مجله علمی پژوهشی دانشگاه علوم پزشکی شهید صدوقی یزد
دوره سی و دو، شماره هفت، مهر ۱۴۰۳
تاریخ دریافت:۱۴۰۲/۰۷/۱۰
تاریخ پذیرش:۱۴۰۳/۰۴/۱۳

شناسایی عوامل موثر و پیش‌بینی بیماری ایسکمیک قلبی با استفاده از روش‌های یادگیری ماشین و داده های طرح سلامت یزد (YaHS)

جمال زارع‌پور احمدآبادی*[۱]، فاطمه زارع مهرجردی[۲]، مهدیه قنبری[۳]، مسعود میرزایی[۴]

مقاله پژوهشی

مقدمه: بیماری‌های قلبی و عروقی، از شایع‌ترین بیماری‌هایی است که آمار مرگ و میر بالایی را در جهان به خود اختصاص داده است. مقاله حاضر به شناسایی عوامل مختلف موثر بر بیماری های ایسکمیک قلبی و شناسایی افراد مستعد به آن با استفاده از انواع روش‌های یادگیری ماشین پرداخته است.

روش بررسی: پژوهش حاضر بر روی داده‌های طرح سلامت مردم یزد (یاس) انجام شده است. در این مطالعه، داده‌های مربوط به سلامت، بیماری‌ها و عوامل خطر مختلف نزدیک به ۱۰۰۰۰ نفر در قالب یک پرسش‌نامه با ۳۰۰ سوال مختلف جمع‌آوری شده است. در این پژوهش، با استفاده از محاسبه هم‌بستگی متغیرهای مختلف جمع‌آوری شده در طرح یاس، عوامل مهمی که با بیماری‌های ایسکمیک قلبی مرتبط هستند، جستجو شده‌اند. سسپس با استفاده از عوامل و الگوریتم‌های یادگیری ماشین، مشخصات افراد مستعد بیماری قلبی شناسایی شدند.

نتایج: نتایج ارزیابی‌ها نشان می‌دهد عواملی مانند سن، سابقه بیماری قلبی خانوادگی، فشارخون، دیابت، کلسترول خون، استرس، اضطراب، افسردگی، کیفیت زندگی، کیفیت خواب، فعالیت فیزیکی، مصرف دخانیات و تغذیه فرد با بیماری ایسکمیک قلبی ارتباط دارند. هم‌چنین از بین روش‌های یادگیری ماشین مختلف روش نزدیک‌ترین همسایگی با ۵ خوشه، روش شبکه عصبی عمیق و پرسپترون چند لایه با معیار ارزیابی فراخوان به مقادیر ۹۹/۹۴٪، ۹۹/۸۸٪ و ۹۹/۱۱٪ به ترتیب بهترین عملکرد را در شناسایی افراد بیمار داشته‌اند.

نتیجه‌گیری: کنترل عواملی از جمله فشارخون، دیابت، کلسترول، استرس، اضطراب، افسردگی، بهبود دادن عواملی مانند کیفیت زندگی، وضعیت خواب، فعالیت فیزیکی، الگوی تغذیه افراد و ترک مصرف دخانیات در ارتقا سلامت ساکنان یزد موثر است. از سوی دیگر شناسایی افراد مستعد بیماری های ایسکمیک قلبی با استفاده از روش‌های یادگیری ماشین نسبت به روش‌های سنتی غربالگری که با مراجعه به مراکز درمانی و پزشک و انجام آزمایش‌های مختلف صورت می‌گیرد، سریع‌تر و با صرف هزینه کمتر انجام می‌شود.

واژه‌های کلیدی: بیماری ایسکمیک قلبی، یادگیری ماشین، عوامل موثر، عوامل خطر

ارجاع: جمال زارع‌پور احمدآبادی، فاطمه زارع مهرجردی، مهدیه قنبری، مسعود میرزایی. **شناسایی عوامل موثر و پیش‌بینی بیماری ایسکمیک قلبی با استفاده از روش‌های یادگیری ماشین و داده‌های طرح سلامت یزد (YaHS)**. مجله علمی پژوهشی دانشگاه علوم پزشکی شهید صدوقی یزد ۱۴۰۳؛ ۳۲ (۷): ۷۹-۸۰۶۷.

۱-گروه علوم کامپیوتر، دانشگاه یزد، یزد، ایران.
۲- گروه مهندسی کامپیوتر، دانشگاه میبد، میبد، یزد، ایران.
۳- مرکز بازتوانی قلب، مرکز تحقیقات قلب و عروق، پژوهشکده بیماریهای غیرواگیر، دانشگاه علوم پزشکی شهید صدوقی، یزد، ایران.
۴- پژوهشکده بیماریهای غیرواگیر، مرکز تحقیقات قلب و عروق، دانشگاه علوم پزشکی شهید صدوقی، یزد، ایران.
*(نویسنده مسئول)؛ تلفن: ۰۹۱۶۲۵۳۹۵۴۳، پست الکترونیکی: zarepourjamal@yazd.ac.ir، صندوق پستی : ۸۹۱۵۸۱۸۴۱۱

[DOI: 10.18502/ssu.v32i7.16571]

مقدمه

افزایش بار بیماری‌های غیر واگیر (NCDs) یک چالش بهداشت عمومی در سراسر جهان است. هر سال، ۳۸ میلیون نفر به علت این بیماری‌ها جان خود را از دست می‌دهند. این مرگ‌های زودرس به‌طور اصلی در کشورهای در حال توسعه رخ می‌دهد و در سال‌های اخیر افزایش یافته است. بیش از ۸۲٪ از این مرگ‌ها به بیماری‌های غیر واگیر منتسب می‌شود .عوامل خطر اصلی شامل اضافه وزن، فشار خون بالا، فعالیت بدنی کم، اعتیاد به مواد مخدر و دیس‌لیپیدمی هستند (۱). تخمین زده می‌شود که حدود یک سوم ساکنان شهرستان یزد از سندرم متابولیک (MetS) رنج می‌برند؛ ۸۵٪ از جمعیت (۷۴-۲۰ ساله) حداقل یک و ۵۹٪ دارای دو عامل خطر مرتبط برای بیماری مزمن هستند (۲). عواملی مانند کلسترول خون بالا، دیس‌لیپیدمی، فشار خون بالا و سیگار کشیدن گزارش شده است. (۳). مطالعه اخیر در مورد شیوع چاقی و اضافه وزن در بزرگسالان بالای ۳۰ سال در یزد نشان داد که شیوع چاقی و اضافه وزن رو به افزایش است. درصد اضافه وزن در زنان بیشتر از مردان و در مناطق شهری در مقایسه با روستاها بیشتر است (۴). سیگار مهم‌ترین عامل قابل پیشگیری مرگ و میر در سراسر جهان است و عامل خطر شایع در جمعیت یزد است (۵). با افزایش جمعیت و تغییرات ناسالم در سبک زندگی و محیط، انتظار می‌رود که میزان بروز مرگ و میر ناشی از بیماری‌های غیرواگیر از جمله بیماری قلبی در آینده افزایش یابد. آموزش خودمراقبتی شامل سبک زندگی سالم بیماران قلبی، کنترل فاکتورهای خطر و مصرف صحیح داروها می‌تواند از بروز حوادث جدی در این افراد جلوگیری کند. یکی از مدل‌های کم هزینه برای آموزش خودمراقبتی بیماران مزمن، آموزش غیرحضوری، تامین نیازهای مختلف افراد به صورت برخط و مشاوره‌های مجازی این افراد می‌باشد. در مطالعه‌ای نجفقلی‌زاده و همکاران (۶) به بررسی عوامل خطرزای قلبی و عروقی در مردان سالمند فعال و کم‌تحرک پرداختند. آن‌ها این پژوهش را در سال ۱۳۹۴ در شهر رشت بر روی دو گروه افراد سالمند فعال و سالمند کم‌تحرک انجام دادند. آن‌ها دریافتند که عوامل خطرزای بیماری‌های قلبی و عروقی در سالمندان فعال در مقایسه با سالمندان کم‌تحرک کمتر است ولی اکثر سالمندان فعال نیز دارای حداقل یک یا چندین عامل خطرزای قلبی و عروقی هستند. در سال‌های اخیر نقش عوامل روانی از قبیل استرس، اضطراب و افسردگی در بروز و تشدید بیماری‌های قلبی و عروقی بیش از پیش مورد توجه قرار گرفته است. برای این منظور فلاح و همکاران (۷) در پژوهشی به بررسی نقش عوامل روانی و فعالیت فیزیکی بر خطر ابتلا به بیماری‌های قلبی با استفاده از معادلات ساختاری در افراد بزرگسال شهرستان یزد پرداخته‌اند. آن‌ها دریافتند که شیوع افسردگی، اضطراب و استرس در شهر یزد بالا است و فعالیت فیزیکی نقش میانجی بین سازه‌های روانشناختی و خطر ابتلا به بیماری‌های قلبی را دارد. بیماری‌های قلبی و عروقی تحت تاثیر گروهی از عوامل قابل تعدیل و غیر قابل تعدیل از جمله سن بالا، دیابت، فشارخون بالا، سابقه فامیلی، شاخص نمایه‌ی توده بدنی و افزایش وزن است. محمدی و همکاران (۸) در پژوهشی به بررسی برآورد سهم دیابت بر بار بیماری‌های قلبی و عروقی درشهرستان یزد با استفاده از داده‌های دیابت موجود در پایگاه داده مطالعه یاس پرداختند. نتیجه این مطالعه نشان داد، سهم بیماری دیابت در کاهش بار قابل انتساب بیماری قلبی و عروقی در زنان در صورتی‌که شیوع دیابت به صفر برسد، برابر ۲۳/۶ درصد و هنگامی‌که شیوع دیابت در زنان به متوسط کشوری سال ۱۳۸۸ معادل ۲۰/۳ درصد رسانیده شود، برابر ۰/۳ درصد است. سهم بیماری دیابت در کاهش بار قابل انتساب بیماری قلبی و عروقی در مردان شهرستان یزد اگر شیوع دیابت به صفر برسد برابر ۱۰/۶ درصد و در صورتی که شیوع دیابت به متوسط کشوری معادل ۱۷/۷ درصد رسانیده شود برابر ۱/۲ درصد است. در پژوهشی دیگر میرزایی و همکاران (۹) به بررسی عوامل خطر قابل تعدیل بیماری‌های ایسکمیک قلبی در بین ۱۰۰۰۰ نفر از ساکنان پنج منطقه بزرگ شهرداری شهر یزد در محدوده سنی ۲۰-۶۹ سال پرداختند. آن‌ها پس از تجزیه و تحلیل اطلاعات افراد دریافتند که عادات غذایی ناسالم و کم‌تحرکی شایع‌ترین عوامل خطرزای قابل تعدیل بیماری قلب و عروق در

[DOI: 10.18502/ssu.v32i7.16571]

شهر یزد هستند. در پژوهشی دیگر استوارفر و همکاران (۱۰) با شناسایی ریسک قابل انتساب جمعیت (PAR) به بررسی عوامل اصلی خطرزای بیماری‌های قلبی و عروقی پرداختند. نویسندگان این پژوهش بعد از بررسی PAR و فاصله معتبر بیزی مربوطه دریافتند که معمول‌ترین عوامل خطر برای بیماری‌های قلبی و عروقی، فعالیت بدنی ناکافی و چاقی شکمی است. سن به عنوان یک عامل غیرقابل کنترل، قوی‌ترین عامل در تعیین بیماری‌های قلبی است. هم‌چنین بیماری‌های قلبی و عروقی عمدتا به فشارخون بالا نسبت داده می‌شوند. در مجموع از موارد مطرح شده می‌توان دریافت که عواملی مانند کم‌تحرکی، مولفه‌های روانی، سن، افزایش وزن، دیابت، فشارخون، سابقه فامیلی و تغذیه ناسالم در بیماری‌ها قلبی موثر هستند. این عوامل طی سال‌ها بر اساس تجربه افراد خبره و متخصص در زمینه بیماری قلبی و محققان با انجام آزمایش‌ها و با صرف زمان و هزینه به‌دست آمده است. هدف از پژوهش حاضر شناسایی عوامل غیرآزمایشگاهی موثر بر بیماری‌های ایسکمیک قلبی در یک جمعیت بزرگ ساکنان شهرستان یزد با استفاده از ابزارهای هوش مصنوعی و با صرف زمان و هزینه خیلی کم است.

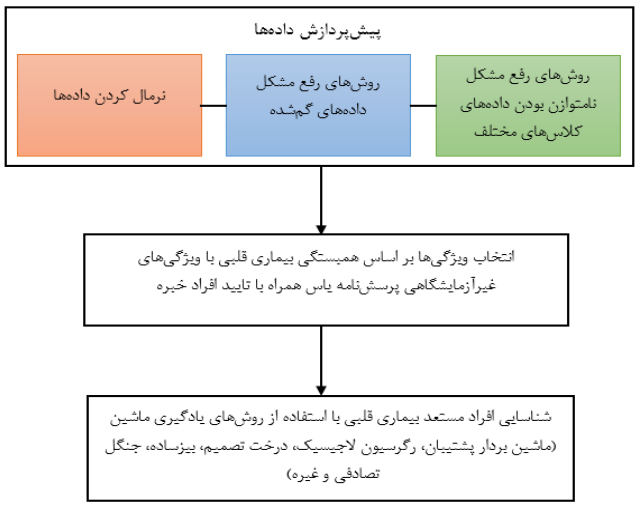
روش بررسی

این مطالعه با استفاده داده‌های کوهورت آینده‌نگر سلامت مردم یزد (یاس) که روی ۱۰۰۰۰ نفر از مردم یزد از سال ۱۳۹۳ درحال انجام است، اجرا شده است. برای این منظور ابتدا از طریق هم‌بستگی (Correlation) ویژگی‌های موجود در پایگاه داده یاس با بیماری قلبی، مهم‌ترین ویژگی‌ها استخراج شده و سپس با استفاده از الگوریتم‌های مختلف یادگیری ماشین افراد مستعد بیماری قلبی شناسایی شده‌اند. در طرح یاس ابتدا یک پرسش‌نامه شامل ۳۰۰ سوال مختلف چندگزینه‌ای در مورد اطلاعات سلامت و بیماری افراد با استفاده از برخی از پرسش‌نامه‌های معتبر جهانی از قبیل IPAQ فعالیت بدنی، DASS-21 روانشناسی و Rose تشخیص بیماری قلبی آماده و اعتبارسنجی شده است. سپس سوالات پرسش‌نامه از تقریبا ۱۰۰۰۰ نفر از ساکنان شهر یزد به‌صورت حضور در منازل افراد طی سال‌های ۱۳۹۴-۱۳۹۳ پرسیده و اطلاعات آنها جمع‌آوری شده است. لازم به ذکر است اطلاعات این افراد، هر پنج سال یکبار مورد ارزیابی مجدد قرار می‌گیرند. جدول ۱ تعداد سوالات مربوط به قسمت‌های مختلف پرسش‌نامه یاس را نشان می‌دهد. نمونه‌گیری مطالعه یاس به‌صورت خوشه‌ای و بر اساس کدپستی انجام شده است. برای این منظور ۲۰۰ خوشه ۵۰ نفری انتخاب شد. جامعه مورد بررسی، افراد ۵ گروه سنی با حد فاصله ۱۰ سال از ۲۰ سال تا ۷۰ سال ساکن شهرستان یزد بوده‌اند. در هنگام نمونه‌گیری سعی شده است تا حد امکان تعداد افراد هر گروه سنی نزدیک به هم باشند. در حین نمونه‌گیری علاوه بر تکمیل پرسش‌نامه، فشارخون، قد و وزن افراد نیز اندازه‌گیری شده است. در پایان نمونه‌گیری، به افراد دعوت‌نامه‌ای جهت حضور در آزمایشگاه مرکزی و تحویل نمونه خون برای انجام آزمایشات خون و ذخیره در بیوبانک برای کارهای تحقیقاتی آینده داده شده است. تنها حدود ۴۰ درصد از افراد در مرحله اول به آزمایشگاه مراجعه نموده و نمونه خون آن‌ها جمع‌آوری شد (۱۱). در سال‌های اخیر، هوش مصنوعی به عنوان یکی از پیشرفت‌های علوم فناوری، به‌صورت گسترده در حوزه‌های مختلفی از جمله علوم پزشکی مورد استفاده قرار گرفته است. یکی از جنبه‌های جالب کاربرد هوش مصنوعی در علوم پزشکی، توانایی قابل‌توجه در پیش‌بینی دقیق و زودهنگام بیماری‌ها و شناسایی عوامل موثر بر آنهاست. الگوریتم‌های هوش مصنوعی می‌توانند با تحلیل داده‌های بزرگ و اطلاعات پزشکی بیماران، الگوهای پنهان در داده‌ها را شناسایی کرده و به پزشکان کمک کنند تا با اقدامات پیشگیرانه، احتمال وقوع بیماری‌ها را کاهش دهند و عوامل موثر بر بیماری‌ها را شناسایی کنند. این اقدامات می‌توانند بهبود سطح سلامتی جامعه را به دنبال داشته باشند. از اینرو هدف این پژوهش استفاده از هوش مصنوعی در تشخیص عوامل موثر بر بیماری قلبی و شناسایی زودهنگام افراد مستعد این بیماری است. پس از بررسی پایگاه داده یاس، ابتدا خلاصه‌ای از روش پیشنهادی در شکل ۱ آورده شد. در ادامه هر یک از مراحل روش پیشنهادی با جزئیات آورده شده است.

[DOI: 10.18502/ssu.v32i7.16571]

جدول ۱: دامنه ویژگی‌های پایگاه داده یاس

عنوان سوالات	شماره سوال‌های پرسشنامه (Test)
سوالات اولیه (سن، جنس، مدت اقامت درمحل، بومی بودن یا نبودن، میزان تحصیلات، تعداد اعضا خانوار و وضعیت تاهل)	۸-۱
فشارخون سیستولیک، فشارخون دیاستول، محدوده نمایه توده بدنی	۱۱-۹
سوالات فعالیت بدنی	۱۷-۱۲
سوالات وضعیت خواب	۲۸-۱۸
سوالات روان	۳۲-۲۹
سوالات وضعیت روحی	۵۳-۳۳
سوالات کیفیت زندگی	۶۱-۵۴
سوالات بیماری‌های مزمن(بیماری قلبی، فشارخون، دیابت، کلسترول خون و غیره)	۱۵۷-۶۲
سوالات سابقه جراحی	۱۷۹-۱۵۸
سوالات دهان و دندان	۱۸۶-۱۸۰
سوالات حوادث	۱۹۶-۱۸۷
سوالات عادات غذایی	۲۳۳-۱۹۷
سوالات وضعیت شغلی	۲۴۱-۲۳۴
سوالات طب سنتی	۲۵۲-۲۴۲
سوالات وضعیت مصرف دخانیات و مواد مخدر	۲۶۳-۲۵۳
سایر سوالات	۲۷۳-۲۶۴
سوالات ویژه زنان	۳۰۰-۲۷۴



شکل ۱: روند نمای روش پیشنهادی جهت شناسایی عوامل موثر بر بیماری قلبی و افراد مستعد این بیماری

[DOI: 10.18502/ssu.v32i7.16571]

پیش‌پردازش داده‌ها

مرحله پیش‌پردازش همانطور که در شکل ۱ آورده شده است شامل ۳ مرحله است:

۱- نرمال‌کردن داده‌ها: در این مرحله برای اینکه همه ویژگی‌ها در دامنه یکسانی قرار بگیرند و تاثیر یکسانی در تشخیص داشته باشند از روش‌های نرمال‌سازی از قبیل Standard Scaler و Robust Scaler استفاده شده است (۱۲).

۲- روش‌های حل مشکل داده‌های گم‌شده (Missing values): در این مرحله ابتدا ویژگی‌هایی که درصد بالایی مقدار خالی (Null) دارند حذف شده و سپس مقادیر خالی هر ویژگی بر اساس مقدار مد هر ویژگی و با توجه به برچسب‌های رکوردهای با کلاس یکسان پر شده‌اند.

۳- روش‌های رفع مشکل نامتوازن بودن داده‌های کلاس‌های مختلف: براساس نتایج به‌دست آمده از بررسی داده‌های پایگاه یاس، پایگاه داده یاس نامتوازن تشخیص داده شده است. بزرگ‌ترین مشکل پایگاه داده نامتوازن، سوق داشتن طبقه‌بند به کلاس با داده اکثریت است. برای حل مشکل نامتوازن بودن در این پژوهش دو روش با نام Oversampling و Smote استفاده شده است. در روش Oversampling، نمونه‌های کلاس حداقل کپی و تکثیر می‌شود تا به تعداد نمونه‌های کلاس حداکثر نزدیک شود. در این روش در واقع هر نمونه کلاس حداقل چندین بار تکرار می‌شود (۱۳). در روش Smote بر خلاف روش اول به جای کپی کردن نمونه‌های کلاس حداقل، با تولید نمونه‌های جدیدی در همسایگی نمونه‌های موجود، متوازن‌سازی داده انجام می‌شود. این روش با استفاده از مفهوم الگوریتم K نزدیک‌ترین همسایگی و با اندازه‌گیری فاصله‌ها چند نمونه مشابه را انتخاب کرده و با استفاده از آن‌ها و در همسایگی آن‌ها نمونه جدید را ایجاد می‌کند (۱۴).

انتخاب ویژگی

مرحله بعدی پس از انجام پیش‌پردازش‌های لازم و آماده‌سازی داده‌ها انتخاب ویژگی‌های مهم و موثر برای تشخیص سالم یا بیمار بودن نمونه‌هاست. برای این منظور از مفهوم هم‌بستگی (correlation) با استفاده از تابع Dataframe.corr موجود در کتابخانه Pandas استفاده شده است. این تابع از روش هم‌بستگی مشهور پیرسون استفاده می‌کند و میزان و نوع وابستگی دو متغیر یا ویژگی را نشان می‌دهد (۱۵). در این پژوهش از مفهوم هم‌بستگی برای بررسی میزان وابستگی تمام ستون ویژگی‌ها با ستون مربوط به بیماری قلبی استفاده شده است. عدد به‌دست‌آمده از همبستگی پیرسون بین ۱ و ۱- متغیر است. اگر عدد به‌دست آمده برابر ۱ باشد بیانگر رابطه مستقیم کامل بین دو متغیر (دو ویژگی) است. رابطه مستقیم بدین معناست که اگر یکی از متغیرها افزایش (کاهش) یابد، دیگری نیز افزایش (کاهش) می‌یابد. اگر عدد به‌دست آمده برابر ۱- باشد نشان دهنده رابطه غیر مستقیم بین دو متغیر است یعنی با افزایش یک متغیر، متغیر دیگر کاهش پیدا می‌کند. اگر عدد به‌دست‌آمده به صفر نزدیک شود، نشان‌دهنده این است که بین دو متغیر رابطه خطی وجود ندارد (۱۶). در پایان پس از استخراج ویژگی‌های مهم به شناسایی افراد مستعد بیماری و سالم با استفاده از انواع روش‌های یادگیری ماشین پرداخته شده است. برای این منظور در اینجا توضیح مختصری از روش‌های یادگیری ماشین استفاده شده آورده شده است.

ماشین بردار پشتیبان (*Support Vector Machine (SVM)*): این روش یکی از مشهورترین روش‌های یادگیری ماشین برای طبقه‌بندی است که عمومیت بسیار بالایی را ایجاد می‌کند. ایده اصلی در این روش یافتن ابرصفحه جداکننده است، به صورتی‌که بیش‌ترین فاصله بین ابرصفحه و نمونه کلاس‌ها ایجاد شود و در نتیجه عمومیت مدل افزایش یابد. این روش ابتدا برای پیدا کردن مرز خطی بین کلاس‌ها مطرح شد، اما این روش در مسائل پیچیده با فضای ویژگی بالا نیز با استفاده از مفهوم کرنل در پیدا کردن مرزهای غیرخطی عملکرد بسیار مناسبی دارد (۱۶).

درخت تصمیم (*Decision tree*): این طبقه‌بندی فضای داده آموزشی را به‌صورت سلسله مراتبی تقسیم می‌کند. این

[DOI: 10.18502/ssu.v32i7.16571]

روش به‌طور مکرر مجموعه داده‌ها را بر اساس معیاری که جداسازی را حداکثر می‌کند، تقسیم می‌نماید. که در آن از شرط بر روی مقدار ویژگی‌ها برای تقسیم داده‌ها استفاده می‌شود تا داده‌ها به‌درستی در گره‌های برگ قرار داده شوند (۱۶,۱۷).

بیز ساده (*Naïve bayes*): این روش یکی از رایج‌ترین و ساده‌ترین روش‌های طبقه‌بند آماری است که بر اساس تئوری بیز عمل می‌کند. در این روش از قضیه بیز برای محاسبه و پیش‌بینی احتمال یک ویژگی مشخص‌شده متعلق به یک کلاس خاص استفاده می‌شود. در بیز ساده ویژگی‌ها مستقل از هم در نظر گرفته می‌شوند (۱۷).

رگرسیون لجستیک (*Logistic regression*): این روش برخلاف نامش به‌منظور طبقه‌بندی استفاده می‌شود. در این روش احتمال عضویت داده در هر یک از کلاس‌ها محاسبه می‌شود و نمونه به کلاسی تعلق می‌گیرد که ماکزیمم احتمال را دارد (۱۸).

K - نزدیک‌ترین همسایه‌ها (*K-Nearest Neighbors (KNN)*): از ساده‌ترین روش‌های طبقه‌بندی است که به‌طور مستقیم از روی داده‌ها و بدون ساخت مدل طبقه‌بندی را انجام می‌دهد. تنها پارامتر قابل تنظیم برای این روش k که تعداد نقاط همسایه است، می‌باشد. برای طبقه‌بندی برچسب داده موردبررسی بر اساس برچسب اکثریت k داده نزدیک به آن تعیین می‌شود (۱۹).

جنگل تصادفی (*Random forest*): یکی دیگر از الگوریتم‌های یادگیری ماشین، الگوریتم جنگل تصادفی است. از این الگوریتم هم در مسائل کلاسه‌بندی و هم مسائل رگرسیون استفاده می‌شود. الگوریتم جنگل تصادفی از خاصیت یادگیری گروهی (Ensemble learning) استفاده می‌کند و از مجموعه‌ای از درختان کم‌عمق تشکیل شده است. نتیجه نهایی با استفاده از رای‌گیری نتایج درختان کم‌عمق ساخته شده مشخص می‌شود. این خاصیت باعث شده تا الگوریتم جنگل تصادفی در برابر نمونه‌های نویزدار و مسائل با داده نامتوازن عملکرد مناسبی داشته باشد (۱۷).

آدابوست (*Adaboost*): روش آدابوست از ترکیب چند مدل ضعیف و با تقویت کردن آن‌ها برای حل مسائل پیچیده ساخته شده است. در این روش مدل‌ها به‌صورت سلسله مراتبی آموزش می‌بینند. هر مدل هدفش رفع ایرادات مدل‌های قبلی است و تمرکزش روی نمونه‌هایی است که مدل‌های قبلی نتوانسته‌اند به درستی طبقه‌بندی کنند. برای این منظور وزن نمونه‌هایی که درست طبقه‌بندی شده اند کمتر می‌شود و وزن نمونه‌هایی که اشتباه طبقه‌بندی شده بیشتر می‌شود. با این‌کار مدل بعدی متوجه می‌شود که کجا باید تمرکز کند و سعی کند چه نمونه‌هایی را درست طبقه‌بندی کند (۲۰).

شبکه عصبی عمیق (*Deep Neural Network*): یادگیری عمیق زیرمجموعه‌ای از یادگیری ماشین است و امروزه در مسائل مختلفی کاربرد دارد. شبکه عصبی عمیق، معماری توسعه‌یافته شبکه عصبی مصنوعی (*Artificial Neural Network (ANN)*) است که با هدف شبیه‌سازی عملکرد نرون‌های مغز انسان برای یادگیری طراحی شده است. شبکه عصبی مصنوعی پایه از سه لایه‌ی اصلی، یک لایه ورودی، یک لایه پنهان و یک لایه خروجی تشکیل شده است. این شبکه ورودی‌ها را با استفاده از وزن‌هایی که لایه‌ها را به هم مرتبط می‌کنند به‌خوبی به خروجی مناسب نگاشت می‌دهد. در واقع مقدار خروجی به‌صورت تابعی از ورودی‌ها به‌دست می‌آید. با افزایش تعداد لایه‌های پنهان می‌توان شبکه‌های عصبی عمیق را به‌وجود آورد. در پژوهش حاضر برای پیاده‌سازی روش پیشنهادی از زبان برنامه‌نویسی پایتون، از کتابخانه‌ی Pandas برای پردازش پایگاه داده یاس، از کتابخانه Scikit-learn برای پیاده‌سازی انواع روش‌های یادگیری ماشین و از کتابخانه Tensorflow برای طراحی شبکه عصبی عمیق پیشنهادی استفاده شده است

نتایج

برای بررسی نتایج روش پیشنهادی ابتدا طبق شکل ۱ مراحل روش کار انجام شده است. برای این منظور ابتدا پایگاه داده یاس مورد بررسی قرار گرفته و پیش‌پردازش‌های لازم انجام شده است. دامنه اعداد ویژگی‌های موجود در پایگاه داده

[DOI: 10.18502/ssu.v32i7.16571]

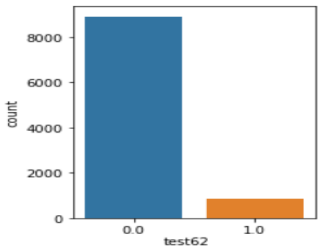
متفاوت است و با انجام نرمال‌سازی داده‌ها محدوده اعداد تمام ویژگی‌ها در دامنه یکسانی قرار داده شده است. سپس ویژگی‌هایی با بیش‌ترین مقدار Null حذف شده و مقادیر گم شده هر رکورد بر اساس مقدار مد ویژگی مورد نظر محاسبه شده است. و در نهایت روش‌های رفع مشکل نامتوازن بودن داده‌های کلاس‌های پایگاه داده یاس انجام شده است. پس از بررسی پایگاه داده یاس و ویژگی مربوط به بیماری قلبی این نتیجه به‌دست آمد که از ۹۹۶۵ نفر موجود در این پایگاه داده، تعداد ۸۵۴ نفر دارای برچسب بیماری ایسکمیک قلبی، تعداد ۸۹۱۳ نفر دارای برچسب سالم (عدد صفر برای افراد سالم در نظر گرفته شده است) و تعداد ۱۹۸ نفر فاقد برچسب هستند. با توجه به اعداد به‌دست آمده تعداد افراد کلاس سالم تقریبا ۱۰ برابر تعداد افراد کلاس بیمار است. برای رفع این مشکل دو روش smote و Oversampling که پیش‌تر توضیح داده شده، معرفی شده است. شکل ۲ نامتوازن بودن پایگاه داده یاس را نشان می‌دهد. در ادامه، مرحله استخراج ویژگی بعد از آماده‌سازی پایگاه داده یاس انجام شده است. پایگاه داده دارای ۳۰۰ ویژگی اطلاعات سلامت از حدود ۱۰۰۰۰ نفر مردم استان یزد است. برای تعیین ویژگی‌های موثر در بیماری قلبی از مفهوم همبستگی بین همه ویژگی‌ها با ویژگی مربوط به بیماری قلبی استفاده شده است و ۵۷ ویژگی با بیش‌ترین هم‌بستگی با بیماری قلبی انتخاب شده است که با تایید افراد خبره همراه بوده است. شکل ۳ نمودار قدرمطلق اندازه مقدار هم‌بستگی این ویژگی‌ها با ویژگی مربوط به بیماری ایسکمیک قلبی را نشان می‌دهد. همانطور که در شکل ۳ نشان داده شده است و با توجه به مطالعات انجام شده می‌توان نتیجه گرفت که سوالاتی از جمله سن به عنوان مهم‌ترین عامل، بیماری دیابت، فشارخون، کلسترول خون بالا، سابقه بیماری قلبی خانوادگی، کیفیت زندگی، فعالیت بدنی، کیفیت خواب، بیماری‌های روانی و تغذیه به عنوان عوامل مهم در بیماری قلبی شناسایی شده‌اند. جدول ۲ ویژگی‌ها یا سوالات مهم شناسایی شده بر اساس مفهوم هم‌بستگی با بیماری قلبی را نشان می‌دهد. مشاهده سوالات موثر بر بیماری قلبی و دسترسی به پرسش‌نامه پاس از طریق لینک http://www.yahs.ir امکان‌پذیر است.

پس از شناسایی ویژگی‌های موثر، داده‌های پایگاه داده به نسبت ۸۰ به ۲۰ در دو گروه آموزش (Train) و آزمایش (Test) تقسیم‌بندی شده و از الگوریتم‌های مختلف یادگیری ماشین برای دسته‌بندی ویژگی‌ها و شناسایی افراد سالم و افراد دارای بیماری قلبی استفاده شده است.

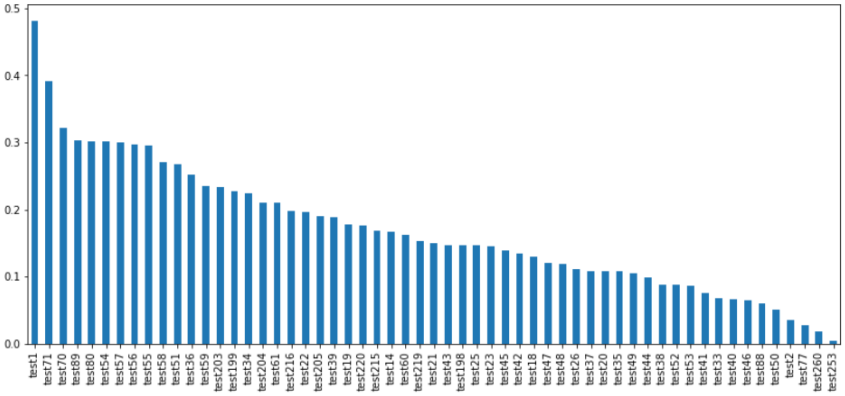
برای ارزیابی روش پیشنهادی از معیارهای ارزیابی مختلف مانند دقت، صحت و فراخوان استفاده شده است. این معیارها با استفاده از چهار مولفه ماتریس آشفتگی محاسبه شده‌اند (۲۱). جدول ۳ فرمول این معیارها را نشان می‌دهد.

در این فرمول‌ها مولفه TP (True Positive) نشان‌دهنده بیمارانی است که توسط روش پیشنهادی درست تشخیص داده شده‌اند. مولفه TN (True Negative) بیان کننده تعداد افراد سالمی است که روش پیشنهادی به درستی تشخیص داده است، مولفه FP (False Positive) تعداد افراد سالمی که روش پیشنهادی به اشتباه به عنوان بیمار شناسایی کرده و مولفه FN (False Negative) بیان‌کننده تعداد افراد بیماری که روش پیشنهادی به اشتباه به عنوان افراد سالم شناسایی کرده است. در این پژوهش سعی بر شناسایی حداکثری افراد مستعد بیماری قلبی است، از این رو معیار فراخوان بسیار مهم است. در ادامه ارزیابی روش پیشنهادی با استفاده از تکنیک Smote برای رفع مشکل نامتوازن بودن داده‌ها و انواع روش‌های یادگیری ماشین به عنوان طبقه‌بند در جدول ۴ آورده شده است. شکل ۴ سطح زیر نمودار بهترین طبقه‌بند پژوهش جاری را نمایش می‌دهد

[DOI: 10.18502/ssu.v32i7.16571]



شکل۲: نامتوازن بودن داده‌های پایگاه داده یاس



شکل۳: ویژگی‌های موثر در بیماری ایسکمیک قلبی بر اساس هم‌بستگی ویژگی‌ها با بیماری قلبی

جدول ۲: ویژگی‌های موثر انتخاب شده بر بیماری قلبی با استفاده از مفهوم هم‌بستگی ویژگی‌ها.

عنوان سوالات	شماره سوال‌ها (Test)
سوالات اولیه(سن، جنس، مدت اقامت درمحل، بومی بودن یا نبودن، میزان تحصیلات، تعداد اعضا خانوار و وضعیت تاهل)	۱ و ۲
سوالات فعالیت بدنی	۱۴
سوالات وضعیت خواب	۱۸، ۱۹، ۲۰، ۲۱، ۲۲، ۲۳، ۲۵ و ۲۶
سوالات وضعیت روحی	۵۳-۳۳
سوالات کیفیت زندگی	۶۱-۵۴
سوالات بیماری‌های مزمن(بیماری قلبی، فشارخون، دیابت، کلسترول خون و غیره)	۷۰، ۷۱، ۷۷، ۸۰، ۸۸، ۸۹
سوالات عادات غذایی	۱۹۸، ۱۹۹، ۲۰۳، ۲۰۴، ۲۰۵، ۲۱۵، ۲۱۶، ۲۱۹، ۲۲۰
سوالات وضعیت مصرف دخانیات و مواد مخدر	۲۵۳، ۲۶۰

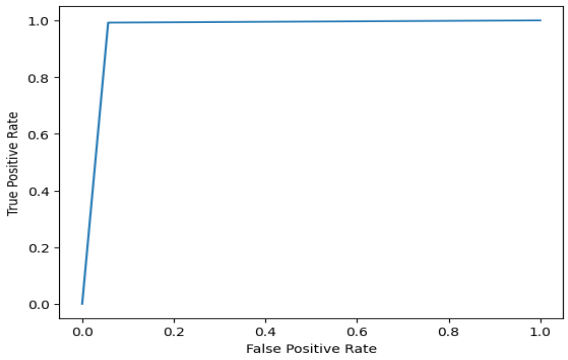
جدول ۳: معیارهای ارزیابی مورد استفاده جهت ارزیابی روش پیشنهادی

معیار	فرمول
دقت (Accuracy)	$ACC = \frac{TP+TN}{TP+TN+FP+FN}$
صحت (Precision)	$P = \frac{TP}{TP+FP}$
فراخوان (Recall)	$R = \frac{TP}{TP+FN}$

[DOI: 10.18502/ssu.v32i7.16571]

جدول ۴: نتایج حاصل از ارزیابی روش پیشنهادی با طبقه‌بندهای مختلف

طبقه‌بند	معیار دقت(%)	معیار صحت(%)	معیار فراخوان(%)
رگرسیون لجستیک	۷۷/۱۳	۷۶/۶۱	۷۸/۷۸
بیزساده	۷۱/۵۲	۷۳/۴۲	۶۸/۴۰
درخت تصمیم	۹۲/۰۶	۹۱/۴۳	۹۳/۰۰
نزدیک‌ترین همسایگی(K=5)	۸۲/۸۰	۷۴/۶۳	۹۹/۹۴
ماشین بردار پشتیبان(کرنل خطی)	۷۷/۴۴	۷۶/۶۷	۷۹/۵۶
ماشین بردار پشتیبان(کرنل غیر خطی(RBF))	۹۵/۰۸	۹۳/۳۸	۹۷/۱۶
پرسپترون چند لایه	۹۵/۷۹	۹۳/۰۲	۹۹/۱۱
جنگل تصادفی	۹۶/۰۴	۹۸/۷۶	۹۳/۳۴
آدابوست	۹۵/۰۰	۹۷/۹۵	۹۲/۳۹
شبکه عصبی عمیق(۸ لایه)	۹۵/۵۸	۹۵/۵۳	۹۹/۸۸



شکل۴: سطح زیر نمودار بهترین طبقه‌بند پژوهش جاری

بحث

هدف پژوهش جاری بررسی عوامل مهم و موثر در بیماری قلبی و شناسایی حداکثری افراد مستعد بیماری قلبی با استفاده از روش‌های یادگیری ماشین است. برای این منظور ابتدا بر روی پایگاه داده مطالعه سلامت مردم استان یزد پیش‌پردازش‌های لازم از قبیل نرمال‌سازی ویژگی‌ها و رفع مشکل داده‌های گم شده و نامتوازن بودن داده‌ها انجام شده است. دو روش Oversampling و Smote برای رفع مشکل نامتوازن بودن داده‌های کلاس‌ها وجود دارد. در پایگاه داده یاس تعداد ۸۵۴ نمونه با بیماری قلبی و ۸۹۱۳ نمونه سالم وجود دارد و نمونه‌های سالم تقریباً ۱۰ برابر نمونه‌های بیمار هستند. از آنجایی که روش Smote با ساخت داده‌های جدید از روی نمونه‌های کلاس حداقل، عمل متوازن‌سازی داده را انجام می‌دهد عملکرد نزدیک‌تری به واقعیت دارد از این‌رو در این پژوهش تنها نتایج این روش آورده شده است. سپس با استفاده از روش هم‌بستگی پیرسون ویژگی‌هایی که بیش‌ترین رابطه را با ویژگی بیماری قلبی دارند از جمله سن، دیابت، فشارخون، کلسترول خون، سابقه بیماری قلبی خانوادگی، سوالات کیفیت زندگی، کیفیت خواب، فعالیت بدنی، بیماری‌های روانشناسی، مصرف دخانیات و تغذیه ناسالم شناسایی شدند. در نهایت پس از استخراج ویژگی‌های مهم، نمونه‌ها با نسبت ۸۰ به ۲۰ به‌عنوان داده‌های آموزش و آزمایش به طبقه‌بندهای مختلف یادگیری ماشین داده شده‌اند. با توجه به نتایج به‌دست آمده این نتیجه به‌دست می‌آید که شبکه پرسپترون چندلایه و شبکه عصبی عمیق به ترتیب با مقادیر دقت ۹۵/۷۹ و ۹۷/۵۸ بهترین عملکرد را در شناسایی حداکثری افرا بیمار دارند. در این زمینه پژوهشی، مقاله مشابه دیگری (۲۲) بر روی پایگاه داده یاس برای شناسایی افراد بیمار با استفاده از روش‌های یادگیری ماشین توسط طباطبایی و همکاران انجام شده است. در این مقاله برای حل مشکل نامتوازن بودن داده‌های سالم و بیمار از

[DOI: 10.18502/ssu.v32i7.16571]

روش بوت استراپ (Bootstrap) استفاده شده است و تعداد داده‌ها متعادل شده‌اند. سپس تعداد ۱۰ سوال از پرسشنامه بر اساس نظر افراد خبره به‌عنوان ویژگی‌های مهم و موثر بر بیماری قلبی انتخاب شده است. در نهایت با استفاده از نرم‌افزار Rapidminer studio و طبقه‌بندهای مختلف موجود در یادگیری ماشین، شناسایی افراد مستعد به بیماری قلبی انجام شده است. بر اساس نتایج به دست آمده روش درخت تصمیم با دقت ۹۱ بهترین عملکرد را داشته است. مقاله دیگر توسط میلان کومار و همکاران (۲۳) بر روی پایگاه داده University of California, Irvine (UCI) برای شناسایی افراد سالم و دارای بیماری قلبی انجام شده است. این پایگاه داده شامل ۱۳ ویژگی آزمایشگاهی از ۳۰۳ نفر است. در این پایگاه داده تعداد افراد سالم و بیمار تقریبا برابر است و مشکل نامتوازن بودن داده‌ها وجود ندارد. برای شناسایی افراد سالم و بیمار از الگوریتم‌های شبکه عصبی مصنوعی، ماشین بردار پشتیبان و درخت تصمیم‌گیری استفاده شده است. بر اساس نتایج به‌دست آمده روش ماشین بردار پشتیبان با دقت ۸۴/۱ بهترین روش برای پیش‌بینی بیماری قلبی عروقی تشخیص داده شده است. در پژوهشی دیگر ملکی و همکاران (۲۴) به بررسی و شناسایی بیماری عروق کرونر با استفاده از ترکیب یک الگوریتم بهینه‌سازی و الگوریتم‌های یادگیری ماشین پرداخته‌اند. برای این منظور آنها از پایگاه داده UCI استفاده کردند. آن‌ها برای کاهش تعداد ویژگی‌ها از الگوریتم بهینه‌سازیHarris Hawks استفاده کردند و از بین ۱۳ ویژگی ۶ ویژگی مهم و موثر بر بیماری قلبی را انتخاب کردند. سپس با استفاده الگوریتم‌های یادگیری ماشین از قبیل شبکه عصبی مصنوعی، درخت تصمیم، نزدیک‌ترین همسایگی و ماشین بردار پشتیبان به پیش‌بینی بیماری پرداختند. بر اساس نتایج به‌دست آمده روش ماشین بردار پشتیبان با دقت ۹۰/۴ بهترین روش برای پیش‌بینی بیماری قلبی عروقی تشخیص داده شده است. از محدودیت‌های پژوهش جاری، می‌توان به سه مورد اشاره کرد: اولین مورد می‌توان به فقدان اطلاعات مربوط به مصرف الکل در پرسش‌نامه که از فاکتورهای خطر ابتلا به بیماری قلبی است، اشاره کرد. دومین محدودیت پژوهش، عدم پاسخ بیماران به برخی از ویژگی‌های مرتبط با فاکتورهای خطر مانند مدت زمان اعتیاد در پرسش‌نامه است. برای رفع این مشکل نمونه‌های فاقد مقدار حذف شده‌اند. سومین محدودیت تعداد زیاد ویژگی‌های انتخابی است، برای کاهش تعداد ویژگی‌ها می‌توان از الگوریتم‌های فرا ابتکاری استفاده کرد تا مهم‌ترین ویژگی‌ها شناسایی شوند.

نتیجه‌گیری

شناسایی عوامل موثر بر بیماری قلبی و افراد مستعد این بیماری برای پیش‌گیری و اتخاذ تصمیمات صحیح جهت ارتقائ سلامت در ایران وجهان ضروری است. در این پژوهش شناسایی افراد مستعد بیماری قلبی با استفاده از پایگاه داده های مطالعه سلامت مردم یزد انجام شد. برای این منظور ابتدا طی مرحله پیش‌پردازش و با استفاده از روش‌های متعادل‌سازی داده‌های مربوط به دو کلاس افراد سالم و بیمار متوازن شده است. در ادامه با استفاده از مفهوم هم‌بستگی بین ویژگی‌ها، مهم‌ترین فاکتورهای موثر در تشخیص بیماری از بین ۳۰۰ ویژگی غیرآزمایشگاهی شناسایی شده است. ویژگی‌های شناسایی شده با استفاده از انواع روش‌های یادگیری ماشین در دو کلاس سالم و جعل دسته‌بندی شده‌اند. از بین روش‌های یادگیری ماشین مختلف روش نزدیک‌ترین همسایگی با ۵ خوشه، روش شبکه عصبی عمیق و پرسپترون چند لایه با معیار فراخوان به مقادیر ۹۹/۹۴، ۹۹/۸۸ و ۹۹/۱۱ به ترتیب بهترین عملکرد را در شناسایی افراد بیمار داشته‌اند. با توجه به بررسی‌های انجام شده می‌توان دریافت که شناسایی عوامل موثر بر بیماری قلبی و افراد مستعد این بیماری با استفاده از روش‌های سنتی غربالگری و انجام آزمایش‌های مختلف از قبیل آنژیوگرافی و آزمایش استاندارد طلایی امری پرهزینه، زمان‌بر و گاها همراه با خطراتی برای فرد بیمار است. پژوهش جاری با استفاده از ویژگی‌های غیر آزمایشگاهی داده های مطالعه سلامت مردم یزد و روش‌های یادگیری ماشین بدون انجام آزمایشات پاراکلینیک و صرف هزینه توانسته است عملکرد مطلوبی را از خود نشان دهد. اپلیکیشن هماتاب از این روش جهت غربالگری بهره می‌برد.

[DOI: 10.18502/ssu.v32i7.16571]

سپاس‌گزاری

این پژوهش و مقاله حاصل از آن با استفاده از پژوهانه (گرنت شماره ۶۳۵۰۴) ششمین دوره "طرح هسته‌های مساله محور احمدی روشن" با عنوان: "استفاده از هوش مصنوعی در ایجاد و توسعه ابزار بر خط آموزش و مشاوره مجازی پیشگیری و خودمراقبتی بیماران قلبی" به راهبری دکتر مسعود میرزایی انجام شده است. بدینوسیله از بنیاد ملی نخبگان و حمایت‌های آن سپاس‌گزاری می‌شود.

حامی مالی: بنیاد ملی نخبگان(گرنت شماره ۶۳۵۰۴)

تعارض در منافع: وجود ندارد.

ملاحظات اخلاقی

پروپوزال این تحقیق توسط دانشگاه علوم پزشکی یزد تایید شده است (کد اخلاق IR.SSU.REC.1400.095).

مشارکت نویسندگان

همه نویسندگان در ارائه ایده، در طراحی مطالعه، در جمع‌آوری داده‌ها، در تجزیه و تحلیل داده‌ها مشارکت داشته و همه نویسندگان در تدوین، ویرایش اولیه و نهایی مقاله و پاسخگویی به سوالات مرتبط با مقاله سهیم هستند.

References:

1-Riazi-Isfahani S, Ghanbari Motlagh A, Hamelmann C. ***Iran's Status of NCDs Prevention and Management Services during COVID-19 Pandemic at PHC Level***. SJKU 2021; 26(5): 50-68. [Persian]

2-Mozaffari-Khosravi V, Mirzaei M, Mozaffari-Khosravi H. ***Prevalence of metabolic syndrome in adults in Yazd 2014-2015: results of Yazd Health Study (YaHS)***. JSSU 2019; 27(11): 2123-31. [Persian]

3-Etaat M, Tabatabaye Z, Jahromi S M, Yosefi P, Sedigh S, Tajiki S. ***Predictors of Blood Pressure in Iranian Women-A Narrative Review***. JSSU 2020; 28(8): 2889-2904. [Persian]

4-Mirzaei M, Sharifnia G, Khazaei Z, Sadeghi E, Fallahzadeh H, Namayandeh SM. ***Prevalence of General Obesity and Central Adiposity and Its Related Factors in Adult Population of Yazd***. JSSU 2017; 25(9): 736-47. [Persian]

5-Marzban A, Karkhaneh M. ***Evaluation of Knowledge and Attitude of Yazd University of Medical Sciences Students to Cigarette Smoking***. Journal of Preventive Medicine 2018; 5(1): 55-63. [Persian]

6-Najafgholizadeh H, Rahmaninia F, Mirzae B. ***Comparison of Some Cardiovascular Risk Factors between Active and Sedentary Elderly Men.*** JQUMS 2017; 21(1): 21-8. [Persian]

7-Fallah MH, Hosseini H, Fallahzadeh H, Mirzaei M. ***The Relationship between Depression, Anxiety, Stress, and Physical Activity with Cardiovascular Disease Risk, Using Structural Equation Modeling in Adults in Yazd City***. The Journal of Tolooebehdasht 2021; 20(3): 59-74. [Persian]

8-Mohammadi M, Mirzaei M, Karami M. ***Potential Impact Fraction of Ischemic Heart Disease Associated with Diabetes Mellitus in Yazd-Iran***. Iranian Journal of Epidemiology 2018; 13(4): 299-307. [Persian]

9-Mirzaei M, Mirzaei M, Sarsangi A R, Bagheri N. ***Prevalence of Modifiable Cardiovascular Risk Factors in Yazd Inner-City Municipalities***. BMC Public Health 2020; 20: 1-8.

10-Ostovarfar M, Fallahzadeh H, Askari M, Ostovarfar J, Mirzaei M. ***Population Attributable Risk (PAR) of***

[DOI: 10.18502/ssu.v32i7.16571]

cardiovascular diseases (CVD) Risk Factors; Bayesian Methods. J Adv Med Biomed Res 2021; 29(134): 161-6.

11- Mirzaei M, Salehi-Abargouei A, Mirzaei M, Mohsenpour M A. ***Cohort Profile: The Yazd Health Study (Yahs): A Population-Based Study of Adults Aged 20–70 Years (Study Design and Baseline Population Data)***. Int J Epidemiol 2018; 47(3): 697-8.

12- Ferreira P, Le DC, Zincir-Heywood N. ***Exploring Feature Normalization and Temporal Information for Machine Learning Based Insider Threat Detection***. In the 15th International Conference on Network and Service Management (CNSM) 2019; 1-7.

13- Mohammed R, Rawashdeh J, Abdullah M. ***Machine Learning with Oversampling and Undersampling Techniques: Overview Study and Experimental Results***. In Proceedings of the 2020 11th International Conference on Information and Communication Systems (ICICS), Irbid, Jordan, 7–9 April 2020; IEEE: Piscataway, NJ, USA, 2020; 243-2

14- Prasetiyo B, Muslim MA, Baroroh N. ***Evaluation Performance Recall and F2 Score of Credit Card Fraud Detection Unbalanced Dataset Using SMOTE Oversampling Technique***. Journal of Physics: Conference Series 2021; 1918(4): 1-5.

15- Benesty J, Chen J, Huang Y, Cohen I. ***Pearson Correlation Coefficient.*** In: Noise Reduction in Speech Processing, Springer, Heidelberg 2009; 37-40.

16- Rustam F, Khalid M, Aslam W, Rupapara V, Mehmood A, Choi GS. ***A Performance Comparison of Supervised Machine Learning Models for Covid-19 Tweets Sentiment Analysis***. Plos one 2021; 16(2): e0245909.

17- Ray S. ***a Quick Review of Machine Learning Algorithms***. International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (Com-It-Con) 2019; 35-9.

18- Dreiseitl S, Ohno-Machado L. ***Logistic Regression and Artificial Neural Network Classification Models: A Methodology Review***. Journal of Biomedical Informatics 2002; 35(5): 352-9.

19- Singh A, Thakur N, Sharma A. ***A Review of Supervised Machine Learning Algorithms***. In 2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom) 2016; 1310-15.

20- Mahesh B. ***Machine Learning Algorithms-A Review***. International Journal of Science and Research (IJSR) 2020; 9: 381-386.

21- Al-Qershiz OM, Khoo BE. ***Evaluation of Copy-Move Forgery Detection: Datasets and Evaluation Metrics***. Multimedia Tools and Applications 2018; 77(24): 31807-33.

22- Tabatabaei SMR, Saadatjoo F, Mirzaei M. ***The Prediction Model for Cardiovascular Disease Using Yazd's Health Study Data (Yahs)***. JSSU 2019; 27(3): 1346-60. [Persian]

23- Kumari M, Godara S. ***Comparative Study of Data Mining Classification Methods in Cardiovascular Disease Prediction.*** International J Computer Sci Trends Techno 2011; 2(2): 304 -8.

24- Maleki S, Zare Mehrjerdi Y, Shishebori D, Mirzaei M. ***Predicting Coronary Artery Diseases Using Effective Features Selected by Harris Hawks Optimization Algorithm and Support Vector Machine.*** Journal of Industrial and Systems Engineering, 14(Special issue: 18th International Industrial Engineering Conference) 2022; 14: 40-47.

[DOI: 10.18502/ssu.v32i7.16571]

Journal of Shahid Sadoughi University of Medical Sciences
Vol 32| NO 7 | Sep 2024

Received: 2 Oct 2023
Accepted: 3 July 2024

Identification of Effective Factors and Prediction of Ischemic Heart Disease Using Machine Learning Methods and Data from the Yazd Health Study (YaHS)"

Jamal Zarepour Ahmadabadi[*1], Fatemeh Zare Mehrjardi[2], Mahdieh Ghanbary[3], Masoud Mirzaei[4]

Original Article

Introduction: Ischemic heart diseases are one of the most common diseases that cause high mortality worldwide. This article has identified various factors affecting heart disease and identified susceptible people using various machine learning methods.

Methods: The current research was conducted on the Yazd Health Study (YaHS) database. YaHS was conducted on adults aged 20-70 years who were residents of Yazd Greater Area and collected information on the health and various diseases of nearly 10,000 people in the form of a questionnaire with 300 different questions. In this research, by using the correlation of questions with heart disease, the most important factors of heart disease have been identified. By using the identified factors and machine learning algorithms, susceptible people with heart disease have been identified.

Results: The results of the evaluations have shown that factors such as age, family history of heart disease, blood pressure, diabetes, blood cholesterol, stress, anxiety, depression, quality of life, quality of sleep, physical activity, smoking, and diet have an effect on heart disease. Likewise, among the different machine learning methods, the nearest neighbor method, the deep neural network method, and the multi-layer perceptron method with a recall criterion of 99.94%, 99.88%, and 99.11% have performed the best in the identifying sick people, respectively.

Conclusion: According to the findings of the research, it can be understood that by controlling factors such as blood pressure, diabetes, blood cholesterol, stress, anxiety, and depression, changing factors such as quality of life, sleep status, physical activity, and eating patterns of people and quitting smoking, it is possible to move towards improving the health of society. On the other hand, the identification of people prone to heart disease using machine learning methods is done faster and at a lower cost than the traditional methods that are done by referring to medical centers and doctors and performing various tests.

Keywords: Heart disease, Machine learning, Disease factors, Risk factors.

Citation: Zarepour Ahmadabadi J, Zare Mehrjardi F, Ghanbary M, Mirzaei M. **Identification of Effective Factors and Prediction of Ischemic Heart Disease Using Machine Learning Methods and Data from the Yazd Health Study (YaHS).** J Shahid Sadoughi Uni Med Sci 2024; 32(7): 8067-79.

[1]Department of Computer Science, Yazd University, Yazd, Iran.
[2]Department of Computer Engineering, Meybod University, Meybod, Yazd, Iran.
[3]Cardiac Rehabilitation Center, Yazd Cardiovascular Research Center, Non-Communicable Diseases Research Institute, Shahid Sadoughi University of Medical Sciences, Yazd, Iran.
[4]Non-Communicable Diseases Research Institute, Yazd Cardiovascular Research Center, Shahid Sadoughi University of Medical Sciences, Yazd, Iran.
***Corresponding author: Tel:** 09162539543**, email:** zarepourjamal@yazd.ac.ir

[DOI: 10.18502/ssu.v32i7.16571]