد نژاد دیابت" (مربوط به میزان ابتلا به یک فرد از نظر ارثی یا ژنتیکی که دیابت را بالاتر از حد دارد)
سن	وضعیت سن در سال

جدول ۲- نتایج پیاده‌سازی الگوریتم جنگل تصادفی

تعداد	معیار ارزیابی	الگوریتم جنگل تصادفی
۱	میزان دقت(Correctly)	٪ ۹۹,۸٦
	تعداد نمونه های درست طبقه بندی شده	۷٦۷
۲	میزان نادرستی(Incorrectly)	٪ ۰,۱۳۰۲
	تعداد نمونه های نادرست طبقه بندی شده	۱
۳	دسترسی به کیفیت مورد اطمینان در کارایی(Kappa)	۰,۹۹۷۱
٤	میزان میانگین خطای واقعی(Mean absolute error)	۰,۱۱٦۸
۵	میزان میانگین مربع خطا(Root mean squared error)	۰,۱۵۸۹
٦	میزان خطای واقعی نسبی(Relative absolute error)	٪ ۲۵,٦۹۸٤
۷	میزان ریشه‌ی خطای مربع نسبی(Root relative squared error)	٪ ۳۳,۳۲۸٦
۸	تعداد نمونه‌ها(Total Number of Instances)	۷٦۸

معیارهای ارزیابی

با توجه به جدول ۲ تعداد نمونه‌هایی که درست طبقه‌بندی شده اند برابر با ۷٦۷ نمونه است. این بدین معنی است که میزان دقت الگوریتم جنگل تصادفی برابر است با ۹۹/۸٦ درصد. برای محاسبه میزان دقت از رابطه‌ی (۸) استفاده می‌گردد:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (۸)$$

با توجه به جدول ۲ تعداد نمونه‌های که نادرست طبقه‌بندی شده اند نیز برابر با ۱ نمونه است. این معیار بدین معنی است که میزان خطای طبقه‌بندی برابر با ۰/۱۳۰۲ درصد است. میزان خطای براساس رابطه‌ی (۹) محاسبه می‌گردد:

$$Error = 100 - \frac{TP+TN}{TP+TN+FP+FN} \quad (۹)$$

در جدول (٤) جزئیات طبقه‌بندی توسط الگوریتم جنگل تصادفی نشان داده شده است.

جدول ۳- مقایسه روش پیشنهادی با روش‌های قبلی

الگوریتم بهینه‌سازی موج آب	C5.0 درخت تصمیم	شبکه‌های عصبی- فازی	الگوریتم جنگل تصادفی
٪ ۹۵,٤٦	٪۸۹,۷٤	٪۹۹,٦	٪۹۹,۸٦

جدول ۴- جزئیات طبقه‌بندی الگوریتم جنگل تصادفی

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
۰/۹۹۸	۰	۱	۰/۹۹۸	۰/۹۹۹	۱	**Negative**
۱	۰/۰۰۲	۰/۹۹٦	۱	۰/۹۹۹	۱	**Positive**
۰/۹۹۹	۰/۰۰۱	۰/۹۹۹	۰/۹۹۹	۰/۹۹۹	۱	

ستون اول $(TP\ Rate)$: بیانگر میزان درستی طبقه‌بندی توسط الگوریتم جنگل تصادفی به ازای هر نوع از کلاس است.

ستون دوم $(FP\ Rate)$: میزان نادرستی طبقه‌بندی داده‌ها.

ستون سوم $(Precision)$: بیانگر صحت طبقه‌بندی هر کدام از طبقه‌های موجود در مجموعه داده است که براساس رابطه‌ی (۱۰) محاسبه می‌گردد.

$$Precision = \frac{TP}{TP+FP} \quad (۱۰)$$

ستون چهارم $(Recall)$: نسبت مقداری موارد صحیح طبقه‌بندی شده توسط الگوریتم از یک کلاس به تعداد موارد حاضر در کلاس مذکور که براساس رابطه‌ی (۱۱) محاسبه می‌گردد.

$$ReCall = \frac{TP}{TP+FN} \quad (۱۱)$$

ستون پنجم $(F-Measure)$: با توجه به محاسبات انجام گرفته برای معیارهای Precision و Recall، در این مرحله می‌توان مقدار کمیت وزن‌دار F-Measure را محاسبه نمود F-Measure، پارامتر مناسبی برای ارزیابی کیفیت طبقه‌بندی است و همچنین توصیف کننده‌ی میانگین وزن‌دار مابین دو کمیت Precision و Recall است. برای یک الگوریتم طبقه‌بندی کننده در شرایط ایده‌آل، مقدار این کمیت برابر با یک است و در بدترین وضعیت برابر با صفر است. این پارامتر با استفاده از رابطه (۱۲) محاسبه می‌گردد.

$$\mathrm{F-Measure} = 2 * \frac{Precision*ReCall}{Precision+ReCall} \quad (۱۲)$$

ستون ششم (ROC Area): بیانگر میزان درستی و نادرستی طبقه‌بندی مطابق با ROC را بیان می‌کند.

بحث

روش پیشنهادی میزان دقت الگوریتم را تا حد امکان بالا برد و نسبت به روش‌های قبلی مثل الگوریتم درخت تصمیم c4.5 و الگوریتم شبکه عصبی از میزان دقت بسیار بالایی برخوردار است.

نتیجه گیری

دیابت نوعی بیماری است که به‌دلیل افزایش سطح قندخون ایجاد می‌شود. دیابت تأثیرات مخربی بر سلامتی افراد دارد و در صورت

تشخیص دیر هنگام، می‌تواند صدمات جبران‌ناپذیری به حس‌بینایی، کلیه‌ها، قلب، شریان‌ها و غیره وارد کند. برای میزان دقت تشخیص بیماری دیابت از الگوریتم جنگل تصادفی استفاده کردیم. این الگوریتم به‌علت اینکه یک الگوریتم ترکیبی است و از هم پیوستن چندین درخت تصمیم ایجاد شده و به‌صورت جنگل تشکیل شده است، می‌تواند یک الگوریتم بسیار کارآمد در بین الگوریتم‌های دیگر داده‌کاوی باشد. برای پیاده‌سازی، یک مجموعه داده که شامل ۷۶۸ بیمار که هر بیمار ۸ ویژگی در آن بررسی شده است، استفاده کردیم و در آخر از الگوریتم جنگل تصادفی استفاده کردیم و توانستیم میزان دقت تشخیص بیماری دیابت را تا ۹۹/۸۶ درصد بالا ببریم.

پیشنهادات آتی

با ترکیب الگوریتم‌های مختلف داده‌کاوی مثل ترکیب الگوریتم جنگل چرخشی و الگوریتم j48 میزان دقت تشخیص دیابت را افزایش داد.

سپاسگزاری

از تمامی افرادی که در نوشتن این مقاله ما را همراهی نمودند تشکر و قدردانی می‌گردد.

مآخذ

۱. فیوضی، محمد؛ قره‌خانی، اعظم؛ و حدادنیا، جواد. ارائه‌ی یک سیستم ترکیبی هوشمند به‌منظور تشخیص بیماری دیابت. *بیستمین کنفرانس مهندسی برق ایران*، ۱۳۹۱.

۲. ذباح، ایمان؛ اسکندری، اسما؛ سرداری، زهرا؛ نوقندی، ابوالفضل. تشخیص بیماری دیابت با استفاده از شبکه‌ی عصبی مصنوعی و عصبی- فازی. *مجله‌ی دانشگاه علوم پزشکی تربت حیدریه*، ۱۳۹۷؛ ۶(۲): ۲۰-۱۰.

۳. مشکوتی، الهام؛ معینی، علی. تشخیص بیماری دیابت با استفاده از ماشین بردار پشتیبان. *کنفرانس بین‌المللی یافته‌های نوین پژوهشی در مهندسی برق و علوم کامپیوتر*، ۱۳۹۴.

4. Taherian Dehkordi S, Khatibi Bardsiri A, and Zahedi MH. Prediction and Diagnosis of Diabetes Mellitus using a Water Wave Optimization Algorithm. *Journal of AI and Data Mining* 2019; 7(4): 617-630.

۵. عامری، حکیمه؛ علیزاده، سمیه؛ برزگری، اکبر. استخراج دانش از داده‌های بیماران دیابتی با استفاده از روش درخت تصمیم C5.0. *فصلنامه‌ی مدیریت سلامت*، ۱۳۹۲؛ ۱۶(۵۳):۵۸-۷۲

۶. رافع، رضا؛ و اربابی، محمد. استفاده از تکنیک‌های داده‌کاوی جهت تشخیص دیابت با استفاده از چربی خون. *مجله‌ی علمی دانشگاه علوم پزشکی ایلام*، ۱۳۹۴؛ ۲۳(۴):۲۳۹-۲۴۷

۷. ضیاءالدینی، سلیمه؛ ابارقی، مینا. ارائه یک روش جدید برای تخمین مقادیر گمشده در مجموعه داده. *مجله‌ی مدل‌سازی در مهندسی دانشگاه سمنان*، ۱۳۹۷؛ ۱۶(۵۵): ۱۶۲-۱۵۵.

۸. نیک آبادی، محسن شفیعی؛ و عظیمی، سید علی. *پیش‌بینی تقاضا در زنجیره‌ی تأمین با استفاده از الگوریتم‌های یادگیری ماشین*. *مجله‌ی مدل‌سازی در مهندسی دانشگاه سمنان*، ۱۳(۴۱): ۱۳۶-۱۲۷.

9. T Ioannis Kavakiotis, Olga Tsave, Athanasios Salifoglou, Nicos Maglaveras, Ioannis Vlahavas, and Ioanna Chouvardab, Release year 2017, Machine Learning and Data Mining Methods in Diabetes Research, Comput Struct Biotechnol J, pages 104–116.

10. Venkatesan N, Priya G. A Study of Random Forest Algorithm with implemetation using Weka. *International Journal of Innovative Research in Computer Science and Engineering (IJIRCSE)* 2015; 1(6): 156-162.

11. Cutler A, Richard Cutler D, Stevens JR. Random Forests. In: Zhang C., Ma Y. (eds) *Ensemble Machine Learning*. Springer, Boston, MA, 2012; pp 157-175.

12. Diagnosing Diabetes with Weka & Machine Learning, 2018. Available in: https://www.alexstrick.com/blog/diagnosing-diabetes-with-weka-machine-learning decreased abdominal visceral adipose tissue in overweight and obese adults. *The American journal of clinical nutrition* 2012; 95(1):101-8.

Diagnosis of Diabetes Using a Random Forest Algorithm

Sadegh Mosharrafzadeh [*1], Bahman Ravaei[1], Ehsanollah Koozegar[2]

1. Computer Engineering, Faculty of Engineering, Yasouj University, Iran
2. Computer Engineering, East Guilan Faculty of Engineering, University of Guilan, Iran

ABSTRACT

Background: Diabetes is the fourth leading cause of death in the world. And because so many people around the world have the disease, or are at risk for it, diabetes can be called the disease of the century. Diabetes has devastating effects on the health of people in the community and if diagnosed late, it can cause irreparable damage to vision, kidneys, heart, arteries and so on. Therefore, it is necessary to have methods to diagnose this disease in the early stages. In this article, data mining is used to diagnose diabetes.
Methods: The main algorithm used in this paper is the random forest algorithm. To evaluate the efficiency of the proposed algorithm in diagnosing diabetes, a data set was used that included 768 samples (patients) and had 8 characteristics. Because the stochastic forest algorithm is a hybrid algorithm created from several decision trees, it achieves high accuracy in diagnosing diabetes.
Results: Using this algorithm, we were able to increase the accuracy of diabetes diagnosis to 99.86%.
Conclusion: Diabetes is the fourth leading cause of death in the world. Different algorithms have been used to diagnose this disease. We tried to use an algorithm that has a very high degree of accuracy compared to other algorithms for diagnosing this disease.

Keywords: Disease, Data Mining, Diabetes, Random Forest Algorithm

* Yasuj - Daneshjoo St. - Yasuj University, Postal Code: 75918-74934, Phone Number: 07431000, Email: info@yu.ac.ir.