مجله دانشگاه علوم پزشکی شهرکرد/ دوره ۱۳، شماره ۴/ مهر و آبان ۱۳۹۰/ ۲۷-۱۸

استفاده از روش تحلیل مولفه‌های اصلی برای افزایش صحت پیش‌بینی سندرم متابولیک در مدل‌های شبکه عصبی مصنوعی و رگرسیون لجستیک

دکتر مرتضی سدهی*[1]، دکتر یداله محرابی[2]، دکتر عباس خدابخشی[3]

[1] مرکز تحقیقات گیاهان دارویی-دانشگاه علوم پزشکی شهرکرد، شهرکرد، ایران، [2] پژوهشکده علوم غدد درون ریز متابولیک- دانشگاه علوم پزشکی شهید بهشتی، تهران، ایران، [3] گروه بهداشت محیط- دانشگاه علوم پزشکی شهرکرد، شهرکرد، ایران.

تاریخ دریافت: ۸۹/۱۰/۱٤ اصلاح نهایی: ۹۰/۱/۲۷ تاریخ پذیرش: ۹۰/۳/۱۵

چکیده:

زمینه و هدف: در فرآیند مدل‌سازی، زمانی‌که بین متغیرهای کمکی همبستگی‌های نسبتا قوی وجود داشته باشد، هم‌خطی‌چندگانه ایجاد شده و باعث کاهش کارآیی مدل می‌گردد. هدف از این مطالعه استفاده از تحلیل مولفه‌های اصلی برای تعدیل اثر هم‌خطی‌چندگانه در مدل‌های رگرسیون لجستیک و شبکه عصبی مصنوعی و بررسی تاثیر آن بر صحت و دقت پیش‌بینی سندرم متابولیک بود.

روش بررسی: در این مطالعه توصیفی – تحلیلی تعداد ۳٤۷ نفر از افراد شرکت کننده در مطالعه آینده نگر قند و لیپید تهران که در فاز اول مطالعه بر اساس تعریف پانل درمان بالغین (ATPIII) مبتلا به سندرم متابولیک نبودند انتخاب شدند. ابتدا مدل‌های رگرسیون لجستیک و شبکه عصبی مصنوعی با استفاده از متغیرهای کمکی اولیه و سپس با استفاده از مولفه‌های اصلی به داده‌ها برازش گردید و پیش‌بینی بر اساس این مدل‌ها انجام شد. از تحلیل راک و آماره کاپا برای مقایسه قدرت پیش‌بینی مدل‌ها استفاده گردید.

یافته‌ها: برای مدل‌های رگرسیون لجستیک، رگرسیون لجستیک با مولفه‌های اصلی، شبکه عصبی مصنوعی و شبکه عصبی مصنوعی با مولفه‌های اصلی به‌ترتیب مساحت زیر منحنی راک ۰/۷٤۹، ۰/۷۹۰، ۰/۸۹۰ و ۰/۹۲۷ به‌دست آمد، میزان حساسیت مدل‌ها ۰/٤۸۳، ۰/٤۳۵، ۰/۸۳٦ و ۰/۹۱۹، ویژگی آن‌ها ۰/۸۵۷، ۰/۹۱۹، ۰/۸۹۲ و ۰/۹٦٤ و اندازه آماره کاپا برای مدل‌ها ۰/۳۲۲، ۰/۳۸٦، ۰/۷۱۲ و ۰/۸۸٦ به‌دست آمد.

نتیجه‌گیری: تحقیق نشان داد که صحت پیش‌بینی مدل‌های بر اساس مولفه‌های اصلی از مدل‌های مبتنی بر متغیرهای کمکی اولیه بیشتر بوده و بنابراین در هنگام وجود هم‌خطی‌چندگانه، مدل‌های مبتنی بر مولفه‌های اصلی برای پیش‌بینی سندرم متابولیک کاراتر هستند.

واژه‌های کلیدی: تحلیل مولفه‌های اصلی، سندرم متابولیک، پیش‌بینی، شبکه عصبی مصنوعی، رگرسیون لجستیک، هم‌خطی‌چندگانه.

مقدمه:

سندرم متابولیک به مجموعه‌ای از اختلالات متابولیک اطلاق می‌شود که وقوع همزمان آن‌ها در هر شخص بیشتر از خطر وقوع احتمالی هر یک به تنهایی است. مطالعات نشان می‌دهد مرگ و میر ناشی از بیماری‌های قلبی عروقی به طور مشخصی در مبتلایان به سندرم متابولیک بیشتر است. این سندرم به علت ارتباط با دیابت و بیماری‌های قلبی عروقی و نیز به خاطر شیوع بالا در بین جوامع، توجه بسیاری از محققان را به خود جلب نموده است. سازمان جهانی بهداشت (WHO)، فدراسیون بین‌المللی دیابت (IDF) و پانل درمانی بالغین (ATPIII) معیارهایی را برای سندرم متابولیک ارائه کرده‌اند (۱). در تعریف ATPIII از شاخص‌های دور کمر، فشارخون، تری‌گلیسیرید، کلسترول، لیپوپروتیین با دانسیته بالا (HDL) و قند خون ناشتا استفاده شده است. تشخیص سندرم متابولیک بر اساس چاقی، افزایش فشار خون، افزایش سطح تری‌گلیسیرید، پایین بودن کلسترول HDL و افزایش قند خون ناشتا انجام می‌شود. در تعریف ATPIII وجود سه معیار از پنج معیار فوق

*نویسنده مسئول: شهرکرد- رحمتیه- دانشکده بهداشت- گروه آمار زیستی و اپیدمیولوژی-تلفن: ۰۳۸۱-۳۳۳٤٦۷۸، E-mail:sedehi56@gmail.com

الزامی است (۱). این سندرم ۲۳ درصد جهان غرب را مبتلا کرده است (۲). در آمریکا، شیوع خام و تطبیق داده شده بر اساس سن به ترتیب ۲۱/۸ و ۲۳/۷ درصد گزارش شده است (۲). در بین بالغین کره جنوبی شیوع تطبیق داده شده بر اساس سن ۱۴/۲ درصد در مردان و ۱۷/۷ درصد در زنان گزارش شده است (۲). در ایران در مطالعه عزیزی و همکاران شیوع سندرم متابولیک در جمعیت مورد مطالعه ۳۰/۱ درصد و شیوع استاندارد شده بر اساس سن ۳۳/۷ درصد گزارش شده است (۳). فخرزاده و همکاران در مطالعه خود شیوع سندرم متابولیک در جمعیت مورد بررسی را ۲۹/۹ درصد و شیوع تطبیق داده شده با سن را ۲۷/۵ درصد گزارش کرده‌اند (۲). صدر بافقی و همکاران در مطالعه خود ضمن بررسی شیوع سندرم متابولیک به بررسی عوامل موثر در آن نیز پرداخته‌اند. در این مطالعه شیوع سندرم متابولیک در جمعیت مورد بررسی ۳۲/۱ درصد گزارش شده است (۴).

با توجه به مرگ و میر ناشی از این بیماری و بار عظیم اقتصادی ناشی از آن ارائه مدل‌های آماری و ریاضی که با دقت قابل قبول بتواند جهت مدل‌سازی و پیش‌بینی ابتلا به سندرم متابولیک مورد استفاده قرار گیرد، از اهمیت ویژه‌ای برخوردار است (۲). در مطالعات انجام شده تاکنون از روش‌های متعددی برای مدل‌سازی و پیش‌بینی ابتلا به سندرم متابولیک در افراد استفاده شده است. Hadaegh و همکاران در مطالعه خود با استفاده از مدل رگرسیون لجستیک به بررسی عوامل مختلف مرتبط با سندرم متابولیک در بزرگسالان با وزن طبیعی پرداخته‌اند (۵). دانش‌پور و همکاران نیز در مطالعه خود عوامل مرتبط با سندرم متابولیک را با استفاده از تحلیل عاملی (Factor Analysis) مورد بررسی قرار داده‌اند (۶). سدهی و همکاران در مطالعه خود از مدل شبکه عصبی مصنوعی (Artificial Neural Network) برای پیش‌بینی سندرم متابولیک استفاده نموده و مدل خود را با مدل‌های رگرسیون لجستیک و تحلیل ممیزی (Discriminate Analysis) مقایسه نمودند (۷).

مدل شبکه عصبی مصنوعی نسبت به مدل‌های کلاسیک آماری محدودیت‌های کمتری دارد و در بسیاری موارد عملکرد آن نسبت به مدل‌های آماری دقیق‌تر است (۷).

یکی از مشکلاتی که در هر دو روش کلاسیک و شبکه عصبی مصنوعی موجب ناپایداری مدل و کاهش صحت پیش‌بینی مدل می‌گردد، هم‌خطی-چندگانه (Multicolinearity) است. این حالت زمانی رخ می‌دهد که متغیرهای کمکی با یکدیگر همبستگی نسبتاً قوی داشته باشند. وجود هم‌خطی‌چندگانه در مدل‌های رگرسیونی باعث افزایش خطای استاندارد برآورد ضرایب رگرسیونی شده و ممکن است منجر به پیش‌بینی‌هایی خارج از دامنه مورد انتظار شود (۸). در مدل شبکه عصبی مصنوعی، وجود هم‌خطی‌چندگانه باعث پایین آمدن دقت پیش‌بینی مدل شبکه عصبی شده و برآورد وزن‌ها در لایه‌های مختلف شبکه عصبی در هر بار تکرار الگوریتم آموزش، با تغییرات زیادی مواجه می‌شود که این مسئله ممکن است باعث عدم همگرایی (Converge) شبکه عصبی مصنوعی گردد. این مشکل به خصوص زمانی که تعداد متغیرهای کمکی (ورودی) زیاد باشد، بسیار محتمل است (۹).

یکی از روش‌های حل این مشکل، استفاده از روش تحلیل مولفه‌های اصلی (Principle Component Analysis) برای پردازش متغیرهای ورودی، حذف همبستگی بین متغیرهای ورودی و کاهش تعداد آن‌ها جهت بهبود نتایج مدل‌های رگرسیونی و شبکه عصبی مصنوعی می‌باشد (۸).

رگرسیون لجستیک یکی از ابزارهای آماری است که به منظور مدل‌سازی و تحلیل داده‌ها از آن استفاده می‌شود.

از مزایای استفاده از مدل رگرسیون لجستیک علاوه بر مدل‌سازی مشاهدات می‌توان به امکان پیش‌بینی احتمال تعلق هر فرد به هر یک از سطوح متغیر وابسته و همچنین امکان محاسبه مستقیم نسبت شانس

(Odds Ratio(OR)) با استفاده از ضرایب مدل را نام برد (۱۰).

شبکه‌های عصبی مصنوعی برای مسائل تشخیص و طبقه‌بندی و پیش‌بینی که در آن‌ها روابط معمولاً به صورت خطی و یا غیر خطی هستند، مورد استفاده قرار می‌گیرند. فلسفه اصلی محاسبات شبکه عصبی مصنوعی، مدل‌سازی عمده ویژگی‌های مغز و نحوه عملکرد آن در جهت ساخت مدل‌هایی است که بتواند حتی الامکان ویژگی‌های مفید مغز را از خود بروز دهد (۱).

شبکه‌های عصبی از جنبه‌های توپولوژی، ساختاری و روش‌های یادگیری به انواع مختلفی تقسیم می‌شوند و هر یک در کاربردهای خاصی عملکرد مناسبی از خود نشان می‌دهند. شبکه عصبی چند لایه پرسپترون (MLP= Multi-Layer Perceptron) با روش یادگیری پس‌انتشار (Back-Propagation) یکی از متداول‌ترین شبکه‌های کاربردی است. در مباحث نظری اثبات شده که شبکه MLP در صورت انتخاب صحیح ساختار مناسب داخلی، قادر است هر گونه سیستم غیر خطی را مدل کرده و شبیه‌سازی نماید (۱۱).

تفسیر اپیدمیولوژیک شبکه‌های عصبی مصنوعی در مقایسه با مدل‌های آماری مرسوم پیچیده‌تر است، با این وجود، این‌گونه مدل‌ها در زمینه‌های گوناگون علوم پزشکی از جمله اپیدمیولوژی (۱۲)، پیش‌بینی سرطان پروستات (۱۳)، پیش‌بینی حاملگی ناخواسته (۱۴) و پیش‌بینی مرگ پس از جراحی قلب باز (۱۵) به کار گرفته شده‌اند.

تحلیل مولفه‌های اصلی توسط پیرسن در سال ۱۹۰۱ به‌وجود آمد و سپس در سال ۱۹۳۳ توسط هتلینگ گسترش یافت. تحلیل مولفه‌های اصلی از روش‌های آماری چند متغیره است که می‌توان از آن برای کاهش تعداد متغیرها و تفسیر بهتر اطلاعات استفاده کرد. با اعمال این روش، متغیرهای ورودی اولیه به مولفه‌های جدید بدون همبستگی تبدیل می‌شوند، به طوری که مولفه‌های ایجاد شده، ترکیبی خطی از متغیرهای ورودی‌اند (۱۰). تحلیل مولفه‌های اصلی یکی از کاربردی‌ترین روش‌های کاهش ابعاد داده‌ها در مدل‌های چند متغیره است. مولفه‌های اصلی با توجه به خصوصیاتی که دارند برای مقابله با مشکل هم خطی‌چندگانه و کاهش ابعاد داده‌ها مورد استفاده قرار می‌گیرند. در این روش با استفاده از ماتریس مقادیر ویژه (Eigen Values)، مولفه‌های اصلی به صورت ترکیبی خطی از متغیرهای اولیه و مستقل از یکدیگر ساخته شده و در تحلیل داده‌ها به جای متغیرهای اصلی (اولیه) مورد استفاده قرار می‌گیرند (۸).

این تحقیق با هدف بررسی تاثیر استفاده از مولفه‌های اصلی بر افزایش صحت پیش‌بینی ابتلا به سندرم متابولیک در مدل‌های شبکه عصبی مصنوعی و رگرسیون لجستیک مورد بررسی انجام شد.

روش بررسی:

در این مطالعه توصیفی- تحلیلی برای بررسی اثر روش تحلیل مولفه‌های اصلی روی صحت پیش‌بینی مدل رگرسیون لجستیک و مدل شبکه عصبی مصنوعی، از داده‌های مربوط به مطالعه قند و لیپید تهران استفاده گردید. مطالعه قند و لیپید تهران یک مطالعه آینده‌نگر بود که در جمعیت نماینده‌ای از ساکنان منطقه ۱۳ تهران، با هدف تخمین میزان شیوع اختلال‌های متابولیک و شناسایی عوامل خطرساز بیماری‌های قلبی عروقی انجام شد. در آن مطالعه ۱۵۰۰۵ نفر از جامعه شهری تهران به‌صورت تصادفی انتخاب شدند که از بین آن‌ها ۱۰۳۶۸ فرد بالای ۲۰ سال، در سال ۱۳۷۹ در فاز اول مطالعه مورد بررسی قرار گرفتند. در تحقیق حاضر، تعداد ۳۴۷ نفر از افرادی که در فاز اول مطالعه (۱۳۸۱-۱۳۷۹) به سندرم متابولیک مبتلا نبوده، از بخش کوهورت مطالعه قند و لیپید تهران انتخاب و پس از حدود ۳ سال پیگیری در فاز دوم مطالعه دوباره مورد بررسی قرار گرفتند که تعداد ۱۲۲ نفر آن‌ها بر اساس معیار ATPIII به سندرم متابولیک مبتلا شده بودند. جزییات مربوط به این تحقیق در مطالعه عزیزی و همکاران آمده است (۳). متغیرهای کمکی مورد بررسی

در این مطالعه عبارتند از سن، جنس، وضعیت تاهل، سابقه بیماری‌های قلبی عروقی، نمایه‌ی توده بدن (BMI)، لیپوپروتئین با دانسیته پایین (LDL)، لیپوپروتئین با دانسیته بالا (HDL)، کلسترول تام، تری‌گلیسیرید، قند خون ناشتا، قند خون دو ساعته، مصرف سیگار (هرگز، گاهی، همیشه)، فشار خون سیستولیک، فشار خون دیاستولیک و دور کمر که اندازه‌گیری همه آن‌ها در فاز ۱ مطالعه قند و لیپید تهران انجام شده بود. متغیر وابسته (پاسخ) نیز در مطالعه ابتلا به سندرم متابولیک در فاز دوم مطالعه است. برای بررسی بروز همخطی‌چندگانه در این مشاهدات از ماتریس ضرایب همبستگی و آزمون بارتلت و برای بررسی کفایت نمونه از معیار Kaiser Mcyer Olkin (KMO) استفاده گردید. این شاخص اندازه همبستگی جزیی بین متغیرها را بررسی می‌کند و مشخص می‌سازد که آیا واریانس متغیرهای تحقیق، تحت تاثیر واریانس مشترک برخی عامل‌های پنهانی و اساسی است یا خیر. این شاخص بین صفر تا یک قرار دارد و مقادیر بزرگتر از ۰/۶ آن نشان دهنده کفایت نمونه است (۱۰).

با توجه به اینکه در این تحقیق لازم بود توان مدل‌های مختلف مورد استفاده برای پیش‌بینی سندرم متابولیک با یکدیگر مقایسه گردند، از معیار سطح زیر منحنی مشخصه عملکرد (ROC) برای مقایسه مدل‌های مختلف استفاده گردید.

داده‌ها به دو قسمت تقسیم شد. از نیمی از داده‌ها (۱۷۳ نفر) برای برازش مدل‌ها و از نیم دیگر (۱۷۴ نفر) برای بررسی دقت پیش‌بینی و اعتبارسنجی مدل‌ها استفاده گردید. ابتدا بدون در نظر گرفتن هم خطی‌چندگانه بین متغیرهای کمکی، مدل رگرسیون لجستیک و مدل شبکه عصبی مصنوعی به داده‌ها برازش گردید. برای طراحی مدل شبکه عصبی مصنوعی، از یک شبکه پرسپترون دو لایه (۱۵:۱۰:۱) با الگوریتم پس‌انتشار خطا و نرخ یادگیری ۰/۱، تابع انتقال تانژانت هایپربولیک و حداکثر خطای ۰/۰۰۱ جهت برازش مدل استفاده گردید. از آنجا که در روش تحلیل مولفه‌های اصلی، از ماتریس همبستگی برای مشخص کردن مولفه‌های اصلی استفاده می‌گردد، تنها متغیرهای کمی می‌توانند در تحلیل مولفه‌های اصلی مورد استفاده قرار گیرند. بنابراین در مرحله بعد با کنار گذاشتن متغیرهای کمکی کیفی (جنس، تاهل، سابقه بیماری قلبی عروقی و مصرف سیگار)، با ترکیب ۱۱ متغیر کمی باقیمانده، با روش تحلیل مولفه‌های اصلی، ۵ مولفه اصلی بدست آمد که از آن‌ها به همراه متغیرهای کیفی برای برازش مدل‌های جدید رگرسیون لجستیک و شبکه عصبی مصنوعی با ۹ متغیر کمکی (پیشگو) استفاده گردید. برای تحلیل مولفه‌های اصلی و برازش مدل رگرسیون لجستیک از نرم افزار SPSS18 و برای برازش مدل شبکه عصبی مصنوعی از نرم افزار MATLAB7.6 استفاده گردید. دقت پیش‌بینی مدل‌ها با استفاده از سطح زیر منحنی مشخصه عملکرد، آماره کاپا، مقادیر حساسیت و ویژگی و همچنین نسبت درست نمایی مثبت و منفی مقایسه گردید.

یافته‌ها

مقدار نتایج برای آزمون بارتلت (P=۰/۰۰۱۳) و KMO (KMO=۰/۷۳۵) نشان دهنده مناسبت داده‌ها برای انجام تحلیل مولفه‌های اصلی است.

در مجموع ۷۵/۸۵۱ درصد از واریانس کل را ۵ مولفه اصلی مورد استفاده در تحلیل پوشش داد (جدول شماره ۱).

جدول شماره ۱: مقدار واریانس تعریف شده توسط مولفه‌های اصلی

مقدار واریانس / مولفه اصلی	درصد واریانس تعریف شده	درصد تجمعی واریانس
PC1*	۲٤/٤۸۷	۲٤/٤۸۷
PC2	۱٦/۷۳۱	٤۱/۲۱۸
PC3	۱۳/۷۰٤	٥٤/۹۲۲
PC4	۱۰/۷٤٤	٦٥/٦٦٦
PC5	۱۰/۱۸٥	۷٥/۸٥۱

** مولفه های اصلی (Principal Component)*

نتایج پیش‌بینی وضعیت افراد مورد بررسی توسط مدل‌های مختلف در مقایسه با وضعیت واقعی افراد فقط مربوط به ۱۷۴ نفری است که در مجموعه اعتبارسنجی قرار دارند. بر اساس این یافته‌ها مدل رگرسیون لجستیک وضعیت ۷۲/۴ درصد، مدل رگرسیون لجستیک با مولفه اصلی وضعیت ۷۴/۷ درصد، مدل شبکه عصبی مصنوعی وضعیت ۸۷/۴ درصد و در نهایت مدل شبکه عصبی مصنوعی با مولفه اصلی وضعیت ۹۲/۸ درصد افراد سالم یا مبتلا به سندرم متابولیک را به‌درستی پیش‌بینی کردند (جدول شماره ۲).

در مدل شبکه عصبی مصنوعی، استفاده از مولفه‌های اصلی به‌جای متغیرهای اولیه باعث افزایش حساسیت، ویژگی و نسبت درستنمایی مثبت و کاهش نسبت درستنمایی منفی شده در حالیکه در مدل رگرسیون لجستیک استفاده از مولفه‌های اصلی به‌جای متغیرهای اولیه باعث کاهش حساسیت و افزایش ویژگی، نسبت درستنمایی مثبت و همچنین نسبت درستنمایی منفی شده است (جدول شماره ۳).

جدول شماره ۲: طبقه‌بندی افراد مورد بررسی بر اساس پیش‌بینی مدل‌های مختلف

پیش بینی مدل / وضعیت واقعی		سالم		سندرم متابولیک	
		تعداد	درصد	تعداد	درصد
رگرسیون لجستیک	سالم	۹٦	٥٥/۲	۳۲	۱۸/٤
	سندرم متابولیک	۱٦	۹/۲	۳۰	۱۷/۲
رگرسیون لجستیک با مولفه اصلی	سالم	۱۰۳	٥۹/۲	۳٥	۲۰/۱
	سندرم متابولیک	۹	٥/۲	۲۷	۱٥/٥
شبکه عصبی مصنوعی	سالم	۱۰۰	٥۷/٦	۱۰	٥/۷
	سندرم متابولیک	۱۲	٦/۸	٥۲	۲۹/۸
شبکه عصبی مصنوعی با مولفه اصلی	سالم	۱۰۸	٦۲	٥	۲/۹
	سندرم متابولیک	٤	۲/۳	٥۷	۳۲/۸

نتایج پیش‌بینی وضعیت افراد مورد بررسی توسط مدل‌های مختلف در مقایسه با وضعیت واقعی افراد فقط مربوط به ۱۷٤ نفری است که در مجموعه اعتبارسنجی قرار دارند.

جدول شماره ۳: حساسیت، ویژگی، آماره کاپا و سطح زیر منحنی مشخصه عملکرد (ROC) برای مدل‌های مختلف

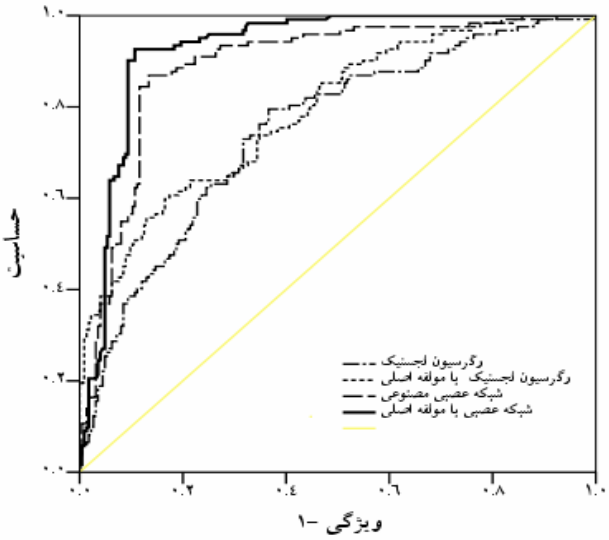
مدل	حساسیت	ویژگی	LR+*	LR-**	آماره کاپا	سطح زیر منحنی راک
رگرسیون لجستیک	۰/٤٨٣	۰/٨٥٧	۳/۳۸	۰/٦۰	۰/۳۲۲	۰/۷٤۹
رگرسیون لجستیک با مولفه اصلی	۰/٤۳٥	۰/۹۱۹	٥/۳۷	۰/٦۱	۰/۳۸٦	۰/۷۹۰
شبکه عصبی مصنوعی	۰/۸۳٦	۰/۸۹۲	۷/۷٤	۰/۱۸	۰/۷۱۲	۰/۸۹۰
شبکه عصبی مصنوعی با مولفه اصلی	۰/۹۱۹	۰/۹٦٤	۲٥/٥۲	۰/۰۸	۰/۸۸٦	۰/۹۲۷

*نسبت درستنمایی مثبت (Positive Likelihood Ratio)
**نسبت درستنمایی منفی (Negative Likelihood Ratio)

بحث:

در مدل شبکه عصبی مصنوعی، استفاده از مولفه‌های اصلی به‌جای متغیرهای اولیه باعث افزایش حساسیت، ویژگی و نسبت درستنمایی مثبت و کاهش نسبت درستنمایی منفی شده در حالیکه در مدل رگرسیون لجستیک استفاده از مولفه‌های اصلی به‌جای متغیرهای اولیه باعث کاهش حساسیت و افزایش ویژگی، نسبت درستنمایی مثبت و همچنین نسبت درستنمایی منفی شده است (جدول شماره ۳).

بر اساس منحنی مشخصه عملکرد (ROC) در بین مدل‌های ارائه شده، مدل شبکه عصبی مصنوعی با مولفه اصلی دارای بیشترین قدرت پیش‌بینی سندرم متابولیک و مدل رگرسیون لجستیک با متغیرهای اولیه، دارای کمترین قدرت پیش‌بینی سندرم متابولیک بود (نمودار شماره ۱).



نمودار شماره ۱: منحنی راک برای مقایسه قدرت پیش بینی مدل ها

با توجه به اهمیت سندرم متابولیک به‌عنوان یک عامل خطرساز مهم برای بیماری‌های قلبی عروقی و دیابت، دسترسی به مدل‌هایی که با دقت بالا بتواند این بیماری را در افراد پیش‌بینی نماید، مورد توجه است. در این پژوهش از تحلیل مولفه‌های اصلی برای بالا بردن دقت پیش‌بینی در مدل رگرسیون لجستیک و شبکه عصبی مصنوعی به منظور کاهش اثر هم‌خطی‌چندگانه استفاده گردید. ایده استفاده از مولفه‌های اصلی به جای متغیرهای اصلی مطالعه، با هدف کاهش ابعاد داده‌ها از این حقیقت ناشی می‌شود که این متغیرها می‌توانند بازتاب دهنده ارتباط بین مشاهدات باشند (۱۰). بدین منظور دو مدل رگرسیون لجستیک و شبکه عصبی مصنوعی ابتدا با در نظر گرفتن متغیرهای اصلی تحقیق، سپس با استفاده از مولفه‌های اصلی به داده‌ها برازش گردید. یافته‌ها نشان داد که بر اساس شاخص‌های مورد استفاده برای مقایسه مدل‌ها، مدل شبکه عصبی مصنوعی با مولفه‌های اصلی نسبت به همه مدل‌های دیگر برتری داشت. استفاده از مولفه‌های اصلی در مدل رگرسیون لجستیک باعث کاهش حساسیت و افزایش ویژگی شد، در حالی که در مدل شبکه عصبی مصنوعی استفاده از مولفه‌های اصلی حساسیت و ویژگی را افزایش داد. ضمن اینکه با استفاده از مولفه‌های اصلی، مقدار آماره کاپا و سطح زیر منحنی مشخصه عملکرد در هر دو مدل شبکه عصبی مصنوعی و رگرسیون لجستیک افزایش نشان می‌دهد. بر اساس

شاخص سطح زیر منحنی مشخصه عملکرد که در این مطالعه مبنای اصلی مقایسه دقت پیش‌بینی مدل‌ها بود، به‌ترتیب مدل‌های شبکه عصبی مصنوعی با مولفه‌های اصلی، شبکه عصبی مصنوعی کلاسیک، رگرسیون لجستیک با مولفه‌های اصلی و رگرسیون لجستیک بیشترین دقت پیش‌بینی را برای سندرم متابولیک داشتند. همچنین یافته‌ها نشان می‌دهد علی‌رغم در نظر گرفتن مولفه‌های اصلی در مدل رگرسیون لجستیک، همچنان مدل شبکه عصبی مصنوعی نسبت به مدل رگرسیون لجستیک با مولفه‌های اصلی دارای صحت پیش‌بینی بیشتری است که این موضوع با یافته‌های مطالعات دیگر هم‌خوانی دارد (۷). یافته‌های این تحقیق با نظر پورحسینقلی و همکاران که در مقاله خود نشان داده‌اند استفاده از تحلیل مولفه‌های اصلی در مدل رگرسیون لجستیک با داده‌های هم‌خط می‌تواند باعث بهبود برآوردها گردد، هم‌خوانی دارد. ایشان مدل خود را بر روی داده‌های مربوط به سرطان سینه اجرا کردند (۸). همچنین نتایج این مطالعه نظر Bucinski و همکاران را که در مطالعه خود نشان دادند استفاده از مولفه‌های اصلی در مدل شبکه عصبی مصنوعی باعث افزایش دقت پیش‌بینی سرطان سینه می‌گردد را تایید می‌کند (۱۶).

علی‌رغم تمامی مزایایی که مدل‌های شبکه عصبی دارند، این مدل‌ها دارای محدودیت‌هایی نیز می‌باشند، از جمله این‌که در مدل‌های شبکه عصبی با توجه به اینکه توزیع پارامترهای شبکه مشخص نمی‌باشد، امکان انجام استنباط آماری برای پارامترها نیز وجود ندارد. از معایب دیگر مدل شبکه عصبی این است که بر خلاف مدل‌های رگرسیون، در مدل شبکه عصبی کلاسیک امکان تعیین میزان تاثیر هر یک از متغیرهای کمکی در پیش‌بینی متغیرهای پاسخ وجود ندارد، مگر اینکه از روش‌های بهینه سازی مانند الگوریتم ژنتیک استفاده گردد (۷).

تفسیر مدل‌هایی که بر اساس مولفه‌های اصلی به‌دست می‌آیند قدری پیچیده است، به‌طوری که با به‌کارگیری مولفه‌های اصلی به‌جای متغیرهای اصلی بررسی اثر هر یک از متغیرهای کمکی بر روی متغیر وابسته مشکل می‌شود، با این‌حال روش‌هایی نیز برای تفسیر این مولفه‌ها در مدل کاهش یافته پیشنهاد شده است (۸).

به هر حال، از آنجا که در این پژوهش، هدف، مقایسه دقت پیش‌بینی مدل‌های مختلف است و نه بررسی اثر متغیرهای کمکی بر روی متغیر پاسخ، محدودیت‌های ذکر شده در مدل شبکه عصبی و تحلیل مولفه‌های اصلی در ارتباط با تفسیر اثر متغیرهای کمکی در پژوهش حاضر مشکلی ایجاد نمی‌کند.

مطالعاتی که تاکنون در ارتباط با مدل‌سازی آماری در باره سندرم متابولیک انجام شده است (۶،۵)، به طور عمده با هدف بررسی عوامل موثر بر سندرم متابولیک انجام شده‌اند، اما در مطالعه حاضر، هدف بررسی صحت و دقت پیش‌بینی مدل‌های شبکه عصبی مصنوعی و رگرسیون لجستیک با استفاده از مولفه‌های اصلی در مقایسه با مدل‌های ذکرشده بدون در نظر گرفتن مولفه‌های اصلی برای پیش‌بینی سندرم متابولیک است که تاکنون بررسی نشده است.

استفاده از روش‌های بهینه‌سازی شبکه‌های عصبی مانند الگوریتم ژنتیک و مدل‌های شبکه عصبی بیزی یا استفاده از روش‌های دیگر مقابله با هم خطی چند گانه مانند رگرسیون ستیغی (Ridge Regression)، همچنین انجام یک مطالعه شبیه‌سازی برای بررسی قابلیت تعمیم نتایج این مطالعه در ارتباط با استفاده از مولفه‌های اصلی به‌جای متغیرهای کمکی برای تعدیل اثر هم‌خطی‌چندگانه در مدل‌های کلاسیک و مدل شبکه عصبی مصنوعی در حالت کلی از مواردی است که علاقمندان می‌توانند در این زمینه مطالعاتی را انجام دهند.

نتیجه‌گیری:

تحقیق نشان داد که دقت و صحت پیش‌بینی مدل‌های رگرسیون لجستیک و شبکه عصبی مصنوعی بر اساس مولفه‌های اصلی نسبت به مدل‌هایی که بر اساس متغیرهای هم‌خط اولیه برازش شدند، بیشتر بوده و بنابراین استفاده از تحلیل مولفه‌های اصلی زمانی‌که متغیرهای کمکی دارای همبستگی باشند، می‌تواند در بالا بردن دقت پیش‌بینی مدل رگرسیون لجستیک و مدل شبکه عصبی مصنوعی موثر باشد.

تشکر و قدردانی:

بدینوسیله از پژوهشکده‌ی علوم غدد درون‌ریز و متابولیسم دانشگاه علوم پزشکی شهید بهشتی که داده‌های این مطالعه را در اختیار پژوهشگران قرار دادند، سپاسگزاری می‌نماییم.

منابع:

1. Hadaegh F, Ghasemi A, Padyab M. [Assessment of different definitions of metabolic syndrome in incidence of diabetes in Iranian urban. Tehran Lipid and Glucose Study. Iran J Diabetes Lipid Disord. 2008; 3: 343-53.]Persian

2. Fakhrzadeh H, Ebrahimpour P, Nouri M. [Survey of prevalence of metabolic syndrome and its risk factors in an urban population. Iran J Diab Lipid disord. 2005; 3: 278-88.]Persian

3. Azizi F, Salehi P, Etemadi A, Zahedi-Asl S. Prevalence of metabolic syndrome in an urban population: Tehran Lipid and Glucose Study. Diabetes Res Clinic Pract. 2003; 61(1): 29-37.

4. Sadr-Bafghi SM, Salari M, Rafiei M, Nemayande SM. [Survey of prevalence of metabolic syndrome and its factors in an urban population. J Medicine Dep Tehran Univ Med J. 2007; 64: 90-6.]Persian

5. Hadaegh F, Zabetian A, Harati H, Azizi F. Metabolic syndrome in normal-weight Iranian adults. Ann Saudi Med. 2007; 27(1): 18-24.

6. Danesh-Pour MS, Mehrabi Y, Hedayati M, Azizi F. [Multivariable survey of factors correlated with metabolic syndrome using factor analysis. Iran J Endocrino Metab. 2006; 30: 139-46.]Persian

7. Sedehi M, Mehrabi Y, Kazemnejad A, Hadaegh F. [Comparison of artificial neural network, logistic regression and discriminate analysis methods in prediction of metabolic syndrome. Iran J Endocrino Metab. 2010; 11(6): 638-46.]Persian

8. Pourhoseingholi MA, Mehrabi Y, Alani-Majd H, Yavari P. [Using latent variables in logistic regression model to eliminate the effect of multicolinearity in analysis of factors associated with breast cancer. Iran J Epidemiol. 2005. 1(2): 41-5.]Persian

9. Menhaj MB. Basics of neural network. Tehran: Prof Hesabi Ins; 2005.

10. Jobson DJ. Applied multivariate data analysis. NewYork: Springer Verlag; 1992.

11. Hagan MT. Neural networks design. Boston: PSW Pub. 1996.

12. Duh MS, Walker AM, Ayania JZ. Epidemiologic interpretation of artificial neural networks. Am J Epidemiol. 1998; 147(12): 112-22.

13. Chakraborty S. Bayesian neural networks for bivariate binary data: an application to prostate cancer study. Stat Med. 2005; 24(23): 3645-62.

14. Sadat Hashemi SM, Kazemnejad A, Lucas C. [Architect of artificial neural network for modeling of multivariate binary responses and its application to predicting type of unwanted pregnancy. [Dissertation]. Tarbiat Modares University; 2003.]Persian

15. Biglarian A. [Application of ANN in determining important predictors of in hospital mortality after coronary artery bypass graft surgery and its comparison with logistic regression. Modares J Med Sci. 2005; 1: 21-30.]Persian

16. Bucinski A, Baczek T, KrysinSki J, Szoszkiewicz R, Zatuski J. Clinical data analysis using ANN and PCA of patients with breast cancer after mastectomy. Rep Pract Oncol Radiother. 2007; 12(1): 9-17.

Journal of Shahrekord University of Medical Sciences
(J Shahrekord Univ Med Sci) 2011 Oct, Nov; 13(4): 18-27.

Original article

Using principal component analysis to increase accuracy of prediction of metabolic syndrome in artificial neural network and logistic regression models

Sedehi M (PhD)*[1], Mehrabi Y (PhD)[2], Khodabakhshi A (PhD)[3]
[1]Medical Plants Research Center, Shahrekord University of Medical Sciences, Shahrekord, Iran, [2]Research Institute for Endocrine Sciences, Shahid Beheshti University of Medical Sciences, Tehran, Iran, [3]Environmental Health Engineering Dept, Shahrekord University of Medical Sciences, Shahrekord, Iran.

Received: 4/Jan/2011 Revised: 16/Apr/2011 Accepted: 5/Jun/2011

Background and aims: In modeling process, correlation between covariates causes multicolinearity that may reduce efficiency of the model. This study was aimed to use principal component analysis to eliminate the effect of multicolinearity in logistic regression and neural network models, and to determine its effect on the accuracy of predicting metabolic syndrome in a sample of individuals participating in the Tehran Lipid and Glucose Study.

Methods: A total of 347 participants from the Cohort section of the Tehran Lipid and Glucose Study (TLGS) were evaluated. The subjects were free of metabolic syndrome, according to the ATPIII criteria, at the beginning. Logistic regression, logistic regression with principal components, neural network and neural network with principal components models were fitted to the data. The ability of the models in predicting metabolic syndrome was compared using ROC analysis and kappa statistics.

Results: The area under receiver operating characteristic (ROC) curve for logistic regression, logistic regression with principal components, neural network and neural network with principal component were estimated as 0.749, 0.790, 0.890 and 0.927 respectively. Sensitivity of the models was calculated as 0.483, 0.435, 0.836 and 0.919 and their specificity as 0.857, 0.919, 0.892 and 0.964 respectively. The kappa statistic for these models was 0.322, 0.386, 0.712 and 0.886 respectively.

Conclusion: the study shows that the prediction accuracy of models based on principal components is better than that of models based on primary covariates, so in the presence of multicolinearity, models based on principal components are efficient for predicting metabolic syndrome.

Keywords: **Artificial neural network, Principal component analysis, Prediction, Multicoline, Metabolic syndrome, Logistic regression.**

Cite this article as: Sedehi M, Mehrabi Y, Khodabakhshi A. [Using principal component analysis to increase accuracy of prediction of metabolic syndrome in artificial neural network and logistic regression models. J Sharekord Univ Med Sci. 2011 Oct, Nov; 13(4): 18-27.]Persian

***Corresponding author:**
Biostatistics and Epidemiology Dept, Health faculty, Rahmatieh, Shahrekord, Iran, Tel: 00983813334678, E-mail:sedehi56@gmail.com