پیش‌بینی بروز حملهٔ حاد قلبی با استفاده از رگرسیون لجستیک (مطالعهٔ موردی: بیمارستانی در ایران)

آرزو نشاطی تنها[۱]، پریا سلیمانی[۲]*

۱. کارشناس ارشد دانشکدهٔ مهندسی صنایع، دانشگاه آزاد اسلامی واحد تهران جنوب

۲. استادیار دانشکدهٔ مهندسی صنایع، دانشگاه آزاد اسلامی واحد تهران جنوب

(تاریخ دریافت ۹۴/۰۲/۰۵ ـ تاریخ دریافت روایت اصلاح‌شده ۹۴/۰۶/۰۸ ـ تاریخ تصویب ۹۴/۱۱/۲۵)

چکیده

سکتهٔ قلبی اولین عامل مرگ‌ومیر در ایران است که بیش از نیمی از بیماران آن قبل از رسیدن به بیمارستان می‌میرند. برخی از حملات قلبی ناگهانی و شدیدند، اما بیشتر آن‌ها به‌آهستگی آغاز می‌شوند و با درد یا ناراحتی خفیفی همراه هستند. تشخیص زودرس این علائم در بیماران برای نجات و درمان موفقیت‌آمیز آنان حیاتی است و این امر اهمیت، نیاز، ضرورت و سودمندی طراحی سیستم‌هایی را برای یاری‌رساندن به پزشک در تشخیص زودهنگام بروز حملات حاد قلبی بیش از پیش مشخص می‌کند. هدف اصلی این پژوهش طراحی و ساخت یک مدل پیش‌بینی بروز حملهٔ حاد قلبی در ایران، براساس فاکتورهای بالینی قابل‌گزارش توسط بیمار در خارج از بیمارستان است؛ یعنی زمانی‌که اطلاعات آزمایش‌های تشخیصی و معاینات پزشکی در دسترس نیستند. این مدل برای کاهش متوسط زمان از شروع علائم هشداردهندهٔ سکتهٔ قلبی تا آغاز درمان قابل‌استفاده است. به‌این‌منظور، داده‌های مربوط به ۷۱۱ بیمار قلبی جمع‌آوری شد و سه مدل با استفاده از رگرسیون لجستیک و یک مدل با استفاده از درخت تصمیم برای پیش‌بینی احتمال بروز حملهٔ حاد قلبی ساخته شد. بهترین مدل رگرسیون لجستیک از لحاظ عملکرد دارای آمارهٔ C، ۰/۹۵۵ و دقت ۹۴/۹ درصد بود و متغیرهای درد شدید قفسهٔ سینه، درد پشت، تعریق سرد، تنگی نفس، حالت تهوع و استفراغ به‌عنوان شاخص‌های اصلی و مؤثر در بروز حملهٔ حاد قلبی شناسایی شدند و در مدل درخت تصمیم با آمارهٔ C، ۰/۹۳۸ متغیرهای مستقل تنگی نفس، تپش قلب، ورم اندام‌ها، تعریق سرد، درد سمت چپ قفسهٔ سینه، درد شدید قفسهٔ سینه، سن و حالت تهوع قادر به پیش‌بینی تغییرات متغیر وابسته (احتمال بروز حملهٔ حاد قلبی) بودند.

واژه‌های کلیدی: بیماری عروق کرونری قلب، پیش‌بینی، حملهٔ حاد قلبی، درخت تصمیم، رگرسیون لجستیک.

مقدمه

بیماری‌های قلبی- عروقی درحال حاضر جزء سه علت اول مرگ‌ومیر و ناتوانی انسان در سراسر دنیا هستند و در آینده به اصلی‌ترین عامل مرگ‌ومیر یا ناتوانی در اغلب کشورها تبدیل می‌شوند. هر سال نزدیک به ۳۲ میلیون مورد سکتهٔ قلبی و مغزی در دنیا رخ می‌دهد که موجب مرگ بیش از ۱۷ میلیون نفر می‌شود. ۶۰ درصد از مرگ‌ومیر در سال ۲۰۰۰ در جهان به‌علت بیماری‌های غیرواگیر بوده است و برآورد می‌شود این میزان تا سال ۲۰۲۰ به ۷۳ درصد برسد. سهم بیماری‌های قلبی- عروقی از این میزان بیش از ۴۸ درصد است، به‌طوری‌که بیش از ۲۰ میلیون مورد از ۶۴ میلیون مرگ در سال ۲۰۱۵ به بیماری‌های قلبی- عروقی مربوط است و درصورتی‌که اقدامات مؤثر انجام نگیرد، موارد مرگ ناشی از بیماری‌های مزمن از سال‌های ۲۰۰۵ تا ۲۰۱۵ حدود ۱۷ درصد افزایش می‌یابد و از ۳۵ میلیون مرگ به ۴۱ میلیون مرگ می‌رسد [۱]. در سال‌های اخیر، به‌علت اقدامات پیشگیرانه و مداخلات مؤثر، روند مرگ‌های ناشی از بیماری‌های قلبی- عروقی در کشورهای توسعه‌یافته رو به کاهش بوده است، درحالی‌که این موارد در کشورهای درحال‌توسعه همچنان سیر صعودی دارد.

توسعهٔ اقتصادی و صنعتی و گسترش ارتباطات موجب ماشینی‌شدن زندگی و به‌دنبال آن سبب تغییراتی در شیوهٔ آن و افزایش بروز بیماری‌های قلبی- عروقی (بیماری‌های عروق کرونر قلب) شده است. این تغییرات شامل مصرف دخانیات، کم‌تحرکی و رژیم غذایی ناسالم است.

* نویسندهٔ مسئول: تلفن: ۰۲۱۷۷۶۳۸۸۵۳ فکس: ۰۲۱۷۷۶۳۸۴۵

Email: p_soleimani@azad.ac.ir

هدف پژوهش پیش رو بررسی این مسئله است که یک مدل ساخته‌شده با استفاده از روش‌های کلاسیک رگرسیون لجستیک و درخت تصمیم، چگونه به‌خوبی و بدون استفاده از تست‌های تشخیصی و یافته‌های معاینات فیزیکی و فقط براساس فاکتورهای تاریخچهٔ بالینی قابل‌گزارش توسط بیمار، بروز حملهٔ قلبی را پیش‌بینی می‌کند.

به‌این‌منظور، داده‌های مربوط به ۷۱۱ بیمار قلبی جمع‌آوری شد. بعد از شناسایی و حذف نقاط پرت و پرنفوذ، تعداد نهایی نمونه‌ها به ۶۶۳ بیمار با میانگین سنی ۶۳/۲۹ سال (انحراف معیار ۱۴/۳۷) رسید. براساس مرور ادبیات و مشورت با پزشکان متخصص قلب و عروق، ۲۸ مشخصهٔ مربوط به فاکتورهای بالینی قابل‌گزارش توسط بیمار بررسی شد. سپس سه مدل رگرسیون لجستیک و یک مدل درخت تصمیم برای پیش‌بینی احتمال بروز حملهٔ حاد قلبی فقط براساس مشخصه‌های بالینی قابل‌گزارش توسط بیمار ساخته شد. عوامل خطر پیش‌بینی‌شده در بروز حملات حاد قلبی در مدل‌های طراحی‌شده در این پژوهش، براساس عوامل بسیار مهم خطر از نظر پزشکان متخصص قلب و عروق در وقوع این عارضه تأیید شد که این موضوع از دستاوردهای مهم پژوهش حاضر است و میزان اطمینان به این مدل‌ها را افزایش می‌دهد.

در ایران، تاکنون پیش‌بینی احتمال بروز حملهٔ حاد قلبی با استفاده از مشخصه‌های بالینی قابل‌گزارش توسط بیمار و بدون انجام‌دادن معاینات فیزیکی و آزمایش‌های تشخیصی در خارج از بیمارستان صورت نگرفته است و محققان توجه چندانی به آن نداشته‌اند؛ بنابراین، این مدل به‌خوبی موضوع مذبور را پیش‌بینی می‌کند و در برنامه‌های کاربردی خارج از بیمارستان- که هنوز اطلاعات آزمایش‌های تشخیصی در دسترس نیستند- برای مشاورهٔ دقیق به بیمار بسیار مفید است. با توجه به میزان مرگ‌ومیر بسیار زیاد ناشی از سکتهٔ قلبی در کشور، حتی کمترین تأثیر بر کاهش متوسط زمان از شروع علائم تا درمان، به نجات جان تعداد زیادی از مردم منجر می‌شود.

درادامه، مطالعات گذشته مرور می‌شود و مدل‌های پیشنهادی ارائه می‌شود. سپس مقایسهٔ مدل‌ها و تحلیل نتایج صورت می‌گیرد. درنهایت، نتیجه بیان می‌شود.

مرور ادبیات موضوع

رگرسیون لجستیک یکی از تکنیک‌های کاربردی برای تحلیل داده‌های طبقه‌بندی‌شده است. رگرسیون لجستیک در اواخر دههٔ ۱۹۶۰ و اوایل دههٔ ۱۹۷۰ به‌عنوان بدیلی برای روش رگرسیون خطی[1] و همچنین تحلیل تابع تشخیصی[2] مطرح شد. زمانی‌که متغیر وابسته در سطح اسمی است و متغیرهای مستقل (پیش‌بین) هم ترتیبی و فاصله‌ای هستند، روش‌های رگرسیون خطی معمولی و تحلیل تشخیصی، مقدار برآوردها را کمتر از مقدار واقعی نشان می‌دهند.

رگرسیون لجستیک شبیه رگرسیون خطی است با این تفاوت که نحوهٔ محاسبهٔ ضرایب در این دو روش یکسان نیست؛ یعنی رگرسیون لجستیک به‌جای حداقل‌کردن مجذور خطاها (OLS)[3] (کاری که رگرسیون خطی انجام می‌دهد)، احتمال وقوع یک واقعه را حداکثر می‌کند و درواقع از خاصیت حداکثر درستنمایی (MLE)[4] استفاده می‌کند. همچنین، در تحلیل رگرسیون خطی برای آزمون برازش مدل و معنی‌داربودن اثر هر متغیر در مدل، به‌ترتیب از آماره‌های F و t استفاده می‌شود، درحالی‌که در رگرسیون لجستیک، از آماره‌های کای اسکوئر (χ^2) و والد (wald) استفاده می‌شود. در رگرسیون لجستیک، احتمال وقوع یک پدیده در داخل محدودهٔ صفر تا یک قرار دارد و رعایت پیش‌فرض نرمال‌بودن متغیرهای مستقل لازم نیست. یکی از اقسام رگرسیون لجستیک، مدل رگرسیون لجستیک باینری است که در این مدل دو طبقه‌بندی برای متغیر پاسخ وجود دارد. اگر بیش از دو طبقه‌بندی وجود داشته باشد، با توجه به جنس متغیر وابسته (اسمی و ترتیبی) مدل‌های رگرسیون لجستیک اسمی و ترتیبی حاصل می‌شود [۲۰].

در ابتدا، این روش به‌طورعمده درمورد کاربردهای پزشکی برای احتمال وقوع یک بیماری استفاده می‌شد، اما امروزه در تمام زمینه‌های علمی کاربرد وسیعی یافته است. هری پی سلکر و همکاران در سال ۱۹۹۵ مقایسه‌ای را بین عملکرد مدل‌های پیش‌بینی رگرسیون لجستیک، درخت تصمیم و شبکهٔ عصبی برای تشخیص بیماری ایسکمی حاد قلبی در بین بیماران بخش اورژانس با استفاده از داده‌های بالینی آن‌ها انجام دادند. از نظر آن‌ها، ممکن بود همهٔ این

روش‌ها پیش‌بینی‌های بسیار خوبی دربارهٔ پیامدهای پزشکی برای کمک به تصمیم‌گیری پزشکان و سیاست‌گذاری در این زمینه انجام دهند، ولی انتخاب بین این روش‌ها باید براساس نیازهای کاربرد خاص مورد نظر صورت می‌گرفت، نه فرض قوی‌تربودن یکی نسبت به دیگری [۲].

وانگ و همکاران در سال ۲۰۰۱ در پژوهش خود مقایسه‌ای بین عملکرد مدل‌های رگرسیون لجستیک و شبکهٔ عصبی مصنوعی برای پیش‌بینی احتمال بروز سکتهٔ قلبی در بیماران براساس فاکتورهای بالینی قابل‌گزارش توسط بیمار (۳۰ مشخصه) انجام دادند. آن‌ها نتیجه گرفتند هر دو مدل منتخب رگرسیون لجستیک و شبکهٔ عصبی مصنوعی عملکرد خوبی داشتند و تفاوت آماری معناداری بین آن‌ها وجود نداشت. برای ساخت مدل‌های رگرسیون لجستیک و شبکهٔ عصبی از روش‌های مختلفی استفاده کردند که هرکدام ده‌بار به‌صورت تصادفی تکرار شدند و مقایسهٔ عملکرد مدل‌ها را با استفاده از شاخص آمارهٔ c (سطح زیر نمودار ROC) انجام دادند. به‌منظور اعتبارسنجی، مدل‌ها را پس از ساخت با استفاده از مجموعهٔ داده‌های به‌دست‌آمده از بیماران مراجعه‌کننده به واحدهای اورژانس بیمارستان ها آزمایش کردند و نتایج پذیرفتنی بود. هدف نهایی این پژوهش استفادهٔ کاربردی از این مدل ها در خارج از بیمارستان- زمانی‌که نتایج آزمایش‌های تشخیصی بیماران در دسترس نیستند- به‌منظور کاهش مرگ‌ومیرهای ناشی از سکتهٔ قلبی بود [۳]. کندی و همکاران در سال ۱۹۹۶ در پژوهشی تشخیص زودهنگام سکتهٔ حاد قلبی را با استخراج و ارزیابی مدل‌های رگرسیون لجستیک با استفاده از داده‌های بالینی و نوار قلب در هنگام مراجعهٔ بیمار بررسی کردند [۴]. دو و همکاران در سال ۱۹۹۷ پژوهشی را با رویکرد توافقی برای تشخیص بیماری عروق کرونر قلب براساس داده‌های بالینی و تست ورزش بیماران ارائه دادند. آن‌ها با استفاده از ترکیب معادلات پیش‌بینی رگرسیون لجستیک براساس متغیرهای تست‌های ورزش و بالینی به‌دست‌آمده از جوامع آماری مختلف بیماران، آن‌ها را به سه گروه کم‌خطر، با خطر متوسط و پرخطر طبقه‌بندی کردند و توانستند ویژگی‌های تشخیصی بسیار عالی را با استفاده از داده‌های ساده و اندازه‌گیری‌های به‌دست‌آمده به‌دست آورند. این روش توافقی بهترین اجرا را با استفاده از یک ماشین‌حساب قابل‌برنامه‌ریزی یا برنامهٔ کامپیوتری برای تسهیل فرایند محاسبهٔ احتمال بیماری عروق کرونر با استفاده از سه معادله داشت [۵]. هنریک هارالدسون و همکاران در سال ۲۰۰۴ در پژوهش خود مدلی را با استفاده از شبکهٔ عصبی مصنوعی برای تشخیص خودکار سکتهٔ حاد قلبی در بیماران براساس دوازده اشتقاق نوارهای قلبی (ECGs) ارائه دادند. روش آن‌ها به‌عنوان سیستم پشتیبان تصمیم‌گیری که بینشی خوب و مناسب را برای تشخیص پزشک فراهم می‌آورد قابل استفاده است [۶]. هریسون و کندی در سال ۲۰۰۵ با استفاده از مدل‌های شبکهٔ عصبی مصنوعی پیش‌بینی سندروم حاد عروق کرونر قلب را با استفاده از داده‌های بالینی از زمان ارائه بررسی کردند. این مطالعه تأیید می‌کند شبکه‌های عصبی مصنوعی می‌توانند روش مفیدی برای توسعهٔ الگوریتم‌های تشخیصی برای بیماران دارای درد قفسهٔ سینه پیشنهاد دهد [۱۴]. بیگلریان و همکاران در سال ۲۰۱۰ با استفاده از شبکه‌های عصبی مصنوعی، عوامل قابل‌پیش‌بینی سرطان معده را در مبتلایان تعیین کردند [۷]. در سال ۲۰۱۲، آنوج با استفاده از قوانین فازی وزن‌دهی‌شده براساس سیستم پشتیبانی تصمیم‌گیری بالینی، میزان ریسک ابتلا به بیماری‌های قلبی را پیش‌بینی کرد [۸]. آتکف و همکاران در سال ۲۰۱۲ پژوهشی را برای تشخیص بیماری عروق کرونری قلب براساس پارامترهای بالینی و ژنتیکی با استفاده از شبکه‌های عصبی مصنوعی انجام دادند [۱۰]. راجسواری و همکاران در سال ۲۰۱۲ با استفاده از شبکه‌های عصبی مصنوعی به کاهش شاخصه‌ها برای تشخیص بیماری ایسکمیک قلب (IHD)[5] پرداختند. بیماری ایسکمیک قلب درواقع یک بیماری قلبی- عروقی است که با سرعت نگران‌کننده‌ای در سراسر جهان درحال افزایش است. درواقع، در این تحقیق آن‌ها به‌دنبال کاهش تعداد شاخصه‌های مورد بررسی برای شناسایی بیماری IHDبودند که کاهش شاخصه‌ها، کاهش زمان پزشکان و بیماران و کاهش هزینه‌ها را نیز در پی دارد. اهداف پژوهش آن‌ها عبارت است از: مشاهدهٔ شبکه‌های عصبی، به‌ویژه

شبکه‌های عصبی پس انتشار، مطالعه دربارهٔ بیماری IHD، شناسایی ترکیب‌های مختلف از پارامترها یا ویژگی‌هایی برای تشخیص بیماری IHD. آن‌ها توانستند شاخصه‌های ورودی را از هفده به دوازده شاخصه کاهش دهند. با وجود دوازده شاخصهٔ ورودی، دقت و صحت پیش‌بینی در طول آموزش ۸۹/۴ درصد و در آزمایش این دقت به ۸۲/۲ درصد رسید. با کاهش بیشتر شاخصه‌ها، دقت مدل کاهش پیدا کرد. درنتیجه، آن‌ها دریافتند مناسب‌ترین تعداد شاخصه برای تشخیص بیماری ایسکمیک قلبی همان دوازده شاخصه است [۹]. صفدری و همکاران در سال ۲۰۱۳ در پژوهش خود یک مدل را با استفاده از تکنیک‌های داده‌کاوی برای پیش‌بینی سکته قلبی طراحی کردند [۱۱]. سوچیترا و ماهسواری در سال ۲۰۱۴ پژوهشی را با هدف ساخت یک سیستم ویژه (کارآمد) با طبقه‌بندی‌کنندهٔ کاملاً خودکار برای تشخیص وجود بیماری ایسکمی قلب با استفاده از تکنیک‌های هوش مصنوعی انجام دادند [۱۲].

مدل پیشنهادی

داده‌ها

تحقیق حاضر یک مطالعهٔ تشخیصی است که براساس متغیرهای مستقل، احتمال بروز حملهٔ حاد قلبی را پیشگویی می‌کند. جامعهٔ آماری بیمارانی هستند که طی سال‌های ۱۳۹۲ تا ۱۳۹۳ به اورژانس بیمارستان تخصصی قلب و عروق اکباتان همدان مراجعه کرده‌اند. حجم جامعهٔ آماری مورد بررسی ۷۱۱ بیمار بود که پس از شناسایی داده‌های پرت، پروندهٔ این بیماران دوباره بررسی شد تا مشخص شود داده‌ها اشتباه وارد شده‌اند یا خیر. بعد از این بررسی بعضی مقادیر اصلاح و بقیهٔ داده‌های پرت حذف شد. تعداد نهایی نمونه‌ها به ۶۶۳ بیمار با میانگین سنی ۶۳/۲۹ سال (انحراف معیار ۱۴/۳۷) رسید. داده‌های بیماران هم شامل داده‌های کمی و هم شامل داده‌های کیفی بود.

متغیر وابسته در این مطالعه، بروز حملهٔ حاد قلبی[۶] درنظر گرفته شد. براساس مرور ادبیات و مشورت با پزشکان متخصص قلب و عروق، ۲۸ مشخصهٔ مربوط به فاکتورهای بالینی قابل‌گزارش توسط بیمار بررسی شد که عبارت بودند از: سن،[۷] جنسیت،[۸] سیگار،[۹] سکتهٔ قلبی قبلی،[۱۰] دیابت،[۱۱] فشارخون،[۱۲] چربی خون،[۱۳]: درد شدید قفسه سینه،[۱۴] درد سمت چپ قفسهٔ سینه،[۱۵] درد سمت راست قفسهٔ سینه،[۱۶] درد پشت،[۱۷] درد دست چپ،[۱۸] درد دست راست،[۱۹] تعریق سرد،[۲۰] تنگی نفس،[۲۱] حالت تهوع،[۲۲] استفراغ،[۲۳] بی‌هوشی،[۲۴] تپش قلب،[۲۵] درد فوق المعدی،[۲۶] سابقهٔ بیماری قلبی،[۲۷] ورم اندام‌ها،[۲۸] خواب‌آلودگی،[۲۹] سرگیجه،[۳۰] ضعف و بی‌حالی،[۳۱] سرفه،[۳۲] اضطراب و دل‌شوره[۳۳] و سردرد[۳۴]. تمام متغیرها به‌جز سن، دودویی (باینری) بودند (۰= عدم وجود، ۱= وجود) و برای متغیر جنسیت بیمار کدهای ۰= زن و ۱= مرد درنظر گرفته شد.

رگرسیون لجستیک

مدل رگرسیون لجستیک، مدل خطی تعمیم‌یافته‌ای محسوب می‌شود که از تابع لوجیت به‌عنوان تابع پیوند استفاده می‌کند و خطایش از توزیع چندجمله‌ای پیروی می‌کند. برای انتخاب الگوریتمی با بالاترین دقت برای پیش‌بینی بروز حملهٔ حاد قلبی، سه نوع مدل رگرسیون لجستیک با استفاده از تمام ۲۸ مشخصهٔ بالینی گزارش‌شده توسط بیماران و سه الگوریتم انتخاب متغیرها یعنی Enter، Forward و Backward و با به‌کارگیری معیار $\alpha=0.05$ به‌عنوان معیار ورود و خروج متغیرها به مدل ساخته شد [۳] و به‌ترتیب مدل ۱، مدل ۲ و مدل ۳ نامیده شدند. برای ساختن مدل‌های مذکور، مجموعهٔ دادهٔ نهایی ۶۶۳ بیمار به‌طور تصادفی به دو مجموعه و به نسبت ۸۰ درصد (آموزش) و ۲۰ درصد (ارزیابی) تقسیم شد.

مدل‌ها

مدل ۱ رگرسیون ساخته‌شده با استفاده از الگوریتم انتخاب متغیر Enter است که معادلهٔ تابع پیشنهادی آن به‌صورت معادلهٔ ۱ است:

(۱)

$$\ln\frac{\hat{P}(x)}{1-\hat{P}(x)} = -3.125+2.996\ CP+4.552\ BP+ 4.030\ Sweats+4.985\ SOB-5.668\ Nausea+ 5.329\ Vomiting$$

درمورد مدل ۱ می‌توان گفت متغیرهای مستقل CP، BP، Sweats، SOB، Nausea و Vomiting قادر به پیش‌بینی

تغییرات متغیر وابسته (احتمال بروز حملهٔ حاد قلبی) هستند و توانایی پیش‌بینی آن‌ها در سطح خطای کمتر از ۰/۰۵ معنادار است.

مدل ۲ رگرسیون ساخته‌شده با استفاده از الگوریتم انتخاب متغیر Forward Stepwise:LR است که معادلهٔ تابع پیشنهادی آن به‌صورت معادلهٔ ۲ است:

$$\ln\frac{P\hat{}(x)}{1-P\hat{}(x)} = -1.708 + 0.38\,Age + 1.736\,CP + 1.665\,BP + 2.183\,Sweats + 2.898\,SOB - 1.946\,Cough \quad (۲)$$

مدل ۲ در گام نهم متغیرهای مستقل Age، CP، BP، Sweats، SOB و Cough قادر به پیش‌بینی تغییرات متغیر وابسته (احتمال بروز حملهٔ حاد قلبی) هستند و توانایی پیش‌بینی آن‌ها در سطح خطای کمتر از ۰/۰۵ معنادار است.

مدل ۳ رگرسیون ساخته‌شده با استفاده از الگوریتم انتخاب متغیر Backward Stepwise:LR است که معادلهٔ تابع پیشنهادی آن به‌صورت معادلهٔ ۳ است:

(۳)

$$\ln\frac{P\hat{}(x)}{1-P\hat{}(x)} = -1.225 + 2.481\,HTN + 4.167\,HLP + 2.926\,CP + 3.848\,BP + 2.914\,Sweats + 4.771\,SOB - 5.892\,Nausea + 3.487\,Vomoting - 2.553\,Cough$$

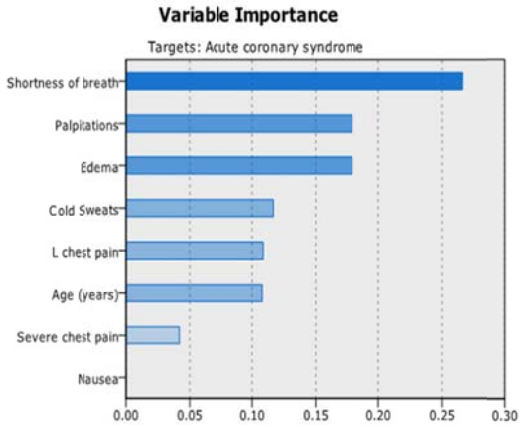
بنابراین، درمورد مدل ۳ می‌توان گفت در گام شانزدهم متغیرهای مستقل Hypertension، Hyperlipidemia، CP، BP، Sweats، SOB، Nausea، Vomiting و Cough قادر به پیش‌بینی تغییرات متغیر وابسته (احتمال بروز حملهٔ حاد قلبی) هستند و توانایی پیش‌بینی آن‌ها در سطح خطای کمتر از ۰/۰۵ معنادار است.

درخت تصمیم

درخت تصمیم یکی از مشهورترین و قدیمی‌ترین روش‌های ساخت مدل دسته‌بندی است. برای تهیهٔ یک مدل کاربردی در این تحقیق از الگوریتم درخت تصمیم CHAID استفاده شده است. در این روش، برای جداسازی و شاخه‌زدن از معیار p-value استفاده می‌شود و چنانچه اختلاف بین دو دسته به‌صورت آماری معنی‌دار باشد، عمل شاخه‌زنی در درخت انجام می‌گیرد و در غیر این‌صورت در درخت شاخه‌ای ایجاد نمی‌شود [۲]. برای پیش‌بینی بروز حملهٔ حاد قلبی در بیماران، مدل دسته‌بند درخت تصمیم با استفاده از تمام ۲۸ مشخصهٔ بالینی گزارش‌شده توسط بیماران و با به‌کارگیری معیار $\alpha = 0.05$ ساخته شد. به‌این‌منظور مجموعهٔ دادهٔ نهایی ۶۶۳ بیمار به‌طور تصادفی به دو مجموعه و به نسبت ۸۰ درصد (آموزش) و ۲۰ درصد (ارزیابی) تقسیم شد.

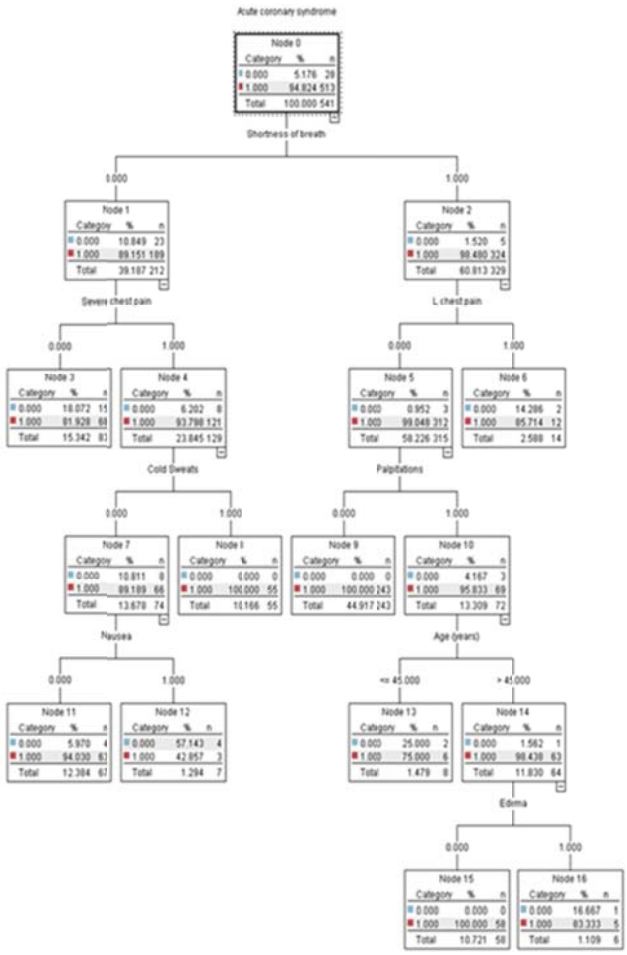
مدل درخت تصمیم

در نمودار ۱، میزان اهمیت متغیرهای مستقل مدل ساخته‌شده با استفاده از درخت تصمیم در پیش‌بینی بروز حملهٔ حاد قلبی نشان داده می‌شود. دلیل اهمیت متغیرهای مستقل در تشخیص این نکته است که مقادیر پیش‌بینی شده توسط مدل به چه میزان با تغییر مقادیر متغیر مستقل تغییر می‌کنند و تا چه میزان نسبت به آن حساس است.



نمودار ۱. میزان اهمیت متغیرهای مستقل مدل درخت تصمیم

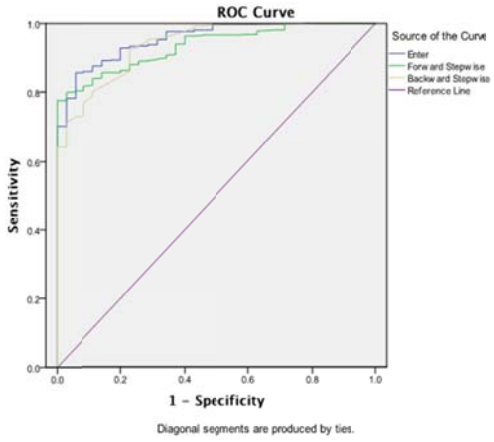
درمورد این مدل می‌توان گفت متغیرهای مستقل SOB، Palpitations، Edema، Sweats، LCP، Age، CP و Nausea قادر به پیش‌بینی تغییرات متغیر وابسته (احتمال بروز حملهٔ حاد قلبی) هستند و توانایی پیش‌بینی آن‌ها در سطح خطای کمتر از ۰/۰۵ معنادار است.



نمودار ۲. درخت تصمیم به‌دست‌آمده برای پیش‌بینی بروز حملهٔ حاد قلبی

محاسبات و تحلیل نتایج

با توجه به اینکه یکی از معیارهای سنجش تناسب مدل‌ها، نمودار ROC است، این منحنی به‌صورت مقایسه‌ای برای هر سه نوع مدل رگرسیون ساخته‌شده در نمودار ۳ و برای مدل درخت تصمیم در نمودار ۴ ارائه شده است [۱۹]. با مقایسهٔ سطح زیر منحنی ROC برای مدل‌های ۱، ۲ و ۳ رگرسیون لجستیک در جدول ۱ و سطح زیر منحنی مذکور برای مدل درخت تصمیم در جدول ۲، مشخص است مدل رگرسیون لجستیک ۱ با الگوریتم انتخاب متغیر Enter(automatic stepwise) با سطح زیر منحنی ۹۵/۵ درصد بهترین عملکرد را در مقایسه با سایر مدل‌های ساخته‌شده دارد.



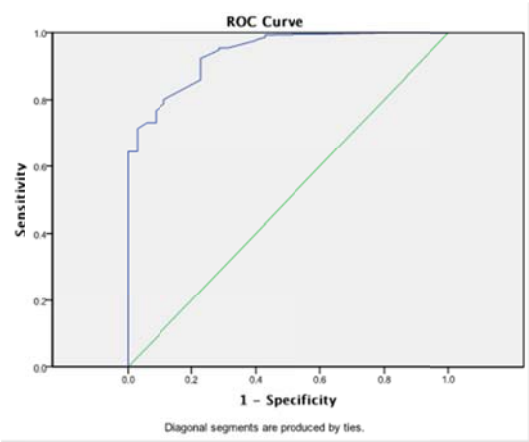
نمودار ۳. نمودار مقایسه‌ای سطح زیر منحنی ROC (معادل CStatistic) برای مدل‌های ۱، ۲ و ۳

جدول ۱. سطح زیر نمودار ROC (معادل C Statistic) برای مدل‌های ۱، ۲ و ۳

Area Under the Curve

Test Result Variable(s)	Area	Std. Error[a]	Asymptotic Sig.[b]	Asymptotic 95% Confidence Interval	
				Lower Bound	Upper Bound
Predicted probability (Enter)	.955	.012	.000	.931	.979
Predicted probability (Forward)	.934	.013	.000	.908	.960
Predicted probability (Backward)	.938	.017	.000	.905	.972

a. Under the nonparametric assumption
b. Null hypothesis: true area = 0.5



نمودار ۴. نمودار سطح زیر منحنی ROC برای مدل درخت تصمیم

جدول ۲. سطح زیر نمودار ROC (معادل C Statistic) برای مدل درخت تصمیم

Area Under the Curve

Test Result Variable(s):Predicted probability

Area	Std. Error[a]	Asymptotic Sig.[b]	Asymptotic 95% Confidence Interval	
			Lower Bound	Upper Bound
.938	.017	.000	.905	.972

a. Under the nonparametric assumption

b. Null hypothesis: true area = 0.5

خلاصهٔ نتایج مدل‌های رگرسیون ۱، ۲ و ۳ در جدول ۳ ارائه می‌شود. با مقایسهٔ این نتایج مشخص می‌شود مدل ۱ با الگوریتم انتخاب متغیر (automatic stepwise)Enter با صحت پیش‌بینی ۹۴/۹ درصد، ضریب تعیین پزودو (R2) بین ۲۳ تا ۶۷ درصد و کای دو ۲/۰۹۶، بهترین مدل پیش‌بینی بین مدل‌های رگرسیون لجستیک ارائه‌شده است.

جدول ۳. خلاصهٔ نتایج مدل‌های ۱، ۲ و ۳

مدل‌های ساخته‌شده	accuracy	Chi-square	Cox & Snell R Square	Nagelkerke R Square
مدل ۱	۹۴/۹	۲/۰۹۶	۰/۲۲۹	۰/۶۷۳
مدل ۲	۹۳/۷	۴/۲۲۵	۰/۱۵۷	۰/۴۷۶
مدل ۳	۹۳/۹	۵/۷۹۷	۰/۲۱۸	۰/۶۴۴

همان‌طورکه پیش از این هم گفته شد، مدل ۱ به‌صورت معادلهٔ ۴ به‌دست آمد:

$$\text{logit}(P) = \ln\frac{P}{1-P} = -3.125 + 2.996\ CP + 4.552\ BP + 4.030\ Sweats + 4.985\ SOB - 5.668\ Nausea + 5.329\ Vomiting \quad (۴)$$

که در آن متغیرهای درد شدید قفسهٔ سینه، درد پشت، تعریق سرد، تنگی نفس، حالت تهوع و استفراغ وارد مدل شدند.

برای مقایسهٔ تأثیر تقسیمات تصادفی مختلف بر ساخت مدل، فرایند ایجاد مجموعه‌های تصادفی از مجموعهٔ داده‌های نهایی با استفاده از ۲۸ مشخصهٔ بالینی و الگوریتم‌های انتخاب متغیر Enter، Forward و Backward، هریک ده بار تکرار شدند [۳] و میانگین مقادیر آمارهٔ C (معادل سطح زیر نمودار منحنی ROC) و دامنهٔ تغییرات آن، حاصل از ده بار تکرار تصادفی هریک از الگوریتم‌های Enter، Forward و Backward در جدول ۴ مشاهده می‌شود.

جدول ۴. تعداد دفعات حضور متغیرهای مستقل به‌عنوان متغیر مؤثر در هریک از ده تکرار تصادفی مدل‌های ۱، ۲ و ۳

نام متغیر مستقل	تعداد دفعات حضور به‌عنوان متغیر مؤثر در ده تکرار تصادفی مدل ۱	تعداد دفعات حضور به‌عنوان متغیر مؤثر در ده تکرار تصادفی مدل ۲	تعداد دفعات حضور به‌عنوان متغیر مؤثر در ده تکرار تصادفی مدل ۳
Age (years)	۳	۳	۲
Sex (1=male)			
Smokes			
Previous MI			
Diabetes			
HTN	۴	۵	۴
HLP			۱
CP	۱۰	۱۰	۱۰
LCP	۵	۲	۴
RCP			
BP	۸	۷	۵
LAP			
R AP			
Sweats	۱۰	۹	۱۰
SOB	۱۰	۱۰	۱۰
Nausea	۹	۶	۱۰
Vomiting	۵	۳	۳
Syncope			
Palpitations		۱	
Epigastric pain			
History of heart disease			
Edema			
Drowsiness			
Dizziness			
Weakness			
Cough		۳	
Anxiety			
Headache			

بایـد توجـه داشـت کـه در هـر سـه مـدل رگرسـیون لجستیک و علی‌الخصوص مدل رگرسیون لجسـتیک نهـایی (مدل ۱)، برخی از متغیرهای مستقل مؤثر در مدل، ضرایب رگرسیونی (β) منفی دارند. برای برخی فاکتورهـای بـالینی گزارش‌شده توسط بیماران از جمله حالت تهوع،[۳۵] سـرفه،[۳۶] سرگیجه،[۳۷] سردرد[۳۸] پس از بررسی و مشاوره با متخصصان مشخص شد این امر تعجـب‌آور نیسـت، زیـرا مطـابق نظـر پزشکان متخصص قلب و عروق، این مشخصه‌ها جزء عوامل خطر اختصاصی بروز حملهٔ حاد قلبی نیستند و منفی‌بـودن این ضرایب از لحاظ پزشکی منطقـی و موجـه اسـت؛ بـرای مثال، وقوع عارضهٔ درد شدید قفسهٔ سینه و سـرفه در یـک بیمار با سابقهٔ بیماری ریوی از احتمال بروز حملهٔ حاد قلبی

می‌کاهد. در مدل ۱، متغیر حالت تهوع نیز چنین وضعیتی دارد.

نتیجه‌گیری

تاکنون تحقیقات و مطالعات زیادی در زمینهٔ کاربرد روش‌ها و تحلیل‌های آماری در پزشکی انجام گرفته است، ولی در این تحقیقات به مسئلهٔ احتمال بروز حملات حاد قلبی در بیماران براساس فاکتورهای بالینی قابل‌گزارش توسط بیمار- به‌ویژه در ایران- کمتر توجه شده است. حجم جامعهٔ آماری مورد بررسی در این پژوهش ۷۱۱ بیمار بود که پس از شناسایی و حذف داده‌های پرت به تعداد ۶۶۳ بیمار رسید. داده‌های بیماران هم شامل داده‌های کمی و هم شامل داده‌های کیفی بود. براساس مرور ادبیات و مشورت با پزشکان متخصص قلب و عروق، ۲۸ مشخصهٔ مربوط به فاکتورهای بالینی قابل‌گزارش توسط بیمار بررسی شد و سه مدل رگرسیون لجستیک و یک مدل درخت تصمیم برای انجام‌دادن پیش‌بینی مذکور ساخته شد. بهترین مدل رگرسیون لجستیک از لحاظ عملکرد دارای آمارهٔ C، ۰/۹۵۵ و دقت ۹۴/۹ درصد بود و متغیرهای درد شدید قفسهٔ سینه، درد پشت، تعریق سرد، تنگی نفس، حالت تهوع و استفراغ، شاخص‌های اصلی و مؤثر در بروز حملهٔ حاد قلبی شناسایی شدند. مدل درخت تصمیم دارای آمارهٔ C، ۰/۹۳۸ بود و متغیرهای تنگی نفس، تپش قلب، ورم اندام‌ها، تعریق سرد، درد سمت چپ قفسهٔ سینه، درد شدید قفسهٔ سینه، سن و حالت تهوع قادر به پیش‌بینی تغییرات متغیر وابسته (احتمال بروز حملهٔ حاد قلبی) بودند.

در هر سه مدل رگرسیون لجستیک و به‌ویژه مدل رگرسیون لجستیک منتخب، برخی از متغیرهای مستقل مؤثر در مدل، مانند حالت تهوع، سرفه، سرگیجه و سردرد ضرایب رگرسیونی (β) منفی داشتند که با بررسی و مشاوره با متخصصان مشخص شد این امر طبیعی است، زیرا مطابق نظر پزشکان متخصص قلب و عروق، این مشخصه‌ها جزء عوامل خطر اختصاصی بروز حملهٔ حاد قلبی نیستند و منفی‌بودن این ضرایب از لحاظ پزشکی منطقی و موجه است.

عوامل خطر پیش‌بینی‌شده در بروز حملات حاد قلبی در مدل‌های طراحی‌شده در این پژوهش، براساس عوامل بسیار مهم خطر از نظر پزشکان متخصص قلب و عروق در وقوع این عارضه تأیید شد که این موضوع از دستاوردهای مهم پژوهش حاضر است و میزان اطمینان به این مدل‌ها را افزایش می‌دهد.

کاربردهای پزشکی

متأسفانه بیماران اغلب به دلایل ناآگاهی از علائم هشداردهندهٔ سکتهٔ قلبی، عوامل احساسی و انکار بیمار و مشاورهٔ ناکافی کارکنان مراقبت‌های بهداشتی، در مراجعهٔ به‌موقع به مراکز درمانی تأخیر فراوانی دارند. شواهد موجود نشان می‌دهد بیمارانی که به علائم هشداردهندهٔ حملات قلبی آگاهی و دانش بیشتری دارند، زودتر درخواست کمک می‌دهند و اگر مشاورهٔ دقیق و سریع درمورد میزان جدی‌بودن علائم خود دریافت کنند، احتمال خطر مرگ آن‌ها به میزان چشمگیری کاهش می‌یابد [۳].

مدل نهایی ارائه‌شده در این پژوهش، بدون استفاده از تست‌های تشخیصی و یافته‌های معاینات فیزیکی و فقط براساس فاکتورهای بالینی قابل‌گزارش توسط بیمار، بروز حملهٔ حاد قلبی را به‌خوبی پیش‌بینی می‌کند. این نوع مدل پیش‌بینی در برنامه‌های کاربردی خارج از بیمارستان- که هنوز اطلاعات آزمایش‌های تشخیصی در دسترس نیست- برای ارائهٔ مشاورهٔ دقیق به بیمار بسیار مفید است؛ برای مثال، ممکن است بیماران از چنین سیستمی به‌طور مستقیم در قالب نرم‌افزاری کاربردی استفاده کنند یا کارکنان اورژانس به‌عنوان پشتیبان در تصمیم‌گیری دربارهٔ وضعیت بیمار به‌صورت تلفنی آن را به‌کار بگیرند.

با توجه به تعداد بسیار زیاد مرگ‌ومیر ناشی از سکتهٔ قلبی در کشور، حتی تأثیری کوچک بر کاهش متوسط زمان از شروع علائم تا درمان، به نجات زندگی بسیاری از مردم منجر می‌شود.

مراجع

1. Samavat, T., Hojjatzadeh, A., Shams, M., Afkhami, A., Mahdavi, A., Bashti, Sh., Pouraram, H., Ghotbi, M., Rezvani, A. (1391). "Prevention and control of cardiovascular disease (for government employees)."*second edition*.
2. Selker, H.P., Griffith, J.L., Patil, S., Long, W.J., D'Agostino, R.B. (1995). "A comparison of performance of mathematicalpredictive methods for medical diagnosis: identifying acute cardiac ischemia among emergency department patients. "*J. Investig. Med*, Vol. 43, PP. 468-476.
3. Wang, S.J., Ohno-Machado, L., Fraser, H.S., Lee Kennedy, R. (2001). "Using patient-reportable clinicalhistory factors to predict myocardial infarction. "*Computers in Biology and Medicine*, Vol, 31, PP. 1-13.
4. Kennedy, R.L., Burton, A.M., Fraser, H.S.,McStay, L.N., Harrison, R.F. (1996)."Early diagnosisof acutemyocardial infarction using clinical and electrocardiographic data at presentation: derivationand evaluation of logistic regression models. "*Eur. Heart J.*, Vol. 17, PP. 1181-1191.
5. Do, D., West, J.A., Morise, A., Atwood, E.,Froelicher, V. (1997). "A consensus approach to diagnosingcoronary artery disease based on clinical and exercise test data. "*Chest*, Vol. 111, PP. 1742- 1749.
6. Haraldsson, H.,Edenbrandt, L.,Ohlsson, M. (2004)."Detecting acute myocardial infarction in the 12-lead ECG using Hermite expansions and neural networks. "*Artificial Intelligence in Medicine*,Vol. 32, PP. 127-136.
7. Biglarian, A., Hajizadeh, E., Kazemnejad, A., Zayeri, F. (2010). "Determining of prognostic factors in gastric cancer patients using artificial neural networks. "*Asian Pac J Cancer Prev*, Vol.11(2), PP. 533-536.
8. Anooj, P.K. (2012). "Clinical decision support system: Risk level prediction of heart disease usingweighted fuzzy rules. "*Journal of King Saud University – Computer and Information Sciences*,Vol. 24, PP. 27-40.
9. Rajeswari, K., Vaithiyanathan, V.,Neelakantan, T.R. (2012). " Feature Selection in Ischemic Heart DiseaseIdentification using Feed Forward Neural Networks. "*Procedia Engineering*, Vol. 41, PP. 1818 – 1823 .
10. Atkov, O.YU.,Gorokhova, S.G., Sboev, A.G., Generozov, E.V., Muraseyeva, E.V., Moroshkina, S.Y., Cherniy, N.N. (2012). "Coronary heart disease diagnosis by artificial neural networks including genetic polymorphisms and clinical parameters. "*Journal of Cardiology*, Vol. 59, PP. 190-194.
11. Safdari, R.,GhaziSaeedi, M.,Arji, G.,Gharooni, M.,Soraki, M.,Nasiri, M. (2013)."A model for predictingmyocardial infarction using data mining techniques. "*Iranian journal of medical informatics*, vol 2, issue 4.
12. Suchithra, Maheswari, P.U. (2014). "Survey on Clinical Decision Support System for DiagnosingHeartDisease. "*International Journal of Computer Science and Mobile Computing*, vol 3, Issue 2, PP. 21-28 .
13. Heden, B., Ohlin, H.,Rittner, R.,Edenbrandt, L. (1997)."Acute myocardial infarction detected in the 12-lead ECG byartificial neural networks. "*Circulation*, Vol. 96, PP. 1798-1802.
14. Harrison, R.F.,Kennedy, R.L. (2005). "Artificial neural network models for prediction of acutecoronary syndromes using clinical data from the time of presentation. "*Ann Emerg Med*, Vol. 46, PP. 431-439.
15. Austin, P.C.,Tu, J.V., Ho, J.E., Levy, D., Lee, D.S. (2013). "Using methods from the data-mining and machine-learning literature for disease classification and prediction: a case study examining classification of heart failure subtypes. "*Journal of Clinical Epidemiology*, Vol. 66, PP. 398-407.
16. Chen, C.M., Hsu, C.Y., Chiu, H.W., Rau, H.H. (2011). "Prediction of survival in patients with liver cancerusing artificial neural networks and classification and regression trees. "*IN Natural Computation (ICNC),Seventh International Conference on* Vol. 2, pp. 811-815. IEEE.
17. Vinterbo, S.,Ohno-Machado, L. (1999). "A genetic algorithm to select variables in logistic regression: example in the domain of myocardial infarction. "*Proceedings of AMIA Annual Fall Symposium*, pp. 984-988.

18. Kurt, I., Ture, M., Kurum, AT. (2008). "Comparing performances of logistic regression, classification and regression tree, and neural networks for predicting coronary artery disease. "*Expert SystAppl*, Vol. 34, PP. 366-374.

19. Zweig, M.H., Campbell, G. (1993). "Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. "*Clin. Chem.*, Vol. 39, PP. 561-577.

20. Scott, M. (2001). "Applied logistic Regression Analysis. "*Second Publication*,*Sage Publication*.

واژه‌های انگلیسی به ترتیب استفاده در متن

1. Linear Regression
2. Discriminant Analysis
3. Ordinary Least Squares
4. Maximum Likelihood Estimation
5. Ischemic Heart Disease(IHD)
6. Acute coronary syndrome
7. Age
8. Sex
9. Smokes
10. Previous MI
11. Diabetes
12. HTN
13. HLP
14. CP
15. LCP
16. RCP
17. BP
18. LAP
19. RAP
20. Sweats
21. SOB
22. Nausea
23. Vomiting
24. Syncope
25. Palpitations
26. Epigastric pain
27. History of heart disease
28. Edema
29. Drowsiness
30. Dizziness
31. Weakness
32. Cough
33. Anxiety
34. Headache
35. Nausea
36. Cough
37. Dizziness
38. Headache