مجله تخصصی اپیدمیولوژی ایران؛ ۱۳۹٤، دوره ۱۱، شماره ٤: صفحات ٥٤-٤٦.

مقاله پژوهشی

# مقایسه روش‌های مختلف یادگیری ماشین در تشخیص پرفشاری خون در بیماران دیابتی با و بدون در نظر گرفتن هزینه‌ها

**مهدی تیموری[۱]، الهام ابراهیمی[۲]، سید محمد علوی‌نیا[۳]**

[۱] استادیار گروه علوم و فناوری شبکه، دانشکده علوم و فنون نوین، دانشگاه تهران

[۲] دانشجوی کارشناسی ارشد فناوری اطلاعات پزشکی، گروه علوم و فناوری شبکه، دانشکده علوم و فنون نوین، دانشگاه تهران

[۳] استادیار مرکز تحقیقات بیماری‌های منتقله به‌وسیله ناقلین، دانشگاه علوم پزشکی و خدمات بهداشتی درمانی خراسان شمالی

نویسنده رابط: مهدی تیموری، نشانی: تهران، خیابان کارگر شمالی، روبروی خیابان دهم، دانشکده علوم و فنون نوین دانشگاه تهران، تلفن ۰۲۱۶۱۱۱۵۷۷۵،

پست الکترونیک: mehditeimouri@ut.ac.ir

تاریخ دریافت: ۹۲/۱۱/۲۲؛ پذیرش: ۹۴/۰۸/۰۹

**مقدمه و اهداف:** بیماران دیابتی همواره در معرض ابتلا به پرفشاری خون هستند. هدف از این تحقیق طراحی یک مدل پیش‌بینی پرفشاری خون در میان افراد مبتلا به دیابت، مبتنی بر هزینه و با در نظر گرفتن توزیع این بیماری در جامعه بود، که تا حد ممکن عملکرد مناسبی داشته باشد.

**روش کار:** در این پژوهش تلاش شد تا با استفاده از روش‌های مختلف یادگیری ماشین، یک مدل پیش‌بینی مبتنی بر هزینه طراحی شود که تا حد ممکن بهترین عملکرد در پیش‌بینی افراد دیابتی در معرض خطر پرفشاری خون را داشته باشد. از میان الگوریتم‌های داده‌کاوی، از الگوریتم‌های درخت تصمیم، ماشین بردار پشتیبانی، شبکه عصبی و نیز رگرسیون لجستیک استفاده شد. برای انجام این پژوهش از داده‌های مربوط به غربالگری بیماران دیابتی برای تشخیص پرفشاری خون در استان آذربایجان شرقی استفاده شد.

**یافته‌ها:** افزایش فشار خون سیستول به میزان ۱۳۰ میلی‌متر جیوه، فرد دیابتی را بیشتر در معرض پرفشاری خون قرار می‌دهد. با رویکرد غیر هزینه‌محور، به شاخص یودن حدود ۶۸ درصد رسیدیم. زمانی که رویکرد هزینه محور به کار بسته می‌شود، بیشترین شاخص یودن (۴۷/۱۱ درصد) مربوط به شبکه عصبی است، هر چند هدف در اینجا حداقل‌سازی هزینه است که در راستای این هدف، درخت تصمیم و رگرسیون لجستیک بهترین عملکرد را دارند.

**نتیجه‌گیری:** در مسائل پیش‌بینی بیماری‌ها در جوامع، حساس به هزینه کردن روش‌ها و در نظر گرفتن توزیع واقعی بیماری در جامعه اهمیت بیش‌تری دارد تا این‌که تنها هدف، کمینه کردن تعداد خطاهای دسته‌بندی روی مجموعه داده‌ی موجود باشد.

**واژگان کلیدی:** فشار خون، دیابت، یادگیری ماشین، دسته‌بندی، روش‌های حساس به هزینه

## مقدمه

امروزه افزایش پرفشاری خون یکی از مهم‌ترین مشکلات سلامت عمومی و رو به گسترش در سطح دنیا به شمار می‌آید. این بیماری ارتباط نزدیکی با بیماری‌های قلبی- عروقی دارد و از این سبب مورد توجه است. بیش از ۷ میلیون نفر در جهان سالیانه از تأثیر مستقیم فشار خون بالا جان می‌بازند که این میزان برابر با حدود ۱۳ درصد کل مرگ‌ها است (۲،۱). مطالعه‌ها در ایران یافته‌های بسیار متفاوتی از میزان شیوع بیماری را نشان داده است، اما به طور کلی می‌توان گفت در ایران ۳۵-۲۵ درصد از افراد میان‌سال مبتلا به پرفشاری خون هستند (۳). بر اساس آمار سال ۱۳۹۱ وزارت بهداشت، درمان و آموزش پزشکی ۲۰ درصد جمعیت کشورمان دارای فشار خون بالا می‌باشند (۴).

سابقه خانوادگی، سن، جنس، مصرف نمک، مصرف الکل، مصرف دخانیات، کم تحرکی و چاقی از عوامل خطر بیماری فشار خون بالا به شما می‌روند (۸-۳،۵). این عوامل تأثیر خود را بر شاخص توده بدنی (BMI)[۱]، دیابت، چربی خون بالا (شامل کلسترول، لیپوپروتئین با تراکم بالا (HDL)، لیپوپروتئین با تراکم پایین (LDL) و تری‌گلیسرید نشان می‌دهند.

بیماران دیابتی همواره در طول عمر خود در معرض ابتلا به پرفشاری خون هستند، تا جایی که بیش از نیمی از مبتلایان به دیابت نوع دو هم‌زمان بیماری فشار خون بالا دارند. فشار خون بالا، دیابت را تشدید و دیابت، فشار خون را بدتر می‌کند و به این ترتیب یک دور باطل بین این دو بیماری برقرار می‌شود و تنها راه

[۱] Body Mass Index

کنترل این دو، درمان توأمان و هم‌زمان آن‌ها است. این بیماران به مراتب بیش‌تر از سایرین در معرض بیماری‌های قلبی- عروقی، اختلالات کلیوی و سکته‌های مغزی هستند. هر چه فشار خون بیمار دیابتی بهتر کنترل شود، خطر رخداد سکته‌های قلبی و مغزی، ضایعات چشمی و کلیوی و عروقی و اعصاب محیطی در او کم‌تر می‌شود (۹). از آنجا که این دو بیماری تهدیدی جدی برای سلامت محسوب می‌شوند و علاوه بر این برای بیمار و خانواده فرد مبتلا نیز پرهزینه است، به همین دلیل پیش‌بینی بیماران دیابتی در معرض خطر بیماری فشار خون بالا ضروری به نظر می‌رسد. بنابراین در بیش‌تر کشورها سعی دارند قبل از مبتلا شدن این افراد به این بیماری آن‌ها را شناسایی نمایند و با مراقبت‌های مناسب پزشکی و رعایت رژیم غذایی و در صورت لزوم درمان دارویی از پیشرفت بیماری جلوگیری کنند. یافتن روش‌های شناسایی افراد با خطر بالای ابتلا به بیماری پرفشاری خون و درمان به موقع آن می‌تواند کمک شایانی به سلامت این افراد و در نتیجه سلامت جامعه نماید.

مطالعه‌های متعدد در زمینه تشخیص پرفشاری خون (۱۸-۱۰) در صدد هستند که بر اساس مجموعه‌ای از آزمایش‌های پزشکی و با استفاده از تحلیل‌های آماری به مدل‌های پیش‌گویی‌کننده دست یابند. در هر کدام از این مطالعه‌ها فاکتورهای متفاوتی مورد بررسی قرار گرفته‌اند. به عنوان مثال در برخی مطالعه‌ها سن، قد، وزن و فشار خون به عنوان فاکتورهای خطر بیماری بررسی شده‌اند و در مطالعه‌های دیگر پارامترهای دیگری همچون جنس و سابقه فامیلی نیز مطرح گردیده‌اند (۱۹،۲۰). در گروهی از مطالعه‌ها بر میزان تأثیر هر یک از عوامل روی بیماری پرداخته می‌شود و با استفاده از ابزار آماری، داده‌کاوی و سایر ابزار موجود، شدت تأثیر هر یک از عوامل مشخص می‌شود. به عنوان مثال با استفاده از روش‌های مختلف داده‌کاوی یا انواع روش‌های رگرسیون به پیش‌بینی این‌که، فرد در حال حاضر در چه وضعیتی (سالم، در معرض خطر یا بیماری) قرار دارد، می‌پردازند (۱۸-۱۶).

هدف از انجام این مطالعه تشخیص خطر ابتلای افراد دیابتی به بیماری پرفشاری خون با استفاده از روش‌های مختلف داده‌کاوی است. به عبارت دیگر با روش‌های مختلف یادگیری ماشین تشخیص داده خواهد شد که یک فرد دیابتی با توجه به تعدادی از مشخصه‌های اندازه‌گیری‌شده در چه وضعیتی (سالم یا بیمار) قرار دارد.

## روش کار

جمعیت مورد مطالعه شامل ۱۹۶۱ بیمار دیابتیک غربالگری شده در شهر تبریز برای تشخیص بیماری پرفشاری خون است. فاکتورهای مورد بررسی عبارت بودند از: جنس، سن، شاخص توده‌ی بدنی، سابقه دیابت در خانواده، سابقه‌ی حاملگی، سابقه دیابت حاملگی، سابقه‌ی سقط جنین، فشار خون سیستول، فشار خون دیاستول، قند خون ناشتا، کلسترول خون، تری‌گلیسرید خون، لیپوپروتئین با تراکم بالا، کراتینین، اوره و آلبومین. به طور خلاصه مجموعه داده‌ی مورد استفاده در این پژوهش شامل دو کلاس بیمار و سالم (غیر مبتلا به فشار خون بالا) و از تعداد ۱۶ مشخصه موجود در آن، ۹ مشخصه آن عددی و بقیه اسمی بودند. در فرایند تولید عوامل خطر برای هر نفر، از نتایج آزمایش خون فرد برای سنجش پارامترهای خونی، از سؤالات شفاهی در مورد اطلاعات شخصی و سوابق بیماری و خانوادگی و از یک بار سنجش فشار خون در نخستین زمان مراجعه برای محاسبه فشار خون استفاده شد. برای تکمیل مجموعه داده و تعیین عدم ابتلا به فشار خون یا ابتلا به آن در افراد، از سه مرتبه اندازه‌گیری فشار خون به فاصله یک روز استفاده شده است. اگر میانگین سه بار فشار خون به فاصله یک روز بیشتر از ۱۴۰ روی ۹۰ بود فرد مبتلا به پرفشاری خون می‌باشد. از میان تمام مشخصه‌های استفاده شده در مدل‌های پیش‌بینی، ۶ شاخص قند خون ناشتا، کلسترول خون، تری‌گلیسرید خون، لیپوپروتئین با تراکم بالا، کراتینین و آلبومین را که طی آزمایش خون مشخص می‌شدند؛ از مجموعه داده حذف شدند. دلیل این موضوع آن است که این آزمایش‌ها هزینه‌بر بوده و امکان انجام آن برای تمام جمعیت کشور منطقی و امکان‌پذیر نیست، هم‌چنین تشخیص فشار خون از روی آزمایش‌ها مفهومی ندارد. از آنجایی مطالعه روی افراد دیابتی انجام شده است، عامل سابقه دیابت در خانواده نیز حذف شد. این مجموعه داده پس از پایش شامل تعداد ۱۲۹۹ نفر سالم (غیر مبتلا به فشار خون بالا) و ۶۶۲ نفر مبتلا به فشار خون بالا بود (جدول شماره ۱).

فرض کنید نتایج نوعی تشخیص بیماری پرفشاری خون در افراد دیابتی به صورت جدول شماره ۱ باشد، که در آن:

- مثبت واقعی (TP[1]): تشخیص‌های درست پرفشاری خون در بیماران دیابتی

[1] True Positive

- مثبت کاذب (FP[۱]): تشخیص‌های غلط پرفشاری خون در بیماران دیابتی (افراد سالمی که بیماری پرفشاری خون برای آنها تشخیص داده شده است)
- منفی واقعی (TN[۲]): تشخیص‌های درست فقدان فشار خون بالا در بیماران دیابتی با استفاده از نتایج آزمایش
- منفی کاذب (FN[۳]): تشخیص‌های غلط فقدان فشار خون بالا در بیماران دیابتی (بیماران دارای فشار خون بالا که سالم تشخیص داده شده‌اند)

می‌باشند. می‌توان نتایج فوق را به صورت کسر FNF + TPF = 1 نیز ارایه کرد که در آن TPF[۴] درصد بیماران دارای فشار خون بالا است که پرفشاری خون آن‌ها درست تشخیص داده شده است و FNF[۵] درصد بیماران دارای فشار خون بالا است که از لحاظ ابتلا به فشار خون بالا به اشتباه سالم تشخیص داده شده‌اند. علاوه بر این، معیارهای زیر نیز می‌توانند برای تحلیل نتایج استفاده شوند:

- حساسیت[۶]: تعداد بیماران پرفشار خونی درست انتخاب‌شده به نسبت کل بیمارانی که واقعاً دارای فشار خون بالا هستند (درحقیقت همان TPF)
- ویژگی[۷]: تعداد افراد سالم (غیر مبتلا به فشار خون بالا) درست انتخاب‌شده به نسبت کل افرادی که واقعاً مبتلا به فشار خون بالا نیستند
- دقت[۸]: تعداد افراد دارای بیماری فشار خون بالا که فشار خون بالای آنها تشخیص داده شده است به نسبت تمام افرادی که دارای فشار خون بالا تشخیص داده شده‌اند
- امتیاز F[۹]: معیاری است که بتواند هر دو معیار ویژگی و دقت را باهم تلفیق کرده و میزان مناسب بودن الگوریتم را با توجه به نمونه‌های مثبت و منفی (از لحاظ ابتلا به پرفشاری خون) و درست تشخیص دادن هردوی این کلاس‌ها مشخص کند. این معیار در رابطه (۱) آورده شده است.

$$\text{F-Score} = \frac{2}{1/\text{Sensitivity} + 1/\text{Precision}} \qquad (۱)$$

- معیار یودن[۱۰]: این معیار به این صورت تعریف می‌شود: J=Sensitivity+Specificity-1.

برای تشخیص افراد در معرض خطر فشار خون بالا (افرادی که از نظر روش مورد استفاده بیمار دارای فشار خون بالا تشخیص داده می‌شوند) از پنج روش مختلف یادگیری ماشین استفاده می‌شود (۲۲)

- رگرسیون لجستیک
- درخت تصمیم
- شبکه عصبی مصنوعی
- ماشین بردار پشتیبانی[۱۱]
- نزدیک‌ترین همسایه

آنچه در طراحی مدل‌های پیش‌بینی بسیار دارای اهمیت است در نظر گرفتن توزیع نمونه‌ها در جامعه یا احتمال پیش از وقوع می‌باشد. در بسیاری از تحقیقات این مسأله به طور کل در نظر گرفته نمی‌شود و شبیه‌سازی‌ها تنها بر روی مجموعه داده موجود، بدون توجه به توزیع واقعی بیماری در جامعه مورد نظر، انجام می‌گردد. در این تحقیق تمامی آزمایش‌ها با در نظر گرفتن احتمال پیش از وقوع بیماری پرفشاری خون در میان جمعیت افراد مبتلا به دیابت (نیمی از افراد دیابتی دچار مشکل فشار خون بالا هستند)، انجام شد.

در این پژوهش با دو رویکرد مختلف، به شبیه‌سازی روش‌های گوناگون یادگیری ماشین بر روی مجموعه داده پرداخته شد که عبارت‌اند از:

- رویکرد غیر هزینه‌محور: در این رویکرد هدف کاهش خطای دسته‌بندی تا حد ممکن است.
- رویکرد هزینه‌محور: هدف در این رویکرد بر مبنای هزینه دعوت افراد به پایش و یا عدم دعوت به پایش است. برای مثال ممکن است که دعوت تعداد افراد بیش‌تر به برنامه پایش (مثلاً ۱۰۰ نفر بیش‌تر) بسیار اقتصادی‌تر از وقتی باشد که تعداد کمی از افراد در معرض خطر (مثلاً ۲۰ نفر) به برنامه پایش دعوت نشوند. به زبانی ساده‌تر، در این روش سعی می‌شود

[۱] False positive
[۲] True Negative
[۳] False Negative
[۴] True Positive Fraction
[۵] False Negative Fraction
[۶] Sensitivity
[۷] Specificity
[۸] Precision
[۹] F-Score
[۱۰] Youden's Index
[۱۱] Support Vector Machine

که هزینه متوسط شامل: ۱- دعوت (اشتباه) افراد سالم به پایش؛ ۲- عدم دعوت (اشتباه) افراد مبتلا به فشار خون به پایش، کمینه شود.

به طور کلی، در مورد هزینه‌های مربوط به یک مدل پیش‌بینی ۳ دسته هزینه وجود دارد که عبارت‌اند از:

۱. هزینه اولیه: این هزینه مربوط به پایش اولیه است که در آن مشخصه‌های کلی مثل جنس، قد، وزن و سایر اقلام اطلاعاتی که از طریق یک پرسشگر جمع‌آوری می‌شوند و هزینه $C_0$ نامیده می‌شوند.

۲. هزینه جمع‌آوری هر یک از مشخصه‌های هزینه‌دار ($C_F$): هزینه‌های این مورد در هر رویکرد متفاوت است و وابسته به این است که از چه مشخصه‌هایی در هر مدل استفاده شود. همان‌طور که اشاره شد در این تحقیق از مشخصه‌های هزینه‌دار استفاده نشده است.

۳. هزینه دسته‌بندی اشتباه نمونه

- هزینه $C_{FN}$: فردی که واقعاً نیازمند دعوت به پایش بوده، اما مدل آن را در دسته عدم دعوت به پایش قرار داده است.
- هزینه $C_{FP}$: فردی که واقعاً نیازمند دعوت به پایش نبوده، اما مدل آن را در دسته دعوت به پایش قرار داده است.

با صرف‌نظر کردن از هزینه اولیه-که برای همه افراد یکسان است- می‌توان هزینه متوسط ناشی از دسته‌بندی و تشخیص اشتباه توسط ماشین را به صورت رابطه (۲) بیان کرد (۲۱):

$$C_{ave} = C_{FP} \times P(FP) + C_{FN} \times P(FN), \qquad (۲)$$

که در آن $C_{ave}$ هزینه متوسط، $C_{FP}$ هزینه مرتبط با تشخیص‌های مثبت کاذب و $C_{FN}$ هزینه مرتبط با تشخیص‌های منفی کاذب است. حال می‌توان احتمال تشخیص‌های مثبت کاذب $P(FP)$ و احتمال تشخیص‌های منفی کاذب را به صورت زیر به دست آورد:

$$P(FP) = P(HBP^-) \times P(T^+ \mid HBP^-) = P(HBP^-) \times FPF$$
$$P(FN) = P(HBP^+) \times P(T^- \mid HBP^+) = P(HBP^+) \times FNF$$

به عبارت دیگر، $P(FP)$ حاصل‌ضرب میزان عدم شیوع بیماری در جامعه یعنی $P(HBP^-)$، در کسر تشخیص‌های نادرست افراد غیر مبتلا به فشار خون بالا (برابر با یک منهای ویژگی) است. هم‌چنین، $P(FN)$ حاصل‌ضرب میزان شیوع بیماری در جامعه یعنی $P(HBP^+)$، در کسر تشخیص‌های نادرست افراد مبتلا به پرفشاری خون (برابر با یک منهای حساسیت) است. چنانچه رابطه (۲) را بازنویسی شود، به رابطه (۳) برقرار خواهد شد که میانگین هزینه مدل را نشان می‌دهد:

$$C_{ave} = C_{FP} \times P(HBP^-) \times FPF + C_{FN} \times P(HBP^+) \times FNF$$
$$= C_{FP} \times \left( P(HBP^-) \times FPF + \frac{C_{FN}}{C_{FP}} \times P(HBP^+) \times FNF \right) \qquad (۳)$$

از آنجایی که هزینه‌های واقعی ناشی از دسته‌بندی نادرست در هر دو کلاس بیمار و سالم به سادگی قابل‌محاسبه نیست، و از طرفی برخی هزینه‌ها معنوی بوده و جنبه‌ی مادی و محاسباتی ندارند، بنابراین به جای برآورد دقیق هر یک از هزینه‌ها، برآوردی از نسبت هزینه‌ها انجام شده و در ادامه با این تناسب آزمایش‌های جدید را انجام می‌دهیم. نسبت $\frac{C_{FN}}{C_{FP}}$ برابر ۱۰ در نظر گرفته می‌شود، البته این نسبت به طور دقیق مشخص نیست، اما به سبب این‌که در پروژه‌های مشابه مربوط به بیماری دیابت نوع ۲ این نسبت برابر ۸/۲ محاسبه شده است و از طرفی بر اساس اطلاعاتی که از بیماری پرفشاری خون موجود بود، عدم تشخیص صحیح این بیماری، انواع هزینه‌های چشم‌گیری برای فرد دربر دارد، بنابراین به دلیل اهمیت بالای این مسأله نسبت ۱۰ به طور برآوردی در نظر گرفته شد. با مهم شدن هزینه تشخیص اشتباه افراد بیمار واقعی، حساسیت افزایش یافته و ویژگی کاهش می‌یابد. هم‌چنین به دلیل عدم اطلاع، می‌توان از ضریب ثابت $C_{FP}$ هم صرف‌نظر کرد. در حقیقت با این کار، نتایج به صورت نسبتی از این مقدار ثابت (اما نامشخص) محاسبه می‌شوند.

برای پیاده‌سازی روش‌های مختلف یادگیری ماشین از نرم‌افزار Weka استفاده شده است.

## یافته‌ها

نتایج روش‌های مختلف یادگیری ماشین با رویکرد غیر هزینه‌محور روی مجموعه داده، در جدول شماره ۲ - مشاهده می‌شود. با توجه به نتایج به دست آمده، با معیار شاخص یودن و امتیاز F، ماشین بردار پشتیبانی و رگرسیون لجستیک عملکرد بهتری دارند، هرچند عملکرد سایر روش‌ها نیز اختلاف کمی با آن‌ها دارد. درخت تصمیم به‌دست آمده در شکل شماره ۱ نمایش

داده شده است. با توجه به این درخت می‌توان گفت:

- ویژگی‌های فشار خون سیستول و فشار خون دیاستول در اولویت بالا قرار دارند. سایر گره‌های این درخت را ویژگی‌های سابقه‌ی سقط‌جنین، سابقه‌ی دیابت خانواده، جنس، سن و شاخص توده بدنی تشکیل داده‌اند.
- در همان ابتدا تمام افرادی که فشار خون سیستول آن‌ها بیشتر از ۱۳۵ است، بیمار تشخیص داده شده‌اند. این امر حساسیت بالای درخت تصمیم در قبال بیماران دیابتی را نشان می‌دهد.
- زنان دیابتی بیش از مردان دیابتی در معرض بیماری فشار خون بالا می‌باشند.
- با افزایش سن احتمال پرفشاری خون در میان بیماران دیابتی افزایش می‌یابد.

نتایج روش‌های مختلف یادگیری ماشین با رویکرد هزینه‌محور نیز در جدول شماره ۲ خلاصه شده است که در این بین درخت تصمیم و رگرسیون لجستیک بهترین وضع را از لحاظ هزینه نسبی (بر حسب $C_{FP}$ ) دارند، هر چند شبکه‌ی عصبی با شاخص یودن ۴۷/۱۱ درصد و امتیاز F برابر ۶۵/۷۷ درصد وضعیت مناسب‌تری نسبت به سایر روش‌ها دارد. در این میان، عملکرد ماشین بردار پشتیبانی با شاخص یودن ۶/۴ بسیار بد است. واضح است که با در نظر گرفتن هزینه و تلاش برای حداقل‌سازی آن، شاخص یودن و امتیاز F در تمامی روش‌ها کم‌تر شده است، اما این مسأله در غربالگری بیماری پرفشاری خون به اندازه‌ی تحمیل هزینه تشخیص نادرست وضع بیماران تأثیر بدی ندارد. در حقیقت همان طور ملاحظه می‌شود حساسیت در حالت اعمال هزینه بسیار بالاتر از حالتی است که هزینه‌ها در نظر گرفته نشده‌اند، چرا که تشخیص وضع صحیح افراد (از نظر ابتلا به فشار خون بالا) در این روش اهمیت بسیار بالاتری دارد. درخت تصمیم به‌دست آمده در شکل شماره ۲ نمایش داده شده است. با توجه به این درخت می‌توان گفت:

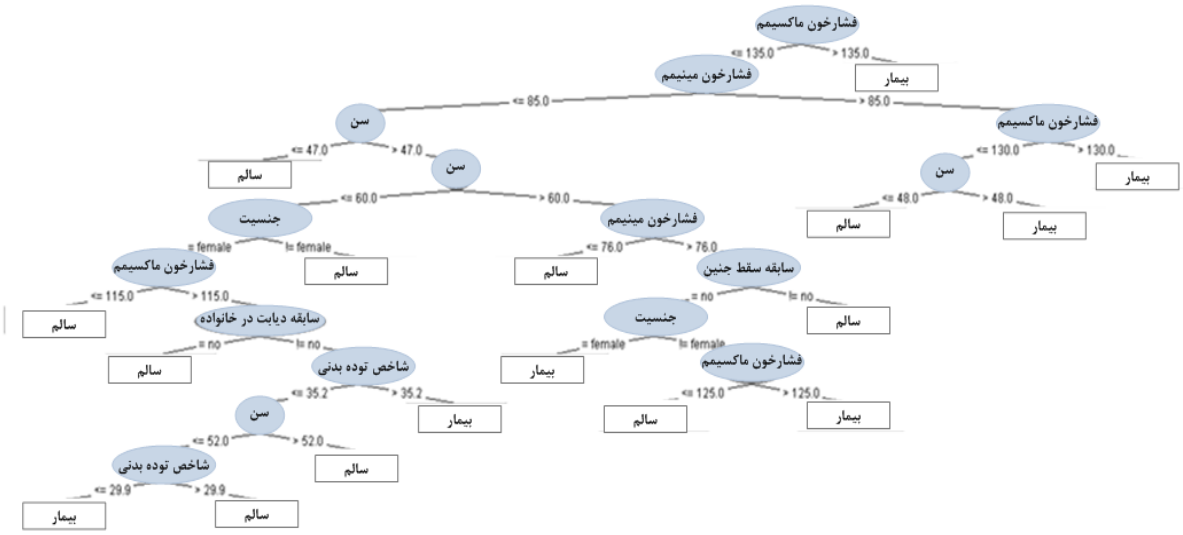
- افرادی که فشار خون سیستول بیش‌تر از ۱۳۰ میلی‌متر جیوه دارند، فرد مبتلا به فشار خون بالا تشخیص داده می‌شوند. این امر نشان از حساسیت بالای این درخت تصمیم و اهمیت بسیار بالای این ویژگی میان بیماران دیابتی می‌باشد.
- افراد مبتلا به دیابت در سنین کمتری نسبت به سایرین به فشار خون بالا مبتلا می‌شوند.
- سن تأثیر بسزایی در ابتلا به بیماری فشار خون بالا دارد.
- سابقه‌ی دیابت در خانواده فرد دیابتی تأثیر عمده‌ای در ابتلای وی به بیماری فشار خون بالا ندارد.

**جدول شماره ۱** – نتایج نوعی تشخیص بیماری فشار خون در افراد دیابتی

| | **وضع واقعی فرد** | |
|---|---|---|
| | سالم (غیر مبتلا به پرفشاری خون) | فشار خون بالا |
| نتیجه تشخیص | | |
| فشار خون بالا | FPF | TPF |
| سالم (غیر مبتلا به پرفشاری خون) | TNF | FNF |
| | ۱۲۹۹ نفر | ۶۶۲ نفر |

**جدول شماره ۲** – نتایج شبیه‌سازی روی مجموعه داده با رویکرد غیر هزینه‌محور

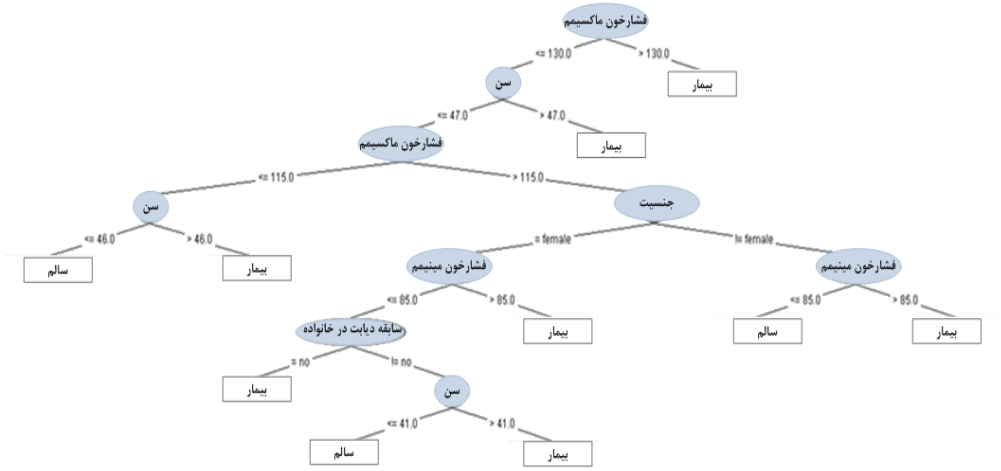
| روش مورد استفاده | حساسیت | دقت | ویژگی | امتیاز F | شاخص یودن |
|---|---|---|---|---|---|
| رگرسیون لجستیک | ۸۳/۲۳ | ۷۴/۹۷ | ۸۵/۸۴ | ۷۸/۸۸ | ۶۹/۷ |
| ماشین بردار پشتیبانی | ۸۳/۰۸ | ۷۵/۸۶ | ۸۶/۵۳ | ۷۹/۳۱ | ۶۹/۶۱ |
| شبکه عصبی | ۸۰/۵۱ | ۷۶/۳۶ | ۸۷/۳۰ | ۷۸/۳۸ | ۶۷/۸۱ |
| درخت تصمیم | ۸۲/۶۳ | ۷۵/۰۳ | ۸۵/۹۹ | ۷۸/۶۵ | ۶۸/۶۲ |




**شکل شماره ۱**- درخت تصمیم با رویکرد غیر هزینه‌محور بر روی مجموعه داده

**جدول شماره ۲** - نتایج شبیه‌سازی بر روی مجموعه داده با رویکرد هزینه‌محور

| روش مورد استفاده | حساسیت | دقت | ویژگی | امتیاز F | شاخص یودن | هزینه نسبی (بر حسب $C_{FP}$) |
|---|---|---|---|---|---|---|
| رگرسیون لجستیک | ۹۸/۳۴ | ۴۶/۵۰ | ۴۲/۳۴ | ۶۳/۱۴ | ۴۰/۶۸ | ۰/۳۷ |
| ماشین بردار پشتیبانی | ۹۹/۷۰ | ۳۵/۲۶ | ۶/۷۰ | ۵۲/۰۹ | ۶/۴ | ۰/۴۸ |
| شبکه عصبی | ۹۵/۹۲ | ۵۰/۰۴ | ۵۱/۱۹ | ۶۵/۷۷ | ۴۷/۱۱ | ۰/۴۵ |
| درخت تصمیم | ۹۷/۵۸ | ۴۸/۶۱ | ۴۷/۴۲ | ۶۴/۸۹ | ۴۵ | ۰/۳۸ |




**شکل شماره ۲**- درخت تصمیم با رویکرد هزینه‌محور بر روی مجموعه داده بدون در نظر گرفتن مشخصه‌های هزینه‌زا

## بحث

همان‌طور که ملاحظه شد، با رویکرد غیر هزینه‌محور، رگرسیون منطقی، ماشین بردار پشتیبانی و درخت تصمیم با شاخص‌های یودن به ترتیب برابر ۶۹/۷۰، ۶۹/۶۱ و ۶۸/۶۲ درصد بهترین عملکردها را دارند، اما عملکرد شبکه عصبی با شاخص یودن ۶۷/۸۱ درصد کمی بدتر است، اما زمانی که رویکرد هزینه‌محور به کار بسته می‌شود، بیش‌ترین شاخص یودن مربوط به شبکه عصبی است (۴۷/۱۱ درصد)، هر چند هدف در اینجا حداقل‌سازی هزینه است که در راستای این هدف، درخت تصمیم و رگرسیون لجستیک بهترین عملکرد را دارند.

در یک مطالعه مشابه با استفاده از مجموعه داده‌ای شامل ۸۰۹ نمونه و ۱۳ مشخصه به پیش‌بینی بیماری فشار خون بالا پرداخته شده است و در نهایت طبق گزارش نویسندگان مقاله، الگوریتم درخت تصمیم با صحت ۸۳ درصد بهترین عملکرد را داشته است (۱۶). در مطالعه‌ای دیگر با استفاده از ماشین بردار پشتیبانی به پیش‌بینی بیماری پرفشاری خون پرداخته شده که در این مطالعه نیز مشابه مطالعه قبل از ۱۳ شاخص استفاده شده که بیش از نیمی از این شاخص‌ها هزینه‌زا هستند و از آزمایش خون افراد به دست می‌آیند. در این مطالعه در نهایت با استفاده از ماشین بردار پشتیبانی با کرنل خطی به صحت ۸۵ درصد رسیده‌اند (۱۷). در مطالعه‌ای دیگر با استفاده از ۹ مشخصه شامل سن، جنس، سابقه فشار خون بالا در خانواده، کشیدن سیگار، تری‌گلیسرید خون، اوریک اسید، لیپوپروتئین چرب، شاخص توده بدنی و کلسترول خون به دسته‌بندی نمونه‌ها پرداخته شده است و در نهایت درخت تصمیم با صحت ۸۳ درصد بهترین نتیجه بوده است (۱۸). همان‌طور که از بحث فوق پیداست، در بیشتر تحقیقات همه‌ی توجه معطوف به پیش‌بینی هرچه دقیق‌تر افراد و افزایش صحت بوده است و از دو مسئله بسیار مهم هزینه جمع‌آوری داده‌ها و اعمال احتمال پیش از وقوع در پژوهش غفلت شده است، موضوعی که در پژوهش حال حاضر به آن توجهی ویژه شده است.

امید است که بتوان این طرح را در کنار طرح‌های مشابه در حال اجرا در وزارت بهداشت، درمان و آموزش پزشکی به عنوان آغاز به کار در یادگیری ماشین در حوزه‌ی پزشکی به صورت عملی در خصوص داده‌های بومی ایران در نظر گرفت.

این طرح این قابلیت را دارد که به صورت کشوری انجام شده و در اختیار عموم افراد ذی‌نفع قرار گیرد، اما متأسفانه در حال حاضر به دلیل نبود داده‌های کشوری تنها می‌توان به دسته‌بندی نمونه‌های استان آذربایجان شرقی پرداخت. در صورتی که داده‌های استاندارد و کشوری در اختیار گروه قرار گیرد می‌توان سامانه برخط ملی برای مردم، پزشکان و تصمیم‌گیرندگان ایجاد کرد.

در این پژوهش برای به دست آوردن هزینه‌های ریالی غربالگری بیماری فشار خون بالا نیاز بود تا هزینه ریالی دقیق تشخیص نادرست بیماری فشار خون بالا وجود داشته باشد، اما به سبب این‌که محاسبه این مورد تاکنون در کشورمان و روی این بیماری به طور خاص صورت نگرفته است، بنابراین، این امر غیرممکن بوده و انجام آن غیرعملی شد. از این‌رو محاسبه‌ی چنین مبالغی به طور اکید مورد نیاز بوده و انجام آن به شدت توصیه می‌شود

## نتیجه‌گیری

در مسائل پیش‌بینی بیماری‌ها در جوامع، حساس به هزینه کردن روش‌ها و در نظر گرفتن توزیع واقعی بیماری در جامعه اهمیت بیش‌تری دارد تا این‌که تنها هدف، کمینه کردن تعداد خطاهای دسته‌بندی روی مجموعه داده‌ی موجود باشد. در این مطالعه بر خلاف مطالعه‌های مرسوم، کمینه کردن هزینه تشخیص مد نظر قرار گرفت، هر چند نشان داده شد که در حالتی که هدف صرفاً کاهش خطا روی مجموعه داده باشد نیز، روش‌های تشخیص روی داده در دسترس عملکرد نسبتاً مناسبی دارند.

## تشکر و قدردانی

مقاله حاضر برگرفته از پایان‌نامه کارشناسی ارشد در رشته مهندسی فناوری اطلاعات پزشکی می‌باشد که به شماره ۲۸۲۴۲/۰۶/۰۲ در دانشکده علوم و فنون نوین دانشگاه تهران تصویب گردید. در خاتمه از مسؤولان محترم اداره دیابت وزارت بهداشت، درمان و آموزش که در گردآوری مجموعه داده‌ی مربوط به بیماران با فشار خون بالا همکاری نمودند، صمیمانه تشکر و قدردانی می‌کنیم.

## منابع


1. Kearney PM, Whelton M, Reynolds, K., Muntner P, Whelton PK, He, J. Global burden of hypertension: analysis of worldwide data. The Lancet. 2005; 365: 217-23.
2. Dickinson HO, Mason JM, Nicolson DJ, Campbell F, Beyer FR, Cook JV, . et al. Lifestyle interventions to reduce raised blood pressure: a systematic review of randomized controlled trials. J of Hypertens. 2006; 24: 215-33.
3. Rezazadehkermani M. Epidemiology and Heterogeneity of Hypertension in Iran: A Systematic Review. Arch of Iranian Med. 2008; 11: 444-52.
4. HamshahriOnline. [2015/03/01]. Available from: http://hamshahrionline.ir/details/185930.
5. Madhukumar S, Gaikwad V. An Epidemiological Study of Hypertension and Its Risk Factors in Rural Population of Bangalore Rural District. Al Ame en J Med Sci. 2012; 5: 264-70.
6. Kannana L, & , Satyamoorthyb TS. An Epidemiological Study of Hypertension In A Rural Household Community. Sri Ramachandra Journal of Medicine. 2009; 2: 9-13.
7. Fagard RH. Epidemiology of hypertension in the elderly. Health Science Journal. 2002; 11: 23-28
8. Yadav S, Boddula R, Genitta G, Bhatia V, Bansal B, Kongara S, et al. Prevalence & risk factors of pre-hypertension & hypertension in an affluent north Indian population. Indian J Med Res. 2008; 128: 712-20.
9. Group UPDS. Tight blood pressure control and risk of macrovascular and microvascular complications in type 2 diabetes: UKPDS 38. BMJ: British Medical Journal. 1998; 317: 703-13.
10. Tazi MA, Abir-Khalil S, Chaouki N, Cherqaoui S, Lahmouz F, Sraïri JE, et al. Prevalence of the Main Cardiovascular Risk Factors in Morocco. Journal of Hypertension. 2003; 21: 897-903.
11. Puavilai W, Laorugpongse D, Prompongsa S, Sutheerapatranont S, Siriwiwattnakul N, Muthapongthavorn N, et al. prevalence and some important risk facors of hypertension in Ban Paew district. Journal of the Medical Association of Thailand. 2011; 94: 1069-76.
12. Chien KL, Hsu HC, Su TC, Chang WT, Sung FC, Chen MF, et al. Prediction models for the risk of new-onset hypertension in ethnic Chinese in Taiwan. Journal of Human Hypertension. 2011; 25: 294-303.
13. Yeh DY, Cheng CH, Chen YW. A predictive model for cerebrovascular disease using data mining. Journal of Bioscience and Bioengineering. 2011; 38: 8970-7.
14. Tazi MA, Abir-Khalil S, Chaouki N, Cherqaoui S, Lahmouz F, Sraïri JE, et al. Prevalence of the Main Cardiovascular Risk factors in Morocco: Results of a National Survey, 2000. Journal of Hypertension. 2011; 21: 897-903.
15. Van Dis I. Cardiovascular risk prediction in the Netherlands. Doctoral Dissertation. University Medical Center Utrecht; 2011.
16. Kareem HR. Hypertension Prediction Using Data Mining. Scholary Research Journal for Interdisciplinary Studies. 2012; 1: 703-8.
17. Samant R, Rao S. A study on Comparative Performance of SVM Classifier Models with Kernel Functions in Prediction of Hypertension. International Journal of Computer Science and Information Technologies. 2013; 4: 818-21.
18. Ture M, Kurt, I., Turhan Kurum A, Ozdamar K. Comparing classification techniques for predicting essential hypertension. Expert Systems with Applications. 2005; 29: 583-88.
19. Tazi MA, Abir-Khalil S, Chaouki N, Cherqaoui S, Lahmouz F, Sraïri JE, et al. Risk factors for hypertension among the adult Moroccan population. Eastern Mediterranean Health Journal. 2009; 15: 827-41.
20. Omorogiuwa A, Ezenwanne EB, Osifo C, Ozor MO, Ekhator CN. Comparative study on risk factors for hypertension in a University setting in Southern Nigeria. International Journal of Biomedicaland Health Sciences, 2009; 5: 103-107.
21. Metz CE, editor Basic principles of ROC analysis. Seminars in Nuclear Medicine;1978; 4: 283-98.
22. Heydari M, Teimouri M, Heshmati M, Alavinia S. Comparison of various classification algorithms in the diagnosis of type 2 diabetes in Iran. International Journal of Diabetes in Developing Countries. 2015:1-7.

**Iranian Journal of Epidemiology 2016; 11(4): 46-54.**

**Original Article**

# Comparison of Various Machine Learning Methods in Diagnosis of Hypertension in Diabetics with/without Consideration of Costs

**Teimouri M[1], Ebrahimi E[2], Alavinia SM[3]**

1- Assistant Professor, Department of Network Science and Technology, Faculty of New Sciences and Technologies, University of Tehran, Tehran, Iran
2- MSc Student in Medical Information Technology, Department of Network Science and Technology, Faculty of New Sciences and Technologies, University of Tehran, Tehran, Iran
3- Assistant Professor, Vector-Borne Diseases Research Center, North Khorasan University of Medical Sciences, Bojnurd, Iran
**Corresponding author:** Teimouri M, mehditeimouri@ut.ac.ir

**Background and Objectives**: Diabetic patients are always at risk of hypertension. In this paper, the main goal was to design a native cost sensitive model for the diagnosis of hypertension among diabetics considering the prior probabilities.

**Methods:** In this paper, we tried to design a cost sensitive model for the diagnosis of hypertension in diabetic patients, considering the distribution of the disease in the general population. Among the data mining algorithms, Decision Tree, Artificial Neural Network, K-Nearest Neighbors, Support Vector Machine, and Logistic Regression were used. The data set belonged to Azarbayjan-e-Sharqi, Iran.

**Results:** For people with diabetes, a systolic blood pressure more than 130 mm Hg increased the risk of hypertension. In the non-cost-sensitive scenario, Youden's index was around 68%. On the other hand, in the cost-sensitive scenario, the highest Youden's index (47.11%) was for Neural Network. However, in the cost-sensitive scenario, the value of the imposed cost was important, and Decision Tree and Logistic Regression show better performances.

**Conclusion**: When diagnosing a disease, the cost of miss-classifications and also prior probabilities are the most important factors rather than only minimizing the error of classification on the data set.

**Keywords:** Hypertension, Diabetes, Machine learning, Classification, Cost sensitive models