مقاله پژوهشی

مجله انفورماتیک سلامت و زیست پزشکی
مرکز تحقیقات انفورماتیک پزشکی
دوره هفتم شماره چهارم. ص ۳۶۷-۳۵۴، ۱۳۹۹

مطالعه موردی تأثیر سوابق بیماری‌های والدین در احتمال ابتلاء به فشارخون بالا با استفاده از تکنیک‌های داده‌کاوی

امین قادری[۱]، ناصر فرج‌زاده[۲]*، الهه بایبوردی[۳]

• دریافت مقاله: ۹۸/۸/۲۳ • پذیرش مقاله: ۹۹/۲/۸

مقدمه: بیماری «فشارخون بالا» از جمله شایع‌ترین عارضه‌های سلامتی به شمار می‌رود. از این بیماری به دلیل تأثیر آن در بروز بیماری‌های خطرناک‌تر دیگری از جمله بیماری‌های قلبی-عروقی و انواع سکته و همچنین نداشتن علائم و نشانه‌های خاص، به‌عنوان مرگ خاموش یا قاتل خاموش یاد می‌شود. ازاین‌رو تشخیص به موقع، کنترل و درمان آن اهمیت زیادی در سیستم‌های بهداشتی داشته و موجب پیشگیری از بروز بیماری‌های تأثیرپذیر از آن خواهد شد. براساس پژوهش‌های صورت گرفته، نشان داده شده است که رابطه محکمی بین سابقه برخی از بیماری‌ها در والدین و ابتلاء فرزندان آن‌ها به فشار خون بالا، وجود دارد.

روش: در این پژوهش با استفاده از داده‌های جمع‌آوری شده از دو مرکز بهداشت واقع در استان اردبیل و به‌کارگیری تکنیک‌های داده‌کاوی و نرم‌افزار Rapid Miner، تأثیر سوابق بیماری‌های والدین در احتمال ابتلاء به فشارخون بالا در فرزندان مورد بررسی قرار گرفته است.

نتایج: نتایج حاصل حاکی از این نکته ارزشمند است که استفاده از اطلاعات پیشینه بیماری والدین مانند فشارخون بالا، عروق کرونر زودرس، هیپرکلسترولمیا و تیروئید، نقش بسزایی در پیش‌بینی درست بیماری فشارخون در فرزندان دارد و باعث افزایش ۵٪ کارایی مدل پیش‌بینی کننده‌ای که از روی نمونه‌های جمع‌آوری شده و با استفاده از تکنیک‌های داده‌کاوی ایجاد شده است، می‌شود.

نتیجه‌گیری: با این که وجود برخی بیماری‌ها به خصوص فشارخون بالا در والدین خطر ابتلاء به فشارخون بالا در فرزندان را بیشتر می‌کند؛ اما در نهایت ویژگی‌های فردی تعیین کننده اصلی بروز این عارضه هستند.

کلید واژه‌ها: فشارخون، هوش مصنوعی، یادگیری ماشین، داده‌کاوی، پیش‌بینی، پیشینه بیماری والدین

•ارجاع: قادری امین، فرج‌زاده ناصر، بایبوردی الهه. مطالعه موردی تأثیر سوابق بیماری‌های والدین در احتمال ابتلاء به فشارخون بالا با استفاده از تکنیک‌های داده‌کاوی. مجله انفورماتیک سلامت و زیست پزشکی ۱۳۹۹؛ ۷(۴): ۶۷-۳۵۴.

۱. کارشناسی ارشد فناوری اطلاعات، دانشکده فناوری اطلاعات و مهندسی کامپیوتر، دانشگاه شهید مدنی آذربایجان، تبریز، ایران
۲. دانشیار، دانشکده فناوری اطلاعات و مهندسی کامپیوتر، دانشگاه شهید مدنی آذربایجان، تبریز، ایران
۳. متخصص پزشکی اجتماعی، استادیار، جهاد دانشگاهی، تبریز، ایران

*** نویسنده مسئول:** ناصر فرج‌زاده
آدرس: تبریز، ۳۵ کیلومتری جاده تبریز-مراغه، دانشگاه شهید مدنی آذربایجان، دانشکده فناوری اطلاعات و مهندسی کامپیوتر
•شماره تماس: ۳۱۴۵۲۰۴۴-۰۴۱
• **Email:** n.farajzadeh@azaruniv.ac.ir

مقدمه

فشارخون بالا عارضه‌ای است که از نشانه‌های همراه و گاهی ایجاد کننده بیماری‌های خطرناکی همچون قلبی-عروقی، نارسایی‌های مزمن کلیوی، دیابت و سکته مغزی به شمار می‌آید [۱]. طبق آمار سازمان بهداشت جهانی، فشارخون بالا عامل ۷/۵ میلیون مرگ (۱۲/۸٪ از کل مرگ‌ومیر جهانی) است. همچنین، طبق آمارهای این سازمان، حدود ۴۰٪ از افراد بالای ۲۵ سال به این عارضه مبتلا هستند [۲]. به همین دلیل، مطالعات وسیعی در خصوص شناسایی عوامل بروز این بیماری و راه‌های کنترل و درمان آن صورت گرفته است [۴، ۳]. به‌طورکلی این عوامل را می‌توان به چهار دسته تقسیم‌بندی کرد: ۱) عوامل رفتاری (سبک زندگی)، ۲) عوامل اجتماعی و اقتصادی، ۳) ابتلاء به برخی بیماری‌ها از قبیل بیماری کلیوی، بیماری‌های غدد درون‌ریز و ناهنجاری‌های سلول‌های خونی و ۴) عوامل ژنتیکی [۵].

داده‌های پزشکی و سلامتی که از مراقبت‌های مستمر و خدمات ارائه شده به بیماران ثبت و تولید می‌شوند، از اهمیت ویژه‌ای در بررسی و تحلیل انواع بیماری‌ها برخوردار هستند [۶]؛ اما از آنجایی که حجم این نوع داده‌ها معمولاً بسیار زیاد است و به عنوان داده‌های عظیم شناخته می‌شوند، نمی‌توان با استفاده از روش‌های دستی و مرسوم اقدام به بررسی و تحلیل چنین داده‌هایی کرد. از این‌رو، سعی می‌شود با بهره‌گیری از روش‌های هوشمندی مانند داده‌کاوی و یادگیری ماشین، اطلاعات ارزشمندی را که در این داده‌ها پنهان شده‌اند، استخراج و در جهت بهبود کیفیت تشخیص‌های پزشکی و برنامه‌ریزی‌های مدیریتی در حوزه سلامت، به کار برد [۸، ۷].

در سال‌های اخیر، استفاده از داده‌کاوی برای استخراج الگوهای پنهان در داده‌ها، گسترش چشم‌گیری داشته است [۹] و طبق آمارهای ارائه شده، ۱۷/۲٪ از کل کاربردهای داده‌کاوی در سال ۲۰۱۸ را کاربردهای سلامتی و پزشکی به خود اختصاص داده و در رده دوم بیشترین استفاده از داده‌کاوی قرار گرفته است [۱۰]. بر همین اساس، سازمان بهداشت جهانی از سال ۱۹۹۷ توانایی بالقوه داده‌کاوی در پیشرفت علم پزشکی را به رسمیت شناخته و بر استفاده از دانش استخراج شده از داده‌های پزشکی و نقش آن در پیش‌بینی و تشخیص و پیشگیری از وقوع بیماری‌ها، تأکید کرده است [۱۱].

استفاده از دادکاوی در تشخیص و یا پیش‌بینی بیماری‌های مرتبط با فشارخون نیز مورد توجه بسیاری از پژوهشگران قرار گرفته است. برای مثال، در پژوهش Shin و همکاران [۱۲]، با به‌کارگیری روش کاوش قوانین وابستگی، رابطه بین فشارخون بالا، دیابت نوع ۲ و آسیب مغزی، مورد مطالعه و بررسی قرار گرفته است. در مطالعه‌ای [۱۳]، پژوهشگران با هدف ایجاد سیستم‌های تصمیم‌یار در حوزه پزشکی، رهیافتی سه مرحله‌ای شامل کاهش ابعاد داده‌ها (انتخاب ویژگی)، وزن‌دهی و تعیین اولویت ویژگی‌ها و در نهایت کلاس‌بندی، ارائه کرده‌اند. راهکار ارائه شده توسط آن‌ها توانست با دقت ۹۲/۵۹٪ و ۸۱/۸۲٪ به ترتیب بیماری قلبی و بیماری هپاتیت را تشخیص دهد. در مطالعه‌ای دیگر [۱۴]، شیوع فشارخون بالا در جمعیت هدف (ایالت تلانگانای هند) مورد ارزیابی قرار گرفته و نشان داده شده است که فشارخون بالا وابستگی آماری قابل‌توجهی با مصرف دخانیات، توده بدنی بالا، چاقی شکمی، مصرف بالای نمک و عدم فعالیت فیزیکی دارد. Xu و همکاران [۱۵]، مدل‌های متفاوتی از پیش‌بینی کننده‌های خطرات فشارخون بالا را با به‌کارگیری روش‌های مختلف داده‌کاوی توسعه داده و کارایی آن‌ها را مورد بررسی قرار داده‌اند.

بیشتر مطالعات انجام شده درباره عوامل بروز و پیش‌بینی فشارخون بالا در رابطه با عوامل مرتبط با سبک زندگی بوده است؛ عواملی که کنترل آن‌ها و اقدام در چهارچوب توصیه‌های بهداشتی علاوه بر فشارخون، احتمال وقوع سایر بیماری‌ها را نیز کاهش می‌دهد. بدیهی است که عوامل محیطی و سبک زندگی نقش بسزایی در بروز فشارخون بالا دارند. با این وجود، بررسی فردی، صرف‌نظر از محیط و سبک زندگی، می‌تواند اطلاعات مهمی در خصوص احتمال ابتلاء به فشارخون را آشکار سازد. از جمله این موارد می‌توان به نقش وراثت اشاره کرد.

بنابراین، به نظر می‌رسد که تعیین احتمال ابتلاء به فشارخون در سنین جوانی با تکیه بر اطلاعات وراثتی به دلیل تمرکز اقدامات پیشگیرانه بر روی افرادی که بیشتر در معرض خطر هستند، حائز اهمیت باشد. به عبارتی، می‌توان با بررسی اطلاعات وراثتی، خطر ابتلاء به فشارخون بالا را حتی قبل از تولد فرد تشخیص داده و اقدامات پیشگیرانه را به کار بست.

مطالعات متعدد نشان داده است که انجام اقدامات پیشگیرانه با تمرکز بر افراد در معرض خطر، به طور اساسی موجب تأخیر یا حتی جلوگیری از بروز فشارخون بالا شود [۱۷، ۱۶]. همچنین تشخیص زودهنگام، کنترل و درمان فشارخون بالا، مزایای زیادی برای سلامتی داشته و باعث جلوگیری از اتلاف سرمایه‌های ملی و فردی می‌شود و با توجه به دخالت

عوامل مختلف محیطی از قبیل شرایط زندگی و کار در بروز فشارخون بالا در سنین بالا، می‌توان چنین نتیجه گرفت که مطالعه در مورد یافتن راهکارهایی که بتواند ابتلاء به این بیماری را در سنین پایین پیش‌بینی کند، ارزش دوچندانی دارد.

در همین راستا، در این پژوهش سعی شده است تا با استفاده از روش‌های داده‌کاوی، تأثیر سابقه برخی بیماری‌های والدین در ابتلاء فرزندان‌شان به بیماری فشارخون بالا، مورد بررسی قرار گیرند.

در این پژوهش با بررسی و مقایسه تأثیر سابقه ابتلای والدین به چهار بیماری مختلف، میزان تأثیر آن‌ها در بروز فشارخون بالای فرزندان اندازه‌گیری شد. علاوه‌براین، با مقایسه تأثیر این سوابق با برخی از ویژگی‌های فردی متأثر از سبک زندگی، سعی در تعیین اولویت ویژگی‌های پیش‌بینی‌کننده و ارائه راه‌حل می‌شود.

تاکنون هیچ مطالعه مشابهی در خصوص مقایسه تأثیر بیماری‌های مختلف به‌خصوص با استفاده از روش‌های داده‌کاوی و همچنین تعیین میزان دقیق تأثیر هر ویژگی بر پیش‌بینی ابتلاء به فشارخون بالا، انجام نشده است.

روش

جمع‌آوری داده: برای انجام این مطالعه از داده‌های ثبت شده مراجعین با بازه‌ی سنی ۱۸ تا ۲۹ سال در سامانه‌های اطلاعاتی بیمارستانی (HIS) دو مرکز بهداشت شهید پیله‌رودی و شهید جدی استان اردبیل از اسفند سال ۹۵ تا خرداد ۹۸، استفاده شده است. تعداد کل نمونه‌های جمع‌آوری شده ۱۰۶۲ مورد است که از این تعداد ۸۹۰ نمونه متعلق به دسته (کلاس) فشارخون طبیعی، ۱۱۳ نمونه متعلق به کلاس پیش فشارخون بالا و ۵۹ نمونه متعلق به کلاس فشارخون بالا بودند (جدول ۱).

پس از جمع‌آوری داده‌های خام مورد نظر، با استفاده از روش‌های پیش‌پردازش داده‌ها از قبیل حذف ویژگی‌های غیرضروری، گسسته‌سازی و حذف داده‌های پرت، اقدام به پالایش و نرمال‌سازی آن‌ها شد. به دلیل نامتوازن بودن توزیع اعضاء در کلاس‌های هدف (مشابه اغلب داده‌های برگرفته از دنیای واقعی)، با استفاده از روش‌های متداول داده‌کاوی، تأثیر منفی این عدم توازن در کیفیت طبقه‌بندی حذف شد. در انتها، از چند الگوریتم شناخته شده در داده‌کاوی استفاده و سعی شد عملکرد آن‌ها در پیش‌بینی ابتلاء به فشارخون بالا در فرزندان با و بدون در نظر گرفتن سوابق بیماری والدین، مقایسه و بررسی شود (شکل ۱).

شکل ۱: فرآیند اجرای روش پیشنهادی برای بررسی تأثیر سوابق بیماری والدین در پیش‌بینی خطر ابتلاء به بیماری فشارخون بالا در فرزندان

پیش‌پردازش

از آنجایی که برای محاسبه شاخص (Body Mass Index) BMI از قد و وزن افراد استفاده می‌شود و این شاخص دید بهتری از وضعیت جسمی افراد ارائه می‌دهد، به کارگیری دو ویژگی قد و وزن به تنهایی غیرضروری بوده و به همین دلیل، این دو ویژگی از داده‌ها حذف شد. از طرفی ویژگی‌های فشارخون سیستولیک و دیاستولیک، در واقع نه یک ویژگی، بلکه شاخصی برای تعیین مقدار ویژگی «تشخیص» هستند که به کمک طبقه‌بندی ارائه شده در مطالعه Chobanian و همکاران [۱۸] برای تعیین ابتلای فرد به فشارخون بالا استفاده شد. علاوه‌براین، در برخی موارد مقادیر فشارخون سیستولیک و دیاستولیک برای بیماران ثبت نشده داده‌های گم‌شده (Missing Data) و صرفاً به درج تشخیص براساس آن‌ها بسنده شده است. ازاین‌رو، این دو ویژگی نیز از مجموعه ویژگی‌ها حذف شدند.

با توجه به این که مقادیر مربوط به BMI و سن مراجعین مقادیر پیوسته‌ای هستند، این مقادیر برای پردازش‌های بعدی مطابق جدول ۲ و ۳ گسسته‌سازی می‌شوند. در اینجا منظور از

گسسته‌سازی، گروه‌بندی و تخصیص یک عدد صحیح به هر گروه است. همچنین با استفاده از ویژگی‌های فشار سیستولیک و دیاستولیک ثبت‌شده، طبقه‌بندی فشارخون طبق جدول ۴ انجام شد [۱۸].

جدول ۱: ویژگی‌های جمع‌آوری شده مربوط به داده‌های مورد آزمایش در این پژوهش

ردیف	عنوان ویژگی	مقادیر/ توضیحات
۱	شناسه	به صورت عددی یکتا
۲	جنس	زن: ۰ مرد: ۱
۳	سن	عددی صحیح بین ۱۸ و ۲۹
۴	وزن	عددی با یک اعشار
۵	قد	عددی با یک اعشار
۶	شاخص توده بدنی (BMI)	عددی با دو اعشار
۷	فشارخون سیستولیک	عدد صحیح
۸	فشارخون دیاستولیک	عدد صحیح
۹	حداقل یکی از والدین به بیماری پرفشاری خون مبتلا است	خیر:۰ بله: ۱
۱۰	حداقل یکی از والدین به بیماری عروق کرونری زودرس (مردان زیر ۵۵ سال و زنان زیر ۶۵ سال) مبتلا است	خیر:۰ بله:۱
۱۱	حداقل یکی از والدین به بیماری هیپرکلسترولمیا مبتلا است	خیر:۰ بله: ۱
۱۲	حداقل یکی از والدین به بیماری تیروئید (پرکاری) مبتلا است	خیر:۰ بله: ۱
	تشخیص	فشارخون نرمال: ۱ پیش فشارخون: ۲ فشارخون بالا: ۳

جدول ۲: طبقه‌بندی BMI مراجعین

BMI	طبقه تخصیص داده شده	توضیح
BMI <۱۸/۵	۱	کمبود وزن
۱۸/۵≤ BMI<۲۴/۹	۲	وزن نرمال
۲۵≤ BMI<۲۹/۹	۳	اضافه وزن
BMI≥۳۰	۴	چاقی

جدول ۳: طبقه‌بندی سن مراجعین

سن	طبقه تخصیص داده شده
۱۸ تا ۲۱	۱
۲۲ تا ۲۵	۲
۲۶ تا ۲۹	۳

جدول ٤: طبقه‌بندی فشارخون براساس معیارهای فشار سیستولیک و دیاستولیک

طبقه فشارخون	فشارخون سیستولیک (mmHg)		فشارخون دیاستولیک (mmHg)
نرمال	کمتر از ۱۲۰	و	کمتر از ۸۰
افزایش یافته (پیش فشارخون بالا)	۱۲۰ تا ۱۲۹	و	کمتر از ۸۰
فشارخون بالا	۱۳۰ به بالا	یا	۸۰ به بالا

شایان ذکر است از آنجایی که پیش فشارخون بالا از فاکتورهای اساسی ابتلاء به فشارخون بالا بوده و احتمال بروز فشارخون بالا در افرادی که به پیش فشارخون بالا مبتلا هستند، بسیار بالا است [۲۱-۱۸] و همچنین به دلیل کمبود تعداد رکوردهای کلاس‌های پیش فشارخون بالا و فشارخون بالا، این دو کلاس با یکدیگر ادغام و تحت کلاس «فشارخون بالا» مورد بررسی قرار می‌گیرند.

مقابله با تأثیر کلاس‌های نامتوازن

در کاربردهای دنیای واقعی معمولاً اندازه کلاس‌های مختلف با یکدیگر اختلاف زیادی دارند [۲۲]. این موضوع در داده‌های پزشکی و سلامتی به دلیل تعداد بسیار کم وقوع شرایط بیماری نسبت به سلامتی، شایع‌تر است [۲۳]. نامتوازن بودن کلاس‌ها موجب می‌شود که الگوریتم‌های کلاس‌بندی عملکرد مناسبی نداشته و نتایج به دست آمده از آن‌ها قابل اعتماد نباشند. برای مثال اگر از ۱۰۰۰ نمونه، ۹۰۰ نمونه متعلق به کلاس A و ۱۰۰ نمونه متعلق به کلاس B باشد و الگوریتمی کلیه نمونه‌ها را از کلاس بزرگ‌تر (A) تشخیص دهد، بدین‌ترتیب دقت این الگوریتم ۹۰٪ خواهد بود که به وضوح عدم قابلیت اطمینان به نتایج آن مشهود است.

راهکارهای متنوعی برای برخورد با مسئله داده‌های نامتوازن معرفی شده است [۲۲]. نمونه‌برداری مجدد یکی از کارآمدترین و ساده‌ترین این راهکارها است. این راهکار خود دارای سه روش متفاوت است که عبارت‌اند از: نمونه‌برداری کمتر از کلاس بزرگ‌تر US (Under-Sampling)، نمونه‌برداری بیشتر از کلاس کوچک‌تر OS (Over-Sampling) و نمونه‌برداری ساختگی SS (Synthetic-Sampling) [۲۴، ۲۳].

برای بررسی کارایی هر یک از این سه روش متوازن‌سازی داده‌ها در بررسی سوابق بیماری والدین در پیش‌بینی ابتلاء به فشارخون بالا در فرزندان، سه مجموعه داده متفاوت با استفاده از هر کدام از این روش‌ها ایجاد شد. به این ترتیب که، نخست با پیروی از روش اول (US)، به تعداد نمونه‌های کلاس کوچک‌تر (نمونه‌های مبتلا به فشارخون بالا)، نمونه‌هایی را به صورت تصادفی از کلاس بزرگ‌تر (نمونه‌های سالم) انتخاب شد. سپس، با پیروی از روش نمونه‌برداری بیشتر از کلاس کوچک‌تر (OS)، تعداد نمونه‌های کلاس کوچک‌تر را چهار بار کپی کرده و به مجموعه کلاس کوچک‌تر اضافه شد. با این کار تعداد نمونه‌های کلاس کوچک‌تر (۷۷۵ نمونه) تقریباً برابر با تعداد نمونه‌های کلاس بزرگ‌تر (۸۰۱ نمونه) شد. در انتها، با استفاده از روش Synthetic Minority Over-sampling Technique (SMOTE) [۲۵]، به تعداد ۸۰۱ نمونه ساختگی از کلاس کوچک‌تر ایجاد کرده و به کلاس کوچک‌تر اضافه شد. به این ترتیب، تعداد نمونه‌های کلاس کوچک‌تر با کلاس بزرگ‌تر برابر می‌شود (جدول ۵).

جدول ۵: خلاصه‌ای از مشخصات مجموعه‌های داده‌ای ایجاد شده

عنوان مجموعه داده‌ای	شرح	تعداد نمونه از کلاس کوچک‌تر (مبتلایان به فشارخون بالا)	تعداد نمونه از کلاس بزرگ‌تر (افراد سالم)	تعداد کل نمونه‌ها
داده‌های اولیه	داده‌های خام جمع‌آوری شده	۱۷۲	۸۹۰	۱۰۶۲
کاهش یافته (US)	نمونه‌برداری به تعداد کم از کلاس بزرگ‌تر	۱۵۵	۱۶۰	۳۱۵
افزایش یافته (OS)	تکرار نمونه‌برداری از کلاس کوچک‌تر تا زمان متوازن شدن اعضای کلاس‌ها	۷۷۵	۸۰۱	۱۵۷۶
داده‌های ساختگی (SS)	تولید نمونه‌های ساختگی از کلاس کوچک‌تر تا زمان متوازن شدن اعضای کلاس‌ها	۸۰۱	۸۰۱	۱۶۰۲

معیارهای ارزیابی

با توجه به میزان حساسیت موضوع پژوهش، باید از معیارهای مناسبی برای ارزیابی کارایی طبقه‌بندی کننده‌ها، استفاده شود. به عنوان مثال، معیارهایی مانند دقت و نرخ خطا در مورد داده‌های نامتوازن، معیارهای درستی به شمار نمی‌آیند [۲۶]. از این‌رو، با پیروی از پژوهش‌های ارائه شده [۲۹-۲۶، ۲۲، ۲۳] که بر روی کلاس‌های نامتوازن صورت گرفته‌اند، از معیارهای F_1-score، G-mean (Geometric Mean) و Area Under Curve (AUC) برای ارزیابی عملکرد الگوریتم‌های کلاس‌بندی استفاده شد. روابط ۱ تا ۳، نحوه محاسبه هر یک از این معیارها را نشان می‌دهد.

$$F1 = \frac{2\,TP}{2\,TP + FP + FN} \quad (۱)$$

$$G - mean = \sqrt{\frac{TP}{TP + FN} \times \frac{TN}{TN + FP}} \quad (۲)$$

$$AUC = \frac{1}{2}\left(\frac{TP}{TP + FN} + \frac{TN}{TN + FP}\right) \quad (۳)$$

در روابط ۱ الی ۳ موارد زیر وجود دارد:

TP (True Positive): نمونه مورد نظر بیمار است و توسط مدل به درستی بیمار تشخیص داده شده است.

TN (True Negative): نمونه مورد نظر سالم است و توسط مدل به درستی سالم تشخیص داده شده است.

FP (False Positive): نمونه مورد نظر سالم است و توسط مدل به اشتباه بیمار تشخیص داده شده است.

FN (False Negative): نمونه مورد نظر بیمار است و توسط مدل به اشتباه سالم تشخیص داده شده است.

الگوریتم‌های مورد استفاده

در این پژوهش، با استفاده از داده‌های جمع‌آوری شده و یکی از پنج الگوریتم پایه‌ای داده‌کاوی ۱) بیز ساده (Naïve Bayes) NB، ۲) درخت تصمیم (Decision Tree) DT، ۳) استنتاج قاعده (Rule Induction) RI، ۴) کا-نزدیک‌ترین همسایه (KNN) و ۵) ماشین بردار پشتیبان (Support Vector Machine) SVM، مدل‌های پیش‌بینی کننده متفاوتی ساخته شده و عملکردشان با استفاده از معیارهای ذکر شده، مورد بررسی و ارزیابی قرار می‌گیرد. هر کدام از این الگوریتم‌ها، ویژگی‌های مشخصی دارند که به آن‌ها برای استفاده در کاربردهای مختلف ارجحیت می‌بخشد [۷، ۸، ۳۰]. هدف این مطالعه از انتخاب و استفاده از این الگوریتم‌ها، ارائه نتایج پایه‌ای برای پژوهش حاضر به کمک رایج‌ترین الگوریتم‌های طبقه‌بندی است تا بتوان در کارهای آتی به آن اتکا کرده و سعی در بهبود کارایی آن‌ها نمود.

آزمایش‌ها

برای بررسی درستی فرضیه این پژوهش، از نرم‌افزار داده‌کاوی رپیدماینر (Rapid Miner) و روش اعتبارسنجی متقابل ده‌قسمتی (10-fold Cross Validation) استفاده شد. در اعتبارسنجی متقابل ده-قسمتی، داده‌های گردآوری شده به صورت تصادفی به ۱۰ زیرمجموعه تقریباً مساوی تقسیم شدند. سپس نه زیرمجموعه از ۱۰ زیرمجموعه برای ساخت مدل پیش‌بینی کننده بر اساس یکی از الگوریتم‌های نامبرده شده، اختصاص داده شد و یک مجموعه باقی‌مانده جهت ارزیابی و محاسبه معیاری‌های ذکر شده، مورد استفاده قرار گرفت. این عمل ۱۰ بار تکرار شده و در هر تکرار یکی از زیر مجموعه‌ها به ترتیب برای اعتبارسنجی انتخاب شد. در انتها، نتایج به دست آمده، میانگین‌گیری شده و گزارش شدند.

برای ارزیابی میزان تأثیر ویژگی‌هایی مرتبط با سوابق بیماری والدین در پیش‌بینی ابتلاء به فشارخون بالا در فرزندان، سه آزمایش به شرح زیر ترتیب داده شد:

۱) استفاده از همه ویژگی‌ها (اطلاعات فردی و سوابق بیماری والدین)

۲) استفاده از فقط ویژگی‌های فردی

۳) استفاده از فقط ویژگی‌های سوابق بیماری والدین

در ادامه، نتایج به دست آمده از ارزیابی الگوریتم‌ها با در نظر گرفتن سه مجموعه ویژگی، ارائه شد.

آزمایش اول

در این آزمایش از تمامی ویژگی‌های در دسترس برای آموزش و ساخت مدل‌های مختلف با الگوریتم‌های متفاوت، استفاده شده است. نتایج عملکرد مدل‌های ساخته شده در جدول ۶ نشان داده شده است.

جدول ٦: نتایج آزمایش اول (استفاده از تمامی ویژگی‌ها)

الگوریتم	داده‌های اولیه			داده‌های متوازن شده								
				ساختگی (SS)			افزایش یافته (OS)			کاهش یافته (US)		
	F_1	G-Mean	AUC	F_1	G-Mean	AUC	F_1	G-Mean	AUC	F_1	G-Mean	AUC
NB	۲۲/۱۳	۳۸/۶۱	۰/۵۵۳	۳۷/۵۷	۶۴/۴۶	۰/۶۶۴	۴۱/۲۶	۶۶/۹۴	۰/۶۸۰	۴۳/۴۱	۷۲/۲۸	۰/۷۳۴
DT	۰/۰۰	۰/۰۰	۰/۵۰۰	۲۸/۱۵	۱۶/۳۳	۰/۵۰۸	۱۶/۷۱	۳۸/۲۴	۰/۴۹۶	۳۴/۰۲	۶۰/۶۱	۰/۶۱۲
RI	۰/۰۰	۰/۰۰	۰/۵۰۰	۳۳/۱۵	۶۰/۰۰	۰/۶۰۴	۳۲/۳۸	۵۶/۱۶	۰/۵۹۷	۳۴/۵۸	۶۱/۹۸	۰/۶۲۲
KNN	۲۱/۵۷	۴۲/۷۵	۰/۵۳۲	۲۶/۱۹	۵۲/۶۸	۰/۵۳۱	۱۸/۹۳	۳۹/۷۵	۰/۵۱۸	۱۷/۶۵	۳۸/۳۹	۰/۵۱۰
SVM	۰/۰۰	۰/۰۰	۰/۵۰۰	۳۸/۱۱	۶۵/۷۱	۰/۶۶۶	۳۹/۵۸	۶۷/۴۱	۰/۶۷۹	۳۹/۴۶	۶۷/۲۸	۰/۶۷۸

مقادیر جدول ۶ نشان می‌دهد که بسته به روش مورد استفاده در توازن داده‌ها، الگوریتم‌ها کارایی متفاوتی از خود نشان می‌دهند. به‌عبارتی، می‌توان چنین نتیجه گرفت که کارایی مدل ساخته شده هم به الگوریتم مورد استفاده و هم به روشی که برای متوازن کردن داده‌ها استفاده می‌شود، بستگی دارد. همچنین ملاحظه می‌شود که عملکرد الگوریتم‌ها در استفاده از مجموعه داده‌های اولیه، بسیار ضعیف است در حالی که، نتایج مربوط به مجموعه US، در هر سه معیار بیشترین مقدار را دارد. با توجه به نتایج ارائه شده در این آزمایش، الگوریتم NB بر روی مجموعه US، به‌عنوان بهترین الگوریتم برای آزمایش اول انتخاب می‌شود.

آزمایش دوم

در این آزمایش، فقط از ویژگی‌های فردی (جنس، سن و شاخص BMI) داده‌های جمع‌آوری شده برای آموزش و ساخت مدل استفاده می‌شود. جدول ۷ نتایج این آزمایش را نشان می‌دهد.

با توجه به این جدول، نتیجه می‌شود که الگوریتم SVM با اعمال بر روی مجموعه OS، بهترین عملکرد را دارد. همچنین ملاحظه می‌شود که عملکرد این الگوریتم روی مجموعه US با اختلاف ناچیزی در رتبه دوم قرار دارد. در نتیجه برای آزمایش دوم مجموعه داده‌ای افزایش یافته به کمک الگوریتم SVM به عنوان بهترین عملکرد انتخاب می‌شوند. نتایج این آزمایش نیز نشان می‌دهد که استفاده از هر یک از سه مجموعه داده‌ای متوازن شده، باعث بهبود عملکرد مدل می‌شود.

آزمایش سوم

در آزمایش سوم و آخر، فقط از ویژگی‌های سابقه بیماری والدین استفاده می‌شود. نتایج حاصل از این آزمایش در جدول ۸ ارائه شده است. مقادیر جدول ۸، برتری عملکرد الگوریتم SVM را در مقایسه با سایر الگوریتم‌ها برای مجموعه‌های داده‌ای متوازن شده، نشان می‌دهد. الگوریتم SVM برای هر سه مجموعه داده‌ای متوازن شده، عملکرد مشابهی (با اختلاف ناچیز) نشان می‌دهد. با این‌حال، می‌توان این سه مجموعه داده‌ای را از نظر معیارهای ارزیابی طبقه‌بندی به ترتیب OS، SS و US، در نظر گرفت. مشابه دو آزمایش قبل، مجموعه داده‌ای اولیه نیز با الگوریتم KNN، ضعیف‌ترین عملکرد را دارد. در این آزمایش، الگوریتم SVM و مجموعه داده‌ای US، به‌عنوان بهترین الگوریتم و روش توازن، انتخاب می‌شوند.

جدول ۷: نتایج آزمایش دوم (استفاده از فقط ویژگی‌های فردی)

الگوریتم	داده‌های خام			روش‌های ایجاد کلاس‌های متوازن								
				ساختگی (SS)			افزایش یافته (OS)			کاهش یافته (US)		
	F_1	G-Mean	AUC	F_1	G-Mean	AUC	F_1	G-Mean	AUC	F_1	G-Mean	AUC
NB	۱۶/۲۹	۳۱/۷۸	۰/۵۳۵	۳۶/۸۹	۶۳/۶۵	۰/۶۵۶	۳۸/۰۴	۶۵/۶۶	۰/۶۶۵	۳۷/۸۸	۶۵/۳۰	۰/۶۶۵
DT	۰/۰۰	۰/۰۰	۰/۵۰۰	۲۸/۰۵	۱۴/۹۰	۰/۵۰۵	۲۰/۱۲	۴۴/۶۹	۰/۴۹۳	۳۲/۰۵	۵۷/۸۳	۰/۵۹۴
RI	۰/۰۰	۰/۰۰	۰/۵۰۰	۳۵/۱۴	۶۲/۰۲	۰/۶۲۳	۲۸/۳۵	۵۰/۹۲	۰/۵۷۰	۳۳/۴۳	۶۰/۷۵	۰/۶۰۹
KNN	۲۰/۰۰	۴۱/۳۳	۰/۵۲۱	۲۷/۸۳	۵۴/۵۷	۰/۵۴۸	۱۸/۹۳	۳۹/۷۵	۰/۵۱۸	۱۷/۷۸	۳۹/۱۷	۰/۵۰۵
SVM	۰/۰۰	۰/۰۰	۰/۵۰۰	۳۳/۸۰	۶۰/۹۸	۰/۶۱۱	۳۹/۵۹	۶۷/۳۹	۰/۶۸۲	۳۹/۵۳	۶۷/۳۱	۰/۶۸۲

جدول ۸: نتایج آزمایش سوم (استفاده از فقط ویژگی سابقه بیماری والدین)

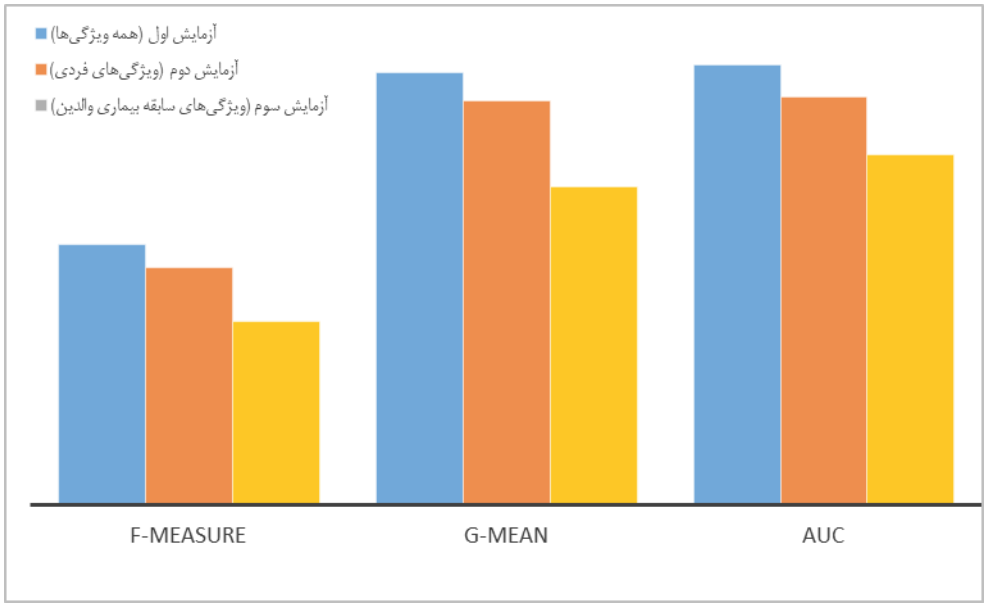
الگوریتم	داده‌های خام			روش‌های ایجاد کلاس‌های متوازن								
				ساختگی (SS)			افزایش یافته (OS)			کاهش یافته (US)		
	F_1	G-Mean	AUC	F_1	G-Mean	AUC	F_1	G-Mean	AUC	F_1	G-Mean	AUC
NB	۹/۱۴	۲۲/۶۷	۰/۵۱۷	۲۸/۱۴	۲۹/۸۱	۰/۵۱۳	۲۶/۹۸	۴۹/۰۹	۰/۵۱۶	۲۷/۰۸	۵۱/۶۳	۰/۵۲۴
DT	۰/۰۰	۰/۰۰	۰/۵۰۰	۲۸/۰۵	۱۴/۹۰	۰/۵۰۵	۲۰/۱۲	۴۴/۹	۰/۴۹۳	۰/۰۰	۰/۰۰	۰/۵۰۰
RI	۰/۰۰	۰/۰۰	۰/۵۰۰	۲۵/۶۹	۵۲/۰۳	۰/۵۳۰	۲۱/۴۱	۴۳/۰۰	۰/۵۲۹	۲۳/۶۸	۴۹/۰۶	۰/۴۹۱
KNN	۲۰/۰۶	۴۱/۳۵	۰/۵۲۲	۲۳/۴۷	۴۹/۴۷	۰/۵۰۶	۱۸/۹۳	۳۹/۷۵	۰/۵۱۸	۱۷/۸۳	۳۹/۲۰	۰/۵۰۶
SVM	۳/۱۶	۱۳/۰۹	۰/۵۰۰	۳۰/۲۴	۵۲/۵۶	۰/۵۸۳	۳۰/۴۹	۵۳/۱۹	۰/۵۸۴	۲۹/۸۷	۵۲/۹۰	۰/۵۸۰
میانگین	۸/۷۰	۲۲/۹۲	۰/۵۰۸	۲۷/۲۵	۵۱/۹۷	۰/۵۲۸	۲۴/۲۳	۴۹/۴۰	۰/۵۲۸	۲۳/۲۴	۴۸/۲۴	۰/۵۲۰

تحلیل نتایج آزمایش‌ها

همچنان که پیش‌تر نیز ذکر شد، هدف از این مطالعه، بررسی تأثیر استفاده از اطلاعات بیماری‌های والدین در تشخیص و پیش‌بینی ابتلاء به بیماری فشارخون بالا در فرزندان است. برای این کار، موارد بهترین عملکرد سه آزمایش با یکدیگر مقایسه شد (شکل ۲).

در آزمایش اول، الگوریتم NB با اعمال بر مجموعه US، بهترین عملکرد را از خود نشان داد. در آزمایش دوم الگوریتم SVM با مجموعه OS، بالاترین رتبه را با توجه به مقادیر معیارهای ارزیابی، کسب کرد. در آزمایش سوم نیز، مشابه آزمایش دوم، الگوریتم SVM با مجموعه OS، بهترین عملکرد را داشت.

مقایسه این سه آزمایش به وضوح نشان می‌دهد که نتایج به دست آمده از آزمایش اول، نسبت به دو آزمایش دیگر در هر سه معیار بهتر است. از طرفی نتایج دو آزمایش اول و دوم نزدیک به هم و با اختلاف نسبتاً زیادی با آزمایش سوم هستند که این نشان دهنده اهمیت و تأثیر زیاد ویژگی‌های فردی در پیش‌بینی ابتلاء به فشارخون بالا است.



شکل ۲: مقایسه نتایج به دست آمده از سه آزمایش ترتیب داده شده

میزان تأثیر هریک از ویژگی‌ها

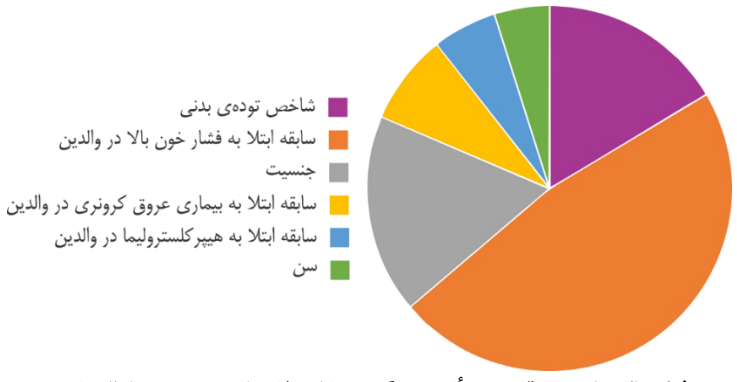
در این قسمت میزان تأثیر هر یک از ویژگی‌ها در عملکرد مدل پیش‌بینی کننده، مورد بررسی قرار می‌گیرد [۳۱،۳۲]. به همین منظور، از معیاری به‌نام بهره اطلاعاتی که روی مجموعه US محاسبه می‌گردد، استفاده می‌شود. بهره اطلاعاتی معیاری برای ارزیابی ویژگی‌ها است که بر اساس بی‌نظمی‌تعریف می‌شود؛ بی‌نظمی یک متغیر نشان دهنده میزان تصادفی بودن و به همین ترتیب ناشناختگی مقادیر آن است. به این ترتیب، با افزایش بی‌نظمی، اطلاعاتی که از تعیین قطعی مقدار آن متغیر (ویژگی) به دست می‌آید، بیشتر می‌شود. نتیجه این محاسبات در جدول ۹ و خلاصه آن در شکل ۳، ارائه شده است.

همان‌طور که قابل پیش‌بینی بود، جدول ۹ یکسان نبودن تأثیر انواع بیماری در والدین را نشان می‌دهد. برای نمونه، سابقه ابتلاء به فشارخون بالا در والدین تقریباً ۸ برابر تعیین کننده‌تر از هیپرکلسترولمیا در پیش‌بینی احتمال بروز فشارخون فرزندان است. در حالی که، بیماری تیروئید والدین هیچ تأثیری در عملکرد پیش‌بینی ندارد.

دقت در نتایج ارائه شده در شکل ۲ و جدول ۹، این واقعیت را آشکار می‌سازد که با وجود تأثیر غیرقابل‌انکار عوامل ژنتیکی و سابقه بیماری والدین در ابتلاء فرزندان به بیماری فشارخون، عوامل محیطی و سبک زندگی نقش تعیین‌کننده‌تری در پیش‌بینی ابتلاء به این بیماری بازی می‌کنند که این موضوع با نتایج پژوهش‌های Manios و همکاران و Xu و همکاران [۳۳،۳۴] همخوانی دارد؛ با این‌که فرزندان افراد مبتلا به بیماری‌های مورد بررسی، بیشتر در معرض خطر فشارخون بالا هستند، اما انتخاب‌های روزمره و محیطی که در آن قرار دارند، تعیین کننده نهایی ابتلاء آن‌ها به بیماری فشارخون بالا هستند. در واقع هدف از این تحقیق، اثبات تأثیر عوامل وراثتی در پیش‌بینی ابتلاء به فشارخون بالا به کمک روش‌های داده‌کاوی است تا به وسیله آن افراد در معرض خطر شناسایی شده و اقدامات پیشگیرانه صورت پذیرد. در نتیجه، این افراد با انتخاب سبک زندگی فعال، تغذیه مناسب و دوری از محیط‌های پرتنش و استرس‌زا، می‌توانند از وقوع این بیماری جلوگیری کنند.

جدول ۹: مقایسه میزان تأثیر هر یک از ویژگی‌ها با استفاده از معیار بهره اطلاعاتی

عنوان ویژگی	مقدار نرمال شده بهره اطلاعاتی (درصد)
شاخص توده بدنی	۱۶/۴
سابقه ابتلا به فشارخون بالا در والدین	۴۷/۴
جنسیت	۱۷/۶
سابقه ابتلا به بیماری عروق کرونری در والدین	۸/۰
سابقه ابتلا به هیپرکلسترولمیا در والدین	۵/۷
سن	۴/۹
سابقه ابتلا به تیروئید در والدین	۰



شکل ۳: خلاصه مقایسه تأثیر ویژگی‌ها با استفاده از معیار بهره اطلاعاتی

بحث و نتیجه‌گیری

در پژوهش حاضر سعی شد با استفاده از پنج الگوریتم پایه‌ای داده‌کاوی، نقش سابقه بیماری والدین در ابتلاء به فشارخون بالا در فرزندان بررسی شود. به همین منظور، اطلاعات فردی و سابقه بیماری والدین افراد ۱۸ تا ۳۰ سال (جمع‌آوری شده از دو مرکز بهداشت استان اردبیل)، مورد استفاده قرار گرفت. این رده سنی از آن جهت دارای اهمیت است که اولاً شانس کنترل و مقابله با وقوع این عارضه بیشتر است؛ دوماً، سایر فاکتورهای محیطی از جمله استرس، تأثیر کمتری در این برهه از زندگی افراد دارند. از طرفی، احتمال بروز فشارخون بالا در این سنین (به دلیل سبک زندگی فعال‌تر) نسبتاً کم بوده و در صورت ابتلاء، تأثیر عوامل ژنتیکی و وراثتی بیشتر است [۳۵، ۲۱].

نتایج ارائه شده در این پژوهش، تأثیر استفاده از اطلاعات سوابق بیماری والدین در افزایش درصد درست پیش‌بینی ابتلاء فرزندان به بیماری فشارخون بالا را تأیید می‌کند؛ استفاده از اطلاعات سوابق بیماری والدین، عملکرد دسته‌بندی را در حدود ۵٪ بهبود می‌بخشد که با نتایج به دست آمده مطالعات متعددی [۳۹-۳۶،۴] همخوانی دارد. از طرفی با تحلیل نتایج به دست آمده، لزوم استفاده از برخی اطلاعات فردی از جمله جنسیت، سن و شاخص توده بدنی نیز در پیش‌بینی ابتلاء به فشارخون بالا آشکار می‌گردد. البته تأثیر این ویژگی‌های فردی در بروز فشارخون بالا یکسان نیست. با استفاده از معیار بهره اطلاعاتی، مقدار این تأثیرات محاسبه گردید و نشان داده شد که دو ویژگی شاخص توده بدنی و ابتلای حداقل یکی از والدین به فشارخون بالا، بیشترین تأثیر و ابتلاء حداقل یکی از والدین به بیماری تیروئید، کمترین تأثیر (ناچیز) را داشت.

لازم به ذکر است که براساس آخرین مطالعات حاضر، تاکنون پژوهشی مشابه با پژوهش حاضر در مورد بررسی تأثیر سابقه بیماری والدین در ابتلای فرزندان به فشار خون بالا انجام نشده است. با این حال می‌توان با تفکیک جنبه‌های مختلف پژوهش حاضر به دو بخش ۱) عوامل بروز فشار خون و ۲) استفاده از داده‌کاوی، مقایسه‌ای با سایر پژوهش‌های صورت گرفته در این خصوص انجام داد.

عوامل بروز فشارخون بالا

Burke و همکاران در مطالعه‌ای [۴۰] بر روی جمعیتی ۹ تا ۱۸ ساله، تأثیر سابقه فشارخون بالا را در ابتلای فرزندان به این بیماری بررسی کردند. این مطالعه تأثیر زیاد فشارخون بالای والدین در پیش‌بینی فشارخون سیستولیک فرزندان به این بیماری را (حتی در ۹ سالگی) نشان داد. همچنین نقش این اطلاعات را در پیش‌بینی فشارخون فرزندان به صورت مستقل از ویژگی‌هایی مثل BMI و تناسب اندام، آشکار ساخت. نقطه قوت این مطالعه تفکیک تأثیر جنسیت والدین مبتلا به فشارخون بالا در بروز آن در فرزندان از جنس مشخص است. Goldstein و همکاران [۳۵] نیز با بررسی فشارخون ۲۲۰ نفر طی دو دوره ۲۴ ساعته، نتیجه گرفتند که ابتلای هر دو والدین به فشارخون بالا بیشتر از ابتلای یکی از والدین در میزان بالا بودن فشار خون افراد مؤثر است.

Ko و همکاران با تحلیل آماری داده‌های مربوط به ۱۵۱۳ نفر، علاوه بر تأیید نقش BMI و سایر شاخص‌های تناسب اندام در بروز فشارخون بالا، مقادیر تعیین‌کننده این شاخص‌ها را در پیش‌بینی ابتلاء به بیماری‌های دیابت، فشارخون بالا و چربی خون، مشخص نمودند [۴۱].

یافته‌های پژوهش حال حاضر، یافته‌های مطالعات قبلی را تأیید می‌کنند؛ ۱) نقش انکارناپذیر سابقه فشارخون والدین در ابتلای فرزندان به این بیماری ۲) تأثیر زیاد شاخص‌های تناسب اندام که از سبک زندگی و میزان فعالیت روزمره ناشی می‌شود. نقطه قوت پژوهش حاضر از این منظر، بررسی سابقه سایر بیماری‌ها در والدین و نقش آن‌ها در فشارخون فرزندان است که میزان این تأثیرات با رتبه‌بندی ارائه شد.

استفاده از داده‌کاوی در پیش‌بینی فشارخون بالا

Lee و همکاران در مطالعه‌ای [۴۲]، از نه روش داده‌کاوی برای پیش‌بینی بروز فشارخون بالا از جمله BMI و سابقه فشارخون در خانواده، استفاده کرده‌اند. در این مطالعه روش‌های مختلف داده‌کاوی با استفاده از معیارهای حساسیت (Sensitivity)، تشخیص (Specificity) و نرخ پیش‌بینی (Predictive Rate)، با یکدیگر مقایسه شدند که شبکه‌های عصبی بهترین عملکرد را داشتند. این مطالعه با استفاده از داده‌های مربوط به ۶۹۴ نفر انجام یافته است و در آن متوازن بودن یا نبودن تعداد اعضای کلاس‌ها، مشخص نگردیده است.

Huang و همکاران با استفاده از ۹۸۶۲ نمونه از داده‌های یک پایگاه داده اطلاعات پزشکی، به پیش‌بینی بروز فشارخون بالا به کمک داده‌های مربوط به وقوع هشت بیماری دیگر، پرداخته‌اند [۴۳]. با توجه به عدم توازن کلاس‌های مورد بررسی، روش کاهش یافته برای حذف تأثیر منفی آن به کار گرفته شده و سپس الگوریتم‌های NB و DT استفاده شده‌اند. دقت گزارش شده ۸۳/۵٪ و F-Measure در حدود ۸۲٪ است.

رمضان‌خوانی و همکاران در [۴۴]، به پیش‌بینی بروز

فشارخون بالا با استفاده از داده‌های فردی از جمله فشارخون سیستولیک و دیاستولیک، سن، جنس، اندازه دور کمر و باسن، قند و چربی خون، سابقه بیماری قلبی زودرس و دیابت در خانواده و ...، پرداخته‌اند. آن‌ها داده‌های نامتوازن را با استفاده از روش داده‌های ساختگی، متوازن‌سازی کردند. در این پژوهش، سه نوع درخت تصمیم به کار گرفته شده که با معیارهای ارزیابی G-Mean و AUC با هم مقایسه گردیدند.

با توجه به این که داده‌ها و ویژگی‌های مورد استفاده در مطالعات قبلی و با داده‌ها و ویژگی‌های مورد مطالعه پژوهش حاضر متفاوت‌اند، بنابراین مقایسه عددی نتایج حاصل از اعمال الگوریتم‌های داده‌کاوی مقدور نیست. همچنین در این پژوهش، برای حذف تأثیرات منفی داده‌های نامتوازن از سه روش مختلف و رایج استفاده کرده و نتایج حاصل از آن‌ها برای انتخاب روش برتر مورد بررسی قرار داده شد. در حالی که، در اغلب مطالعات بررسی شده، توازن داده‌ها مشخص نشده و در موارد معدودی که به این نکته توجه گردیده، تنها از یک روش برای مقابله با تأثیرات آن استفاده شده است. علاوه‌براین، بیشتر مطالعات مبتنی بر داده‌کاوی، داده‌های مورد بررسی را در ابتدا به دو دسته داده‌های آموزش و آزمایشی تقسیم کرده‌اند که در نتیجه ممکن است وابستگی به داده‌های آزمایش، اعتبار ارزیابی نتایج را تحت تأثیر قرار دهد. در حالی که در پژوهش حاضر با استفاده از روش اعتبارسنجی متقابل ده–قسمتی، قابلیت اطمینان به نتایج حاصل به مراتب بالاتر است.

مانند سایر پژوهش‌های انجام شده براساس داده‌های سلامت، مهم‌ترین چالش و محدودیت این پژوهش مرحله جمع‌آوری داده‌ها بود. از جمله این موارد می‌توان به نگرانی‌ها در رابطه با حفظ حریم شخصی و اطلاعات پزشکی افراد اشاره کرد. این مورد با حذف اطلاعات هویتی افراد از داده‌های جمع‌آوری شده، مرتفع گردید. چالش مهم دیگر، داده‌های گم شده یا ثبت شده به‌صورت اشتباه توسط مراقبین سلامت بود که جهت رفع آن‌ها اطلاعات قابل تصحیح، اصلاح و سایر موارد حذف گردیدند. چالش مهم بعدی عدم امکان راستی‌آزمایی اطلاعات سابقه بیماری والدین به دلیل جمع‌آوری آن‌ها توسط پرسشنامه، بود. همچنین، به دلیل عدم ثبت اطلاعات جنسیتی والدین، متأسفانه امکان بررسی این ویژگی مقدور نشد.

هدف از این پژوهش، اثبات تأثیر عوامل وراثتی در پیش‌بینی ابتلاء به فشارخون بالا به کمک روش‌های داده‌کاوی جهت شناسایی افراد در معرض خطر و انجام اقدامات پیشگیرانه است. نتایج به دست آمده از آزمون‌های طراحی شده، تأثیر سابقه برخی بیماری‌ها در والدین را به روشنی نمایان ساخت. دقت در نتایج ارائه شده، این واقعیت را آشکار می‌سازد که با وجود تأثیر غیرقابل‌انکار عوامل وراثتی و سابقه بیماری والدین در ابتلاء فرزندان به بیماری فشارخون، عوامل محیطی و سبک زندگی نقش تعیین کننده‌تری در پیش‌بینی ابتلا به این بیماری بازی می‌کنند؛ با این‌که فرزندان افراد مبتلا به برخی از بیماری-های مورد بررسی، بیشتر در معرض خطر فشارخون بالا هستند، اما انتخاب‌های روزمره و محیطی که در آن قرار دارند، تعیین کننده نهایی بروز فشارخون بالا در آن‌ها است. در نتیجه، این افراد با انتخاب سبک زندگی فعال، تغذیه مناسب و دوری از محیط‌های پرتنش و استرس‌زا، می‌توانند از وقوع این بیماری جلوگیری کنند.

تعارض منافع

بدین‌وسیله نویسندگان تصریح می‌نمایند که در مورد پژوهش حاضر هیچ‌گونه تضاد منافعی وجود ندارد.

References

1. Williams B, Mancia G, Spiering W, Agabiti Rosei E, Azizi M, Burnier M, et al. 2018 ESC/ESH Guidelines for the management of arterial hypertension: The Task Force for the management of arterial hypertension of the European Society of Cardiology (ESC) and the European Society of Hypertension (ESH). Eur Heart J 2018;39(33):3021-104. doi: 10.1093/eurheartj/ehy339

2. Global Health Observatory (GHO). data: Raised blood pressure [cited 2018 Jul 18]. Available from: https://www.who.int/gho/ncd/risk_factors/blood_pressure_prevalence_text/en/

3. Wang W, Lee ET, Fabsitz RR, Devereux R, Best L, Welty TK, et al. A longitudinal study of hypertension risk factors and their relation to cardiovascular disease: the Strong Heart Study. Hypertension 2006;47(3):403-9. doi: 10.1161/01.HYP.0000200710.29498.80

4. Kuschnir MC, Mendonça GA. Risk factors associated with arterial hypertension in adolescents. J Pediatr (Rio J) 2007;83(4):335-42. doi: 10.2223/JPED.1647

5. World Health Organization(WHO). A global brief on hypertension: silent killer, global public health crisis: World Health Day 2013 [cited 2021 May 2]. Available from: https://apps.who.int/iris/handle/10665/79059

6. Yoo I, Alafaireet P, Marinov M, Pena-Hernandez K, Gopidi R, Chang JF, Hua L. Data mining in healthcare and biomedicine: a survey of the literature. J Med Syst 2012;36(4):2431-48. doi: 10.1007/s10916-011-9710-5

7. Srinivas K, Rani BK, Govrdhan A. Applications of data mining techniques in healthcare and prediction of heart attacks. International Journal on Computer Science and Engineering 2010;2(2):250-5.
8. Soni J, Ansari U, Sharma D, Soni S. Predictive data mining for medical diagnosis: An overview of heart disease prediction. International Journal of Computer Applications 2011;17(8):43-8. doi: 10.5120/2237-2860
9. Reddy DL, Marriboyina V. A review on data mining from past to the future. International Journal of Computer Applications 2011; 15(7): 19-22.
10. Mayo M. What are the Industries / Fields where you applied Analytics, Data Science, Machine Learning in 2018? [cited 2021 May 2]. Available: https://www.kdnuggets.com/2019/03/poll-analytics-data-science-ml-applied-2018.html
11. Nahar J, Imam T, Tickle KS, Chen YP. Association rule mining to detect factors which contribute to heart disease in males and females. Expert Systems with Applications 2013;40(4):1086-93. https://doi.org/10.1016/j.eswa.2012.08.028
12. Shin AM, Lee IH, Lee GH, Park HJ, Park HS, Yoon KI. Diagnostic analysis of patients with essential hypertension using association rule mining. Healthc Inform Res. 2010;16(2):77-81. doi: 10.4258/hir.2010.16.2.77
13. Polat K, Güneş S. A hybrid approach to medical decision support systems: Combining feature selection, fuzzy weighted pre-processing and AIRS. Comput Methods Programs Biomed 2007;88(2):164-74. doi: 10.1016/j.cmpb.2007.07.013
14. Prathyusha TV, Prasad VG, Saiprasad GS, Nagaraj K. A study of prevalence and certain lifestyle risk factors of essential hypertension in a rural area in Telangana, India. International Journal of Medical Science and Public Health 2016;5(7):1417-23. doi: 10.5455/ijmsph.2016.13102015211
15. Xu F, Zhu J, Sun N, Wang L, Xie C, Tang Q, et al. Development and validation of prediction models for hypertension risks in rural Chinese populations. J Glob Health 2019; 9(2): 020601. doi: 10.7189/jogh.09.020601
16. Wang B, Liu Y, Sun X, Yin Z, Li H, Ren Y, et al. Prediction model and assessment of probability of incident hypertension: the Rural Chinese Cohort Study. Journal of Human Hypertension 2020. doi: 10.1038/s41371-020-0314-8
17. Li C, Sun D, Liu J, Li M, Zhang B, Liu Y, et al. A Prediction Model of Essential Hypertension Based on Genetic and Environmental Risk Factors in Northern Han Chinese. Int J Med Sci 2019;16(6):793-9. doi: 10.7150/ijms.33967
18. Chobanian AV, Bakris GL, Black HR, Cushman WC, Green LA, Izzo JL, et al. The seventh report of the joint national committee on prevention, detection, evaluation, and treatment of high blood pressure: the JNC 7 report. JAMA 2003;289(19):2560-72.
19. Kim SJ, Lee J, Nam CM, Jee SH, Park IS, Lee KJ, et al. Progression rate from new-onset pre-hypertension to hypertension in Korean adults. Circ J 2011;75(1):135-40. doi: 10.1253/circj.cj-09-0948
20. Faselis C, Doumas M, Kokkinos JP, Panagiotakos D, Kheirbek R, Sheriff HM, et al. Exercise capacity and progression from prehypertension to hypertension. Hypertension 2012;60(2):333-8. doi: 10.1161/HYPERTENSIONAHA.112.196493
21. Winegarden CR. From "prehypertension" to hypertension? Additional evidence. Ann Epidemiol 2005;15(9):720-5. doi: 10.1016/j.annepidem.2005.02.010
22. Thanathamathee P, Lursinsap C. Handling imbalanced data sets with synthetic boundary data generation using bootstrap re-sampling and AdaBoost techniques. Pattern Recognition Letters 2013;34(12):1339-47. https://doi.org/10.1016/j.patrec.2013.04.019
23. Zhao Y, Wong ZS, Tsui KL. A framework of rebalancing imbalanced healthcare data for rare events' classification: a case of look-alike sound-alike mix-up incident detection. J Healthc Eng 2018;2018:6275435. doi: 10.1155/2018/6275435
24. Chatterjee S. Deep learning unbalanced training data? Solve it like this [cited 2018 Jul 18]. Available from: https://towardsdatascience.com/deep-learning-unbalanced-trainnig-data-solve-it-like-this-6c528e9efea6
25. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. Journal of Artificial Intelligence Research 2002;16(1):321-57. doi: 10.1613/jair.953
26. Sun Y, Wong AK, Kamel MS. Classification of imbalanced data: A review. International Journal of Pattern Recognition and Artificial Intelligence 2009;23(4):687-719. https://doi.org/10.1142/S0218001409007326
27. Tharwat A. Classification assessment methods. Applied Computing and Informatics 2020. doi: 10.1016/j.aci.2018.08.003
28. Sasaki Y. The truth of the F-measure. Manchester: University of Manchester; 2007.
29. Tan PN, Kumar V, Srivastava J. Selecting the right objective measure for association analysis. Information Systems 2004;29(4):293-313. doi:10.1016/S0306-4379(03)00072-3
30. Palaniappan S, Awang R. Intelligent heart disease prediction system using data mining techniques. IEEE/ACS International Conference on Computer Systems and Applications; 2008 Mar-Apr 31-4; Doha, Qatar: IEEE; 2008. p. 108-15. doi: 10.1109/AICCSA.2008.4493524
31. Novakovic J. Using information gain attribute evaluation to classify sonar targets. 17th Telecommunications forum TELFOR; 2009 Nov 24-26; Serbia, Belgrade: 2009. p. 1351-4.
32. Chaudhari AS, Stevens KJ, Wood JP, Chakraborty AK, Gibbons EK, Fang Z, et al. Utility of deep learning super-resolution in the context of osteoarthritis MRI biomarkers. J Magn Reson Imaging 2020;51(3):768-79. doi: 10.1002/jmri.26872
33. Manios Y, Karatzi K, Moschonis G, Ioannou G, Androutsos O, Lionis C, et al. Lifestyle, anthropometric, socio-demographic and perinatal

correlates of early adolescence hypertension: The Healthy Growth Study. Nutr Metab Cardiovasc Dis 2019;29(2):159-69.
doi: 10.1016/j.numecd.2018.10.007
34. Xu R, Zhang X, Zhou Y, Wan Y, Gao X. Parental overweight and hypertension are associated with their children's blood pressure. Nutr Metab (Lond) 2019;16:35. doi: 10.1186/s12986-019-0357-4
35. Goldstein IB, Shapiro D, Guthrie D. Ambulatory blood pressure and family history of hypertension in healthy men and women. American Journal of Hypertension 2006;19(5):486-91. https://doi.org/10.1016/j.amjhyper.2005.09.025
36. Gearing Fr, Clark Eg, Perera Ga, Schweitzer MD. Hypertension among relatives of hypertensives: progress report of a family study. Am J Public Health Nations Health 1962;52(12):2058-65. doi: 10.2105/ajph.52.12.2058
37. Manuck SB, Proietti JM, Rader SJ, Polefrone JM. Parental hypertension, affect, and cardiovascular response to cognitive challenge. Psychosom Med 1985;47(2):189-200. doi: 10.1097/00006842-198503000-00011
38. Munger RG, Prineas RJ, Gomez-Marin O. Persistent elevation of blood pressure among children with a family history of hypertension: the Minneapolis Children's Blood Pressure Study. J Hypertens 1988;6(8):647-53. doi: 10.1097/00004872-198808000-00008
39. Rebbeck TR, Turner ST, Sing CF. Probability of having hypertension: effects of sex, history of hypertension in parents, and other risk factors. J Clin Epidemiol 1996;49(7):727-34. doi: 10.1016/0895-4356(96)00015-7
40. Burke V, Gracey MP, Beilin LJ, Milligan RA. Family history as a predictor of blood pressure in a longitudinal study of Australian children. J Hypertens 1998;16(3):269-76. doi: 10.1097/00004872-199816030-00003
41. Ko GT, Chan JC, Cockram CS, Woo J. Prediction of hypertension, diabetes, dyslipidaemia or albuminuria using simple anthropometric indexes in Hong Kong Chinese. Int J Obes Relat Metab Disord 1999;23(11):1136-42. doi: 10.1038/sj.ijo.0801043
42. Lee IN, Liao SC, Embrechts M. Data mining techniques applied to medical information. Medical Informatics and the Internet in Medicine 2000;25(2):81-102.
https://doi.org/10.1080/14639230050058275
43. Huang F, Wang S, Chan CC. Predicting disease by using data mining based on healthcare information system. IEEE International Conference on Granular Computing; 2012 Aug 11-13; Hangzhou, China: IEEE; 2018. p. 191-9. doi: 10.1109/GrC.2012.6468691
44. Ramezankhani A, Kabir A, Pournik O, Azizi F, Hadaegh F. Classification-based data mining for identification of risk patterns associated with hypertension in Middle Eastern population: A 12-year longitudinal study. Medicine (Baltimore) 2016;95(35):e4143.
doi: 10.1097/MD.0000000000004143

Original Article

Journal of Health and Biomedical Informatics
Medical Informatics Research Center
2021; 7(4):354-367

A Case Study of the Impact of Parental Diseases on the Probability of Hypertension Using Data Mining Techniques

Ghaderi Niri Amin[1], Farajzadeh Nacer[2]*, Baybordi Elahe[3]

• **Received:** 14 Nov 2019 • **Accepted:** 27 Apr 2020

Introduction: Hypertension is one of the most common health problems. As it has a major impact on other serious diseases such as cardiovascular diseases and strokes, and due to not having any specific symptoms, it is known as a silent killer. Therefore, proper diagnosis, control, and treatment of hypertension is crucial in health care systems and will indeed prevent the development of the other diseases affected by. Studies have shown a strong connection between some diseases of parents and the probability of having hypertension in children.

Method: In this study, using data collected from two health centers in Ardabil province and applying data mining techniques and Rapid Miner software, the impact of parental diseases on the probability of hypertension in children was investigated.

Results: The results indicated that using paternal medical history on hypertension, premature coronary artery disease, hypercholesterolemia, and hyperthyroidism has a significant role in predicting hypertension in children. Moreover, the performance of the predictive model, developed using the collected samples and data mining techniques, was enhanced by 5%.

Conclusion: Some diseases of parents, especially hypertension, increase the risk of hypertension in children. However, individual characteristics are the major determinants of this complication.

Keywords: Hypertension, Artificial Intelligence, Machine Learning, Data mining, Prediction, Paternal Medical History

• **Citation:** Ghaderi Niri A, Farajzadeh N, Baybordi E. A Case Study of the Impact of Parental Diseases on the Probability of Hypertension Using Data Mining Techniques. Journal of Health and Biomedical Informatics 2021; 7(4): 354-67. [In Persian]

1. MSc. in Information Technology, Faculty of Information Technology and Computer Engineering, Azarbaijan Shahid Madani University, Tabriz, Iran
2. Associate Professor, Faculty of Information Technology and Computer Engineering, Azarbaijan Shahid Madani University, Tabriz, Iran
3. Assistant Professor, Community Medicine Specialist, Academic Center for Education, Culture and Research (ACECR), Tabriz, Iran

***Corresponding Author:** Nacer Farajzadeh
Address: Faculty of Information Technology and Computer Engineering, Azarbaijan Shahid Madani University, 35 Km Tabriz-Maragheh Road, Tabriz, Iran
• **Tel:** 04131452044 • **Email:** n.farajzadeh@azaruniv.ac.ir