مجله تخصصی اپیدمیولوژی ایران؛ زمستان ۱۳۹۸، دوره ۱۵، شماره ٤: صفحات ۳۷۰-۳٦۲.

مقاله پژوهشی

مقایسه پیش‌بینی ابتلا به دیابت بارداری با مدل‌های رگرسیون لجستیک، تحلیل ممیزی، درخت تصمیم و شبکه عصبی مصنوعی

منصور رضایی[۱]، نگین فخری[۲]، سوده شهسواری[۳]، فاطمه رجعتی[٤]

[۱] استاد آمار زیستی، مرکز تحقیقات باروری و ناباروری، دانشگاه علوم پزشکی کرمانشاه، کرمانشاه، ایران

[۲] اکارشناسی ارشد آمار زیستی، کمیته تحقیقات دانشجویی، دانشگاه علوم پزشکی کرمانشاه، کرمانشاه، ایران

[۳] استادیار، دانشکده پیراپزشکی، دانشگاه علوم پزشکی کرمانشاه، کرمانشاه، ایران

[۴] دانشیار ، مرکز تحقیقات عوامل محیطی مؤثر بر سلامت، دانشگاه علوم پزشکی کرمانشاه، کرمانشاه، ایران

نویسنده رابط: نگین فخری، نشانی: کرمانشاه، دانشگاه علوم پزشکی کرمانشاه، تلفن: ۰۹۱۸۴۷۶۸۹۳۸، پست الکترونیک: n.fakhri94@yahoo.com

تاریخ دریافت: ۹۸/۰۱/۲۲؛ پذیرش: ۹۸/۰۶/۰۹

مقدمه و اهداف: دیابت بارداری شایع‌ترین اختلال متابولیک دوران بارداری است. در صورت تشخیص زودرس این بیماری می‌توان از برخی عوارض آن جلوگیری کرد. هدف این پژوهش پیش بینی زودرس ابتلا به دیابت بارداری بوسیله مدل های رگرسیون لجستیک، تحلیل ممیزی، درخت تصمیم و شبکه عصبی مصنوعی پرسپترون و مقایسه این مدل‌ها بود.

روش کار: پرونده ۴۲۰ خانم باردار (۱۳۹۱-۱۳۸۹) دارای پرونده در مراکز بهداشتی کرمانشاه، با روش نمونه‌گیری در دسترس بررسی شد. اطلاعات جمعیت شناختی، متغیرهای مربوط به دوره بارداری و نتایج آزمایش‌ها و ابتلا به دیابت بارداری با معیار قند خون ناشتا بزرگ‌تر یا مساوی ۹۲ از پرونده آنان گردآوری شد. پس از برازش چهار مدل فوق به داده‌ها، عملکرد مدل‌ها باهم مقایسه گردید و با توجه به معیارهای صحت، حساسیت و ویژگی بر اساس منحنی ROC، مدل برتر معرفی شد.

یافته‌ها: پس از برازش مدل‌های رگرسیون لجستیک، تحلیل ممیزی، درخت تصمیم و شبکه عصبی مصنوعی به مجموعه داده‌ها، معیار صحت برای مدل‌های مذکور به ترتیب برابر ۰/۸۱، ۰/۸۳، ۰/۷۸ و ۰/۸۳ ، حساسیت ۰/۵۰، ۰/۶۳، ۰/۵۸و ۰/۵۸، ویژگی ۰/۹۶، ۰/۹۳، ۰/۸۷ و ۰/۹۴ و سطح زیر منحنی ROC به ترتیب برابر ۰/۸۶، ۰/۷۸، ۰/۷۳ و ۰/۸۷ محاسبه گردید.

نتیجه‌گیری: در پیش‌بینی و رده‌بندی ابتلا و عدم ابتلا به دیابت بارداری، مدل شبکه عصبی مصنوعی دارای نرخ دسته‌بندی اشتباه کمتر و سطح زیر منحنی ROC بیشتری نسبت به سایر مدل‌ها بود. می‌توان نتیجه گرفت که این مدل دارای پیش‌بینی‌های صحیح‌تر و نزدیک به واقعیت نسبت به سایر مدل‌ها است.

واژگان کلیدی: دیابت بارداری، صحت، حساسیت، ویژگی، منحنی ROC

مقدمه

دیابت بارداری[۱] (GDM) عبارت است از عدم تحمل کربوهیدرات‌ها با شدت‌های مختلف، که یا شروع آن در زمان بارداری است و یا اولین بار در زمان بارداری تشخیص داده می‌شود (۱). این بیماری شایع‌ترین اختلال متابولیک دوران بارداری است (۲) که با عوارض خطرناکی برای مادر و جنین همراه است. فراوانی GDM در نقاط مختلف دنیا بین۱۴-۱ درصد گزارش‌شده است (۳). شیوع GDM در ایالت متحده آمریکا ۳-۱ درصد، در کشورهای آسیایی به‌طور متوسط ۱۰,۹ درصد، در اروپا ۵,۲ درصد (۴) و در ایران درمجموع۹,۴ درصد برآورد شده است (۵). از جنبه‌های مختلف، GDM می‌تواند برای مادر و جنین زیان‌آور باشد. ازجمله عوارض مادری آن می‌توان به افزایش خطر پره‌اکلامپسی، صدمات زایمانی ناشی از ماکروزومی جنین، پلی هیدرآمنیوس و شیوع بیشتر عفونت‌های باکتریایی و قارچی اشاره کرد. همچنین نفروپاتی و رتینوپاتی دیابتی نیز ممکن است همراه با بارداری پیشرفت کند (۶). مطالعات نشان داده است، در جوامعی که دیابت نوع ۲ شیوع بیشتری دارد، دیابت بارداری نیز شایع است، اما خطر و زمان شروع این بیماری کاملاً متغیر است (۱, ۷, ۸). ازجمله عوارض جنینی دیابت بارداری، ماکروزومی جنین، هیپوگلیسمی، هیپوکلسمی، هیپربیلی روبینمی و افزایش موارد مرگ‌ومیر پری ناتال است. بر اساس مطالعات متعدد و شواهد موجود، کنترل قند خون مادر مبتلا می‌تواند خطر ایجاد

[۱] Gestational Diabetes Mellitus

عوارض فوق را به میزان زیادی کاهش دهد. عوارضی مانند مرگ داخل رحمی، هیپوگلسیمی و هیپربیلی روبینمی نوزادی قابل‌کنترل می‌باشند و عوارضی مثل ماکروزومی، دیستوشی شانه و همچنین زایمان سزارین ۵۰ درصد کاهش می‌یابد (۶).

تاکنون آزمایش‌ها متفاوتی است ازجمله قند خون ناشتا و آزمون تحمل گلوکز خوراکی ۵۰ گرم برای تشخیص دیابت بارداری مورداستفاده قرارگرفته است (۹, ۱۰). در برخی مطالعات نیز جهت غربالگری دیابت بارداری آزمون گلوکز خوراکی ۱۰۰ گرم پیشنهادشده و نشان داده‌شده که قند خون دوساعته با نقطه برش بزرگ‌تر یا مساوی ۱۵۳ mg/dl ، بهترین حساسیت و ویژگی را در تشخیص دیابت بارداری دارد (۶). سازمان جهانی بهداشت و کارگروه مطالعات بارداری انجمن بین‌المللی دیابت در سال ۲۰۱۳، داشتن قند خون ناشتا بیشتر یا مساوی ۹۲ را در هفته ۲۶ تا ۳۰ بارداری را به‌عنوان یکی از معیارهای تشخیصی دیابت بارداری توصیه نمود (۹, ۱۱) و به دلیل وجود این متغیر در پرونده مادران باردار این مطالعه، این معیار جهت تشخیص دیابت بارداری در نظر گرفته شد.

با توجه به شیوع بالای دیابت بارداری و اهمیت تشخیص به‌موقع آن برای پیشگیری از عوارض در مادر و جنین، توسعه معیارهای تشخیصی دیابت بارداری که توانایی کشف بارداری‌های پرخطر ناشی از قند خون بالای مادر را داشته باشند، بسیار اهمیت دارد (۱۲). همچنین با توجه به اینکه در صورت تشخیص زودرس دیابت بارداری می‌توان برخی عوامل خطر آن را کنترل نمود، بنابراین استفاده از روش‌هایی که بتواند در ماه‌های اولیه بارداری، ابتلای فرد به دیابت بارداری در ماه‌های آخر بارداری را پیش‌بینی کند، به‌عنوان ابزاری کمکی در تشخیص زودرس دیابت بارداری، کاربردی و مفید به نظر می‌رسد. یکی از رویکردهایی که در این زمینه وجود دارد استفاده از مدل‌های آماری به‌عنوان ابزاری کمکی در کنار تشخیص پزشک جهت تشخیص زودرس دیابت بارداری است.

در علم آمار برای رده‌بندی داده‌ها، مدل‌های مختلفی تحت دودسته کلی مدل‌های کلاسیک و مدل‌های هوش مصنوعی در دسترس هستند. مدل‌های رگرسیون لجستیک و تحلیل ممیزی ازجمله مدل‌های کلاسیک هستند و درخت تصمیم و شبکه عصبی مصنوعی نیز جزء مدل‌های هوش مصنوعی می‌باشند که می‌توانند به‌منظور رده‌بندی داده‌ها مورداستفاده قرار گیرند.

در مطالعات بسیاری جهت پیش‌بینی ابتلا و عدم ابتلا به بیماری‌های مختلف از مدل‌های آماری متفاوتی استفاده‌شده است. در مطالعه Boonjing (۲۰۱۰)، عملکرد مدل‌های رگرسیون لجستیک، درخت تصمیم و شبکه عصبی مصنوعی در پیش‌بینی بیماری قلبی بررسی شد. در این مطالعه حساسیت، ویژگی و صحت پیش‌بینی مدل شبکه عصبی مصنوعی به ترتیب ۸۱,۱، ۷۸,۷ و ۸۰,۲ درصد، برای مدل درخت تصمیم به ترتیب ۸۱,۷، ۰,۷۶ و ۷۹,۳ درصد و برای مدل رگرسیون لجستیک به ترتیب ۸۱,۲ ، ۷۳,۱ و ۷۷,۷ درصد به دست آمد. در این مطالعه شبکه عصبی مصنوعی نرخ خطای کمتر و صحت بیشتری در پیش‌بینی بیماری قلبی نسبت به دو مدل دیگر داشت (۱۳). با توجه به وجود مدل‌های مختلف برای رده‌بندی داده‌ها، برای دستیابی به مدل بهینه باید مقایسه مدل‌ها انجام شود. بررسی مطالعات قبلی در زمینه دیابت بارداری، نشان می‌دهد که تاکنون جهت تشخیص دیابت بارداری استفاده چندانی از مدل‌های آماری مورداستفاده در این مطالعه نشده است. لذا هدف از انجام این مطالعه پیش‌بینی ابتلا به دیابت بارداری با استفاده از چهار مدل رگرسیون لجستیک، تحلیل ممیزی، درخت تصمیم و شبکه عصبی مصنوعی و مقایسه عملکرد مدل‌های مذکور در تشخیص زودرس دیابت بارداری است.

روش کار

داده‌های پژوهش: جامعه موردبررسی مادران باردار مراجعه‌کننده به مراکز بهداشتی درمانی شهرستان کرمانشاه در سال‌های ۱۳۹۱-۱۳۸۹ بودند که جهت دریافت مراقبت‌های دوران بارداری پرونده پزشکی تشکیل دادند. نمونه‌ها به روش نمونه‌گیری در دسترس (آسان) به این شکل انتخاب شدند که از مراکز موجود در هر کدام از ۵ منطقهٔ شمال، جنوب، شرق، غرب و مرکز شهر کرمانشاه، یک مرکز به‌صورت تصادفی انتخاب‌شده و از پرونده‌های موجود در آن مرکز که بین ۲۵۰۰ تا ۳۰۰۰ پرونده بود، حدود ۱۵۰ پرونده موردبررسی قرار گرفت (به دلیل حجم بسیار زیاد پرونده‌ها، امکان بررسی کل پرونده‌ها نبود) و اطلاعات موردنظر توسط همکاران آموزش‌دیده با رعایت موازین اخلاقی از پرونده‌ها استخراج شد. از پرونده‌های موجود، پرونده‌هایی که اطلاعات موردنیاز در آن کامل ثبت‌شده و هیچ داده گمشده نداشتند (۴۲۰ پرونده) بررسی شدند، متغیرهایی که طبق بررسی متون گذشته به‌عنوان عوامل خطر دیابت بارداری معرفی‌شده بودند و در پرونده‌ها ثبت‌شده بود از پرونده‌ها استخراج گشت. متغیرهای سن و تحصیلات مادر و پدر، رتبه بارداری مادر، فاصله از بارداری قبلی مادر، دیابت مادر در اولین مراجعه، میزان هموگلوبین (Hb)، هماتوکریت (Hc)، قند خون ناشتا (FBS)، پلاکت مادر در هفته ۶ تا

۱۰ بارداری، RH خون و گروه خونی مادر، ابتلای مادر به سیفلیس و عفونت ادرار، میزان کراتینین و اوره مادر، یک‌قلو یا چند قلویی جنین، BMI، فشارخون سیستولیک و دیاستولیک مادر در ماه اول بارداری به‌عنوان متغیرهای پیشگو وارد مدل شدند. همچنین ابتلای مادر به دیابت بارداری در هفته ۲۶ تا ۳۰ بارداری با معیار قند خون ناشتا بزرگ‌تر یا مساوی ۹۲، به‌عنوان متغیر پاسخ در نظر گرفته شد. ابتلا به دیابت در اولین مراجعه به‌عنوان معیار خروج مادران از مطالعه در نظر گرفته شد. جهت محرمانه ماندن اطلاعات بیماران، اطلاعات پرونده‌ها بدون نام افراد و تنها با شماره پرونده مورد دسترسی و تجزیه‌وتحلیل قرار گرفت.

با توجه به اینکه در مطالعات مدل‌سازی باید حجم نمونه بین $5m$ و $15m$ باشد که در آن m تعداد متغیرهای مورداستفاده است و در مطالعه حاضر جمعاً ۲۲ متغیر استفاده‌شده بنابراین حداقل حجم نمونه باید در بازه ۱۱۰ تا ۳۳۰ باشد، حجم نمونه این مطالعه (۴۲۰)، حداقل حجم لازم را برآورده می‌کند.

پس از ورود اطلاعات به کامپیوتر، تجزیه‌وتحلیل داده‌ها با استفاده از نرم‌افزار R نسخه۳،۳،۳ و نرم‌افزار SPSS نسخه ۲۵ انجام گرفت.

ابتدا با استفاده از آزمون t برای متغیرهای کمی و آزمون کای دو برای متغیرهای کیفی، متغیرهایی که در دو گروه (افراد مبتلا و غیرمبتلا) تفاوت معنی‌دار داشتند مشخص شدند. علاوه بر متغیرهای معنی‌دار، ۳ متغیر سن، فاصله از بارداری قبلی و هموگلوبین نیز در مطالعات گذشته جزء متغیرهای مهم ابتلا به دیابت بارداری معرفی‌شده (۱۴, ۱۵) و در مجموعه متغیرهای مطالعه حاضر نیز وجود داشت، بنابراین مدل‌سازی ۴ مدل رگرسیون لجستیک، تحلیل ممیزی، درخت تصمیم و شبکه عصبی مصنوعی با مجموعه‌ای شامل ۸ متغیر سن، فاصله از بارداری قبلی، هموگلوبین، هماتوکریت، قند خون ناشتا، شاخص توده بدنی، فشارخون سیستول و فشارخون دیاستول هفته‌های ۶-۱۰ بارداری انجام شد. پس از ساختن چهار مدل آماری با استفاده از ۸ متغیر مذکور، درنهایت عملکرد مدل‌ها با استفاده از شاخص‌های صحت، حساسیت، ویژگی و سطح زیر منحنی ROC باهم مقایسه شدند.

برای دو مدل داده‌کاوی شبکه عصبی و درخت تصمیم که ممکن است در هر بار اجرای مدل، از روش جستجوی متفاوتی استفاده کنند و مقادیر شاخص‌ها متفاوت شوند، مدل‌ها مکرراً (۳۰۰ بار) اجراشده و میانگین شاخص‌ها محاسبه و گزارش شد.

مدل‌های آماری

مدل رگرسیون لجستیک[1] شکلی از رگرسیون است که در آن متغیر وابسته به‌صورت رده‌ای دوسطحی است. متغیرهای مستقل می‌توانند هم در مقیاس کمی و هم در مقیاس رده‌ای باشند. مجموع احتمال تعلق به سطوح برابر ۱ است. این مدل به‌صورت زیر است (۱۶):

$$\text{logit}(p) = \ln\left(\frac{p}{1-p}\right) = \alpha + \beta_1 x_1 + \cdots + \beta_k x_k$$

که در آن:

$$p = p(Y=1) = \frac{e^{\alpha+\beta_1 x_1+\cdots+\beta_k x_k}}{1+e^{\alpha+\beta_1 x_1+\cdots+\beta_k x_k}}$$

در این مدل P احتمال تعلق فرد به رده اول متغیر وابسته است. Xi متغیر مستقل iام و β_i ضریب برآورد شده مدل برای متغیر مستقل iام است (۱۷). از مزایای استفاده از مدل رگرسیون لجستیک علاوه بر مدل‌سازی مشاهده‌ها، امکان پیش‌بینی احتمال تعلق هر فرد به هریک از رده‌های متغیر وابسته و همچنین امکان محاسبه مستقیم نسبت شانس با استفاده از ضرایب مدل وجود دارد.

مدل تحلیل ممیزی[2] بر مبنای مشاهدات دارای ماهیت چند متغیره است و عمل تشخیص و تمیز بین جمعیت‌ها را انجام می‌دهد. هدف کلی در اینجا به وجود آوردن یک ترکیب خطی بین متغیرها بر مبنای اندازه‌های حاصل از افراد به‌صورت زیر است:

$$L = \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$$

می‌توان از این روش برای گروه‌بندی افراد و پیش‌بینی احتمال تعلق فرد به یک گروه خاص استفاده کرد (۱۸). جهت ساختن این ترکیب خطی مطابق با روش آر. اِ. فیشر، از دو رده (جمعیت) نمونه تصادفی گرفته و برای هر یک از افراد نمونه P متغیر تصادفی همبسته به نام‌های $X_1, X_2, \ldots, X_p$ اندازه‌گیری می‌شود. پس از محاسبه کمیت‌های $\bar{X}_1$ ،$\bar{X}_2$ و S^{-1}_{pooled} ، آنگاه تابع ممیز خطی فیشر به‌صورت زیر محاسبه می‌شود:

$$\hat{y} - (\bar{X}_1 - \bar{X}_2)' S^{-1}_{pooled} X$$

حال برای پیش‌بینی تعلق نمونه جدید به یکی از دو رده ۱ و ۲، یک قاعده تخصیص مبتنی بر تابع ممیز فیشر(۱۹)، با تعریف y_0 به‌صورت زیر به کار می‌رود:

$$y_0 = (\bar{X}_1 - \bar{X}_2)^T S^{-1}_{poold} X_0$$

$y_0 \geq \hat{m}$ را به جامعه ۱ اختصاص می‌دهد هرگاه: X_0

$y_0 < \hat{m}$ را به جامعه ۲ اختصاص می‌دهد هرگاه: X_0

که در این روابط $\hat{m}$ به‌صورت زیر است:

[1] Logistic Regression
[2] Discriminant Analysis

$$\hat{m} = \frac{1}{2}(\bar{X}_1 - \bar{X}_2)' S_{pooled}^{-1}(\bar{X}_1 + \bar{X}_2)$$

مدل درخت تصمیم با الگوریتم CART[1] یکی از قوی‌ترین الگوریتم‌های داده‌کاوی است که برای کاوش داده‌ها و کشف دانش کاربرد دارد. در درخت تصمیم، سؤالاتی در مورد متغیرهای پیشگو پرسیده می‌شود که نمونه آموزشی را به قسمت‌های کوچک‌تری تقسیم می‌کند. الگوریتم CART یکی از الگوریتم‌های درخت تصمیم است که معمولاً از معیار ضریب جینی برای تقسیم داده‌ها به قسمت‌های مختلف استفاده می‌کند. در الگوریتم CART فقط سؤالاتی پرسیده می‌شود که دارای جواب دوحالتی است. همچنین در این الگوریتم، ساختار درخت، نسبت به تغییرات یکنواخت متغیرهای مستقل، تغییر نمی‌کند (۲۰). در الگوریتم CART دو مرحله رشد درخت (Maximum Tree) و هرس کردن درخت (Pruning Tree) را داریم. در مرحله رشد درخت، با شروع از گره ریشه، الگوریتم CART تمام متغیرها و تمام مقادیر ممکن متغیرها برای تقسیم شدن را چک می‌کند تا بهترین تقسیم در گره انجام شود. در انتخاب بهترین تقسیم‌کننده، به دنبال ماکزیمم کردن میانگین خلوص دو گره فرزند هستیم (۲۱). در این الگوریتم برای یافتن هر متغیری که باید تقسیم شود و نیز برای یافتن بهترین نقطه تقسیم در هر متغیر، معمولاً از شاخص جینی (Gini Index) استفاده می‌شود. برای یافتن متغیری که باید تقسیم روی آن انجام شود، به ازای مجموعه داده‌های D که به دو مجموعه D_1 و D_2 تقسیم شود برای متغیر A، داریم (۲۲):

$$Gini(D) = 1 - \sum_{i=1}^{c} p_i^2$$
$$Gini_A(D) = \frac{|D_1|}{|D|} \times Gini(D_1) + \frac{|D_2|}{|D|} \times Gini(D_2)$$

برای هر یک از متغیرها، پس از محاسبه Gini Index برای همه حالات، مقدار حداقل انتخاب‌شده و آن متغیر برای گره جاری در نظر گرفته می‌شود. برای یافتن بهترین نقطه تقسیم در متغیری که باید تقسیم شود، از تابع ناخالصی $i(t)$ استفاده می‌کنیم:

$$i(t) = \sum_{k \neq l} p(k|t)p(l|t)$$

در هر گره، الگوریتم CART بهترین نقطه تقسیم برای متغیر x_j یعنی x_j^R را به‌صورت زیر به‌گونه‌ای پیدا می‌کند که حداکثر همگنی را در گره‌های فرزند چپ و راست داشته باشیم.

$$\arg\max_{x_j \leq x_j^R} [i(t_p) - P_l\, i(t_l) - P_r\, i(t_r)]$$

در مرحله هرس کردن درخت، روش معمول هرس کردن بر اساس بهینه کردن تابعی از پیچیدگی درخت و خطای دسته‌بندی اشتباه است (۲۳). تابع هزینه-پیچیدگی به‌صورت زیر است:

$$R_\alpha(T) = R(T) + \alpha(T)$$

که در آن R(T) نرخ خطای دسته‌بندی اشتباه درخت T است. $\alpha(T)$ اندازه پیچیدگی T است (مجموع کل گره‌های پایانی درخت). بهترین درخت هرس شده درختی است که کمترین مقدار $R_\alpha(T)$ را داشته باشد.

مدل شبکه عصبی مصنوعی[2] (ANN) یک سیستم پردازش است که در آن از سیستم‌های عصبی بیولوژیک مانند مغز الهام گرفته است. عضو کلیدی این ساختار جدید، سیستم پردازنده اطلاعات است. شبکه عصبی مصنوعی برای تشخیص، طبقه‌بندی و پیش‌بینی در مجموعه داده‌های بزرگ که در آن‌ها روابط معمولاً به شکل غیرخطی هستند مورداستفاده قرار می‌گیرد (۲۴, ۲۵). شبکه عصبی مصنوعی پرسپترون نوعی از شبکه عصبی است که بر مبنای یک واحد محاسباتی به نام پرسپترون ساخته می‌شود. پرسپترون برداری از ورودی‌ها با مقادیر حقیقی را گرفته و یک ترکیب خطی از این ورودی‌ها را محاسبه می‌کند. اگر مقدار حاصل از یک مقدار آستانه بیشتر بود خروجی پرسپترون برابر با ۱ و در غیر این صورت ۱ - خواهد بود. فرآیند یادگیری در این شبکه‌ها از طریق الگوریتم‌های یادگیری خاصی صورت می‌گیرد که با تنظیم وزن‌های موجود در ارتباطات بین نورون‌ها، اقدام به آموزش شبکه می‌کنند. شبکه عصبی به‌طورمعمول دارای سه لایه ورودی، میانی (مخفی)، و خروجی است که در آن هر لایه ورودی به یک یا تعداد بیشتری لایه میانی مرتبط است و لایه میانی نیز به لایه خروجی مرتبط است. نورون‌ها به هم اتصالاتی دارند و هر اتصال وزنی دارد که بیانگر میزان تأثیر نورون بر نورون دیگر است. شبکه عصبی مصنوعی موردنظر ما پرسپترون چندلایه (MLP) است که در مقایسه با روش‌های دیگر بهتر عمل می‌کند (۲۶). خروجی تمامی واحدهای پردازش از هر لایه به‌عنوان ورودی به واحدهای پردازش لایه بعدی داده می‌شوند. واحدهای پردازش در لایه ورودی همگی خطی هستند ولی در لایه‌های مخفی و خصوصاً لایه خروجی از نورون‌های غیرخطی با تابع تانژانت هیپربولیک و یا سیگموئید و یا هر تابع غیرخطی پیوسته و مشتق‌پذیر دیگری می‌توان استفاده کرد. یادگیری در شبکه‌های عصبی پرسپترون به شکل با نظارت است. در یادگیری با ناظر مجموعه‌ای از زوج داده‌ها به نام نمونه‌های آموزشی به‌صورت زیر داده می‌شود و خروجی

[1] Classification And Regression Tree

[2] Artificial Neural Network

پرسپترون توسط رابطه زیر مشخص می‌شود:

$$O(x1, x2, ..., xn) = \begin{cases} 1 & \text{if} (\sum w_i X_i > 0) \\ -1 & \text{if} (\sum w_i X_i <= 0) \end{cases}$$

یادگیری پرسپترون عبارت است از پیدا کردن مقادیر درستی بری وزن‌های اتصالات بین نورون‌ها. ابتدا مقادیری تصادفی به وزن‌ها نسبت داده‌شده و سپس پرسپترون به تک‌تک مثال‌های آموزشی اعمال می‌شود. اگر مثال غلط ارزیابی گردد، مقادیر وزن‌های پرسپترون تصحیح می‌شوند. این فرآیند تا زمانی که شبکه تمام مثال‌های آموزشی را درست ارزیابی کند ادامه می‌یابد.

یافته‌ها

در این مطالعه ۴۲۰ زن باردار موردبررسی قرار گرفتند. میانگین سنی کل افراد شرکت داده‌شده در مدل‌سازی ۶,۰۳±۲۸,۲۵ سال بود. از این افراد، ۱۳ نفر (۳,۱ درصد) دو یا چند قلو باردار بودند و ۱۵۳ نفر (۳۶,۵ درصد) اولین بارداری خود را تجربه می‌کردند. در نمونه موردبررسی ۱۳۴ نفر (۳۱,۸ درصد) مبتلا و ۲۸۷ نفر (۶۸,۲ درصد) غیر مبتلا به دیابت بارداری بودند. آزمون **t** متغیرهای کمی که در دو گروه (افراد مبتلا و غیرمبتلا) تفاوت معنی‌دار داشتند را مشخص کرد (جدول شماره ۱). آزمون کای دو نشان داد هیچ‌کدام از متغیرهای کیفی در دو گروه تفاوت معنی‌داری ندارند.

مدل‌سازی‌ها با ۸ متغیر سن، فاصله از بارداری قبلی، هموگلوبین و نیز هماتوکریت، قند خون ناشتا، شاخص توده بدنی، فشارخون سیستول و فشارخون دیاستول هفته‌های ۶ تا ۱۰ بارداری انجام شد.

پس از برازش مدل رگرسیون لجستیک به روش گام‌به‌گام و شیوه حذف پسرو، مجموعه‌ای شامل ۴ متغیر برای ورود به مدل انتخاب شدند (جدول شماره ۲) و نهایتاً مدل رگرسیون لجستیک با متغیرهای معنی‌دار ساخته شد (معادله ۱).

معادله ۱ :

$$\ln p/(1-p) = -17.21 - 0.07\,(Age) + 0.15\,(FBS) + 0.16\,(BMI)$$

مدل رگرسیون لجستیک نشان داد که متغیرهای سن، قند خون ناشتا و BMI رابطه معنی‌داری با ابتلا به دیابت بارداری داشتند (P-value<0.05) برای مدل تحلیل ممیزی از آزمون بارتلت جهت تأیید همگنی واریانس‌ها استفاده شد. متغیرها در اختیار مدل قرار گرفتند و آزمون لاندای ویلکس ۵ متغیر را معنی‌دار معرفی کرد (جدول شماره ۳).

آزمون معنی‌داری تابع ممیزی انجام شد

($Chi.square = 191.47$ و $p - value < 0.001$).

معادله مدل تحلیل ممیزی به دست آمد (معادله ۲).

معادله ۲:

$$LD = 0.853\,(FBS) + 0.369\,(BMI) + 0.099\,(DBP) + 0.056\,(Hc) + 0.042\,(SBP)$$

در برازش مدل درخت تصمیم، با توجه به تابع ناخالصی و شاخص $Gini$ ، درخت با عمق ۴ ساخته شد و متغیر FBS هفته ۶ تا ۱۰ بارداری به‌عنوان مهم‌ترین متغیر در ابتلا به دیابت بارداری در گره ریشه قرار گرفت و متغیرهای BMI و سن نیز به‌عنوان متغیرهای پیش‌بینی کننده دیابت بارداری، در گره‌های بعدی در ساخت درخت مورداستفاده قرار گرفتند (شکل شماره ۱).

بهترین مدل شبکه عصبی با یک‌لایه پنهان و ۶ گره در لایه پنهان ساخته شد. تابع فعال‌سازی در لایه پنهان تابع تانژانت هایپربولیک و در لایه خروجی تابع **Softmax** بود. همچنین مقدار SSE مجموعه آزمایشی از مجموعه آموزشی کمتر بوده (۶۲.۲۳ به ۹۱.۰۳) و خطای نسبی در دو مجموعه به هم نزدیک بود (۰.۷ به ۰.۶) که هر دو مورد برازش خوب مدل را نشان می‌دهند. آنالیز حساسیت نشان داد که در مدل شبکه عصبی، متغیرهای FBS ، BMI ، Hc و Hb سه‌ماهه اول بارداری مهم‌ترین عوامل خطر دیابت بارداری هستند.

جهت مقایسه عملکرد ۴ مدل، شاخص‌های ارزیابی عملکرد مدل‌ها محاسبه شد (جدول شماره ۴). مدل بهتر مدلی است که در آن، شاخص‌ها مقادیر بالاتری داشته باشند.

جدول شماره ۱– مقایسه میانگین ( ) متغیرهای کمی پیشگو در دو گروه مبتلا و غیرمبتلا

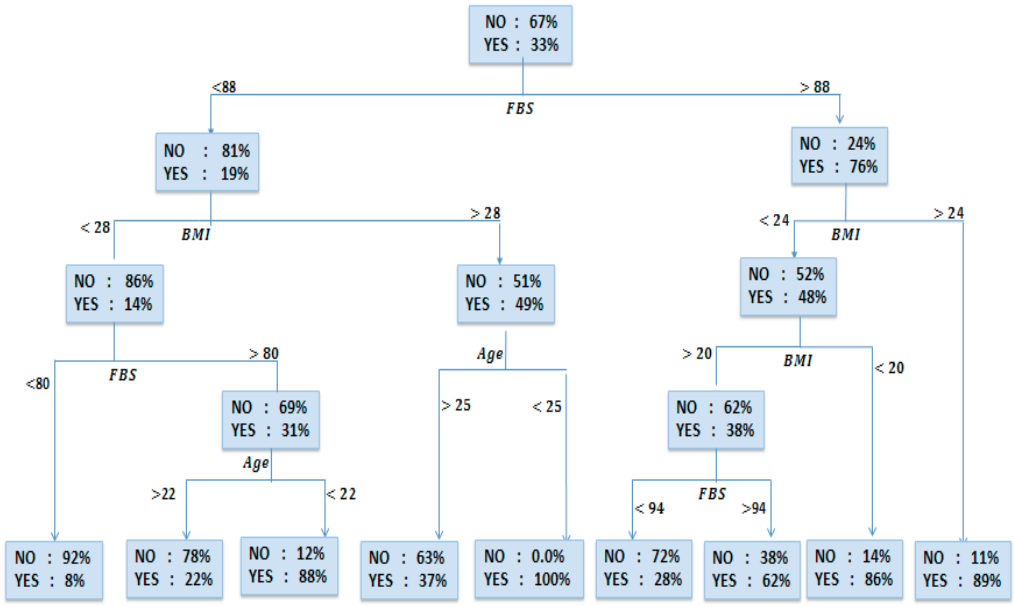
متغیرها	افراد مبتلا	افراد غیر مبتلا	آماره آزمون	P – value
هماتوکریت هفته ۶ تا ۱۰ بارداری	۴۰/۳۶ ± ۳۷/۴	۳۷/۲۸ ± ۳/۱	−۲/۰۳	۰/۰۴۳
قند خون ناشتا هفته ۶ تا ۱۰ بارداری	۸۸/۶۱ ± ۹/۴	۷۶/۵۸ ± ۸/۶	−۱۳/۹۹	< ۰/۰۰۱
فشارخون سیستول ماه اول بارداری	۱۰۳/۴۶ ± ۹/۶	۱۰۱/۱ ± ۱۲/۹	−۲/۰۴	۰/۰۰۶
فشارخون دیاستول ماه اول بارداری	۶۴/۹۹ ± ۷/۵	۶۲/۹۹ ± ۷/۵	−۲/۷۵	۰/۰۰۶
شاخص توده بدنی اولین مراجعه	۲۶/۶۶ ± ۴/۹	۲۳/۶۳ ± ۳/۲	−۸/۱۱	< ۰/۰۰۱

جدول شماره ۲- ضرایب رگرسیونی و اطلاعات متغیرهای باقیمانده در مدل لجستیک بعد از روش گام‌به‌گام

	ضریب بتا (β)	انحراف معیار ضریب β	آماره آزمون	P – value	نسبت شانس
عرض از مبدأ	-۱۷/۲۲	۱/۸۷	-۹/۱۹	< ۰/۰۰۱	-
سن (Age)	-۰/۰۷	۰/۰۲	-۳/۳۷	< ۰/۰۰۱	۰/۹۳
قند خون ناشتا(FBS)	۰/۱۵	۰/۰۱	۹/۴۶	< ۰/۰۰۱	۱/۱۶
فشارخون دیاستول(DBP)	۰/۰۳	۰/۰۱	۱/۸۸	۰/۰۶	۱/۰۳
شاخص توده بدنی(BMI)	۰/۱۶	۰/۰۳	۴/۹۰	< ۰/۰۰۱	۱/۱۷

جدول شماره ۳- نمرات مربوط به متغیرها در مدل تحلیل ممیزی

نام متغیر	ضرایب
قند خون ناشتا (FBS)	۰/۸۵۳
شاخص توده بدنی (BMI)	۰/۳۶۹
فشارخون دیاستول (DBP)	۰/۰۹۹
هماتوکریت (Hc)	۰/۰۵۶
فشارخون سیستول (SBP)	۰/۰۴۲



شکل شماره ۱- نمودار درخت تصمیم

جدول شماره ٤- میزان صحت، حساسیت، ویژگی و سطح زیر منحنی ROC مربوط به ٦ مدل برازش شده

مدل مورداستفاده	صحت	حساسیت	ویژگی	سطح زیر منحنی ROC
رگرسیون لجستیک	۸۱ درصد	۵۰ درصد	۹۶ درصد	۸۶ درصد
تحلیل ممیزی	۸۳ درصد	۶۳ درصد	۹۳ درصد	۷۸ درصد
درخت تصمیم	۷۸ درصد	۵۸ درصد	۸۷ درصد	۷۳ درصد
شبکه عصبی مصنوعی	۸۳ درصد	۵۸ درصد	۹۴ درصد	۸۷ درصد

بحث

با توجه به شیوع بالای دیابت بارداری، دسترسی به روش‌هایی که بتوانند با دقت بالایی این بیماری را در ماه‌های اولیه بارداری پیش‌بینی کنند موردتوجه است. در مطالعه حاضر ۴ مدل آماری مختلف جهت پیش‌بینی زودرس دیابت بارداری استفاده‌شده و عملکرد مدل‌ها مورد مقایسه قرار گرفتند. مدل شبکه عصبی مصنوعی دارای بالاترین صحت بود یعنی نسبت به دیگر مدل‌های موردبررسی میزان دسته‌بندی‌های درست بیشتری داشت. بیشترین ویژگی با مدل رگرسیون لجستیک به دست می‌آید بنابراین این مدل دارای توان بیشتری در تشخیص درست افراد غیر مبتلا است. همچنین بالا بودن حساسیت مدل‌های شبکه عصبی مصنوعی و تحلیل ممیزی در مطالعه ما نشان می‌دهد که این مدل‌ها دارای توان بیشتری در تشخیص درست افرادی هستند که واقعاً مبتلا به دیابت بارداری می‌باشند. سطح زیر منحنی ROC در مدل رگرسیون لجستیک و شبکه عصبی مصنوعی بالاتر از دیگر مدل‌ها بود. به‌طورکلی می‌توان گفت که عملکرد مدل شبکه عصبی مصنوعی بهتر از سایر مدل‌ها و عملکرد درخت تصمیم ضعیف‌تر از دیگر مدل‌های موردبررسی در این مطالعه بود.

بالاتر بودن صحت با مدل شبکه عصبی در پژوهش حاضر هم‌راستا با مطالعه‌ای است که در سال ۱۳۸۸ تحت عنوان مقایسه مدل‌های شبکه عصبی مصنوعی با رگرسیون لجستیک و تحلیل ممیزی در پیش‌بینی سندروم متابولیک با استفاده از داده‌های بانک اطلاعاتی مرکز قند و لیپید تهران (۸۱-۱۳۷۹) به مقایسه سه روش پرداخته شد و یافته‌های آن مطالعه نشان داد که مدل شبکه عصبی مصنوعی نسبت به مدل رگرسیون لجستیک و مدل تحلیل ممیزی از صحت بیشتری برای پیش‌بینی سندروم متابولیک در افراد موردبررسی برخوردار است (۲۷).

نتایج مطالعه حاضر که نشان داد مدل شبکه عصبی مصنوعی و رگرسیون لجستیک دارای سطوح زیر منحنی ROC بالاتری نسبت به دیگر مدل‌ها بودند و نیز آزمون برابری سطوح زیر منحنی ROC نشان داد مدل درخت تصمیم از بقیه مدل‌ها ضعیف‌تر عمل کرده و عملکرد آن تفاوت معنی‌داری با دیگر مدل‌ها دارد (p-value<0.05)، این نتیجه با نتیجه مطالعه Kurt همخوانی دارد. در مطالعه Kurt و همکاران که در سال ۲۰۰۶ بر روی ۱۲۴۵ نفر جهت پیش‌بینی ابتلا به CAD انجام شد، سطح زیر منحنی ROC مدل‌های رگرسیون لجستیک، الگوریتم CART، شبکه عصبی پرسپترون و RBF مورد مقایسه قرار گرفت و نتایج نشان داد که شبکه عصبی مصنوعی و رگرسیون لجستیک دارای سطح زیر منحنی ROC بیشتری نسبت به دیگر مدل‌ها بود و نیز الگوریتم CART سطح زیر منحنی ROC کمتری نسبت به شبکه عصبی پرسپترون و رگرسیون لجستیک داشت (۲۶).

در مطالعه حاضر اگرچه مدل درخت تصمیم با الگوریتم CART نسبت به دیگر مدل‌ها صحت کمتری داشت اما حساسیت این مدل بیشتر از حساسیت مدل رگرسیون لجستیک به دست آمد. یعنی مدل درخت با الگوریتم CART نسبت به مدل رگرسیون لجستیک دارای توان بیشتری در تشخیص درست افرادی است که واقعاً مبتلا به دیابت بارداری هستند همچنین در مطالعه‌ای (۱۳) که به‌منظور پیش‌بینی بیماری قلبی انجام‌شده است، حساسیت برای مدل درخت تصمیم ۸۱,۷ و برای مدل رگرسیون لجستیک ۸۱,۲ درصد به دست آمد که حساسیت بیشتر مدل درخت تصمیم را نسبت به مدل رگرسیون لجستیک نشان می‌دهد و با نتیجه مطالعه ما همخوانی دارد.

از محدودیت‌های مطالعه حاضر، ناقص بودن اطلاعات و گمشدگی زیاد در پرونده‌های در دسترس بود که باعث شد بسیاری از پرونده‌ها به دلیل داشتن داده گمشده، مورداستفاده قرار نگیرند و حجم نمونه موردبررسی کاهش یابد.

نتیجه‌گیری

در پیش‌بینی و رده‌بندی ابتلا و عدم ابتلا به دیابت بارداری، مدل شبکه عصبی مصنوعی دارای نرخ دسته‌بندی اشتباه کمتر و

سطح زیر منحنی ROC بیشتری نسبت به سایر مدل‌ها بود. می‌توان نتیجه گرفت که این مدل در پیش‌بینی زودرس دیابت بارداری دارای پیش‌بینی‌های صحیح‌تر و نزدیک به واقعیت نسبت به سایر مدل‌های مورداستفاده در این مطالعه است.

با توجه به این‌که مطالعات گذشته در زمینه دیابت بارداری که مدل‌های آماری را مقایسه کرده باشند بسیار اندک است لذا پیشنهاد می‌شود در صورت در دسترس داشتن حجم نمونه بیشتر، مدل‌های آماری دیگری همچون جنگل‌های تصادفی و ماشین بردار پشتیبان نیز جهت مدل‌سازی ابتلا به دیابت بارداری بررسی‌شده و با مدل‌های دیگر ازجمله مدل‌های مورداستفاده در مطالعه حاضر مقایسه شوند.

تشکر و قدردانی

این مطالعه از پایان‌نامه دانشجوی کارشناسی ارشد آمار زیستی، دانشکده بهداشت، دانشگاه علوم پزشکی کرمانشاه با عنوان "بررسی مقایسه‌ای مدل‌های رگرسیون لجستیک، تحلیل ممیزی، الگوریتم CART و شبکه عصبی مصنوعی در پیش‌بینی دیابت بارداری" و شماره ۹۷۱۹۸ در سال ۱۳۹۷ استخراج‌شده است. از معاونت تحقیقات و فناوری دانشگاه جهت حمایت مالی سپاسگزاری می‌کنیم.

منابع

1. Seshadri R. American diabetes association gestational diabetes mellitus. Diabetes Care. 2002; 25: S94-S6.
2. Jiménez-Moleón JJ, Bueno-Cavanillas A, Luna-del-Castillo JD, Garcia-Martin M, Lardelli-Claret P, Gálvez-Vargas R. Prevalence of gestational diabetes mellitus: variations related to screening strategy used. European journal of endocrinology. 2002; 146: 831-7.
3. Patil S, Pandey PD, Patange R. Gestational Diabetes Mellitus Diagnosed with 2hr 75g-Oral Glucose Tolerance Test (DIPSI) and Its Adverse Perinatal Outcome. International Journal of Recent Trends in Science and Technology. 2014; 1: 323-30.
4. Engelgau MM, Herman WH, Smith PJ, German RR, Aubert RE. The epidemiology of diabetes and pregnancy in the US, 1988. Diabetes care. 1995; 18: 1029-33.
5. Sayehmiri F, Bakhtiyari S, Darvishi P, Sayehmiri K. Prevalence of Gestational Diabetes Mellitus in Iran: A Systematic Review and Meta-Analysis Study. The Iranian Journal Of Obstetrics, Gynecology And Infertility. 2013; 15: 16-23.
6. Manshori A, Rezaeian M, Bagheri H, Aminzadeh F, Goujani R. Assessment of the appropriate cut-off point in glucose challenge test based on the risk of gestational diabetes in pregnant women. The Iranian Journal Of Obstetrics, Gynecology And Infertility. 2015; 18: 1-8.
7. Bellamy L, Casas J-P, Hingorani AD, Williams D. Type 2 diabetes mellitus after gestational diabetes: a systematic review and meta-analysis. The Lancet. 2009; 373: 1773-9.
8. Kim C, Newton KM, Knopp RH. Gestational diabetes and the incidence of type 2 diabetes. Diabetes care. 2002; 25: 1862-8.
9. Kashi Z, Borzouei S, Akhi O, Moslemi Zadeh N, Zakeri H, Mohammadpour Tahmtan R, et al. Diagnostic value of fasting plasma glucose (FPG) in screening of gestational diabetes mellitus. Iranian Journal of Diabetes and Metabolism. 2006; 6: 67-72.
10. Gavin III JR, Alberti K, Davidson MB, DeFronzo RA. Report of the expert committee on the diagnosis and classification of diabetes mellitus. Diabetes care. 1997; 20: 1183.
11. Font-López KC, Marcial-Santiago AdR, Becerril-Cabrera JI. Validity of blood glucose fasting test as diagnostic for gestational diabetes during the first trimester of pregnancy. Ginecología y Obstetricia de México. 2018; 86: 233-8.
12. Metzger BE, Coustan DR, Committee O. Summary and recommendations of the fourth international workshop-conference on gestational diabetes mellitus. Diabetes care. 1998; 21: B161.
13. Khemphila A, Boonjing V, editors. Comparing performances of logistic regression, decision trees, and neural networks for classifying heart disease patients. Computer Information Systems and Industrial Management Applications (CISIM), 2010 International Conference on; 2010: IEEE.
14. Gharloghi S, Heidarpour S, Rezaei M. The relationship between hemoglobin concentration in the first trimester of pregnancy and gestational diabetes mellitus(persian). 2014.
15. Rahimi M, Dinari Z, Najafi F. Prevalence of gestational diabetes and its risk factors in Kermanshah 2009. Behbood. 2010; 14: 244-50.
16. Menard S. Applied logistic regression analysis: Sage; 2002.
17. Hosmer DW, Lemeshow S. Special topics. Applied Logistic Regression, Second Edition. 2000: 260-351.
18. Makian S, Almodaresi S, Karimi TS. A Comparison among Artificial Neural Network, Discriminant Analysis and Logestic Regression Techniques for Bankruptcy: A Case Study of Kerman's Firms. The Economic Research. 2010; 10: 141-61.
19. Johnson RA, Wichern DW. Applied multivariate statistical analysis: Prentice hall Upper Saddle River, NJ; 2002.
20. Steinberg D, Colla P. CART: classification and regression trees. The top ten algorithms in data mining. 2009; 9: 179.
21. Timofeev R. Classification and regression trees (CART) theory and applications: Humboldt University, Berlin; 2004.
22. Shafer J, Agrawal R, Mehta M, editors. SPRINT: A scalable parallel classi er for data mining. Proc 1996 Int Conf Very Large Data Bases; 1996.
23. Harper PR. A review and comparison of classification algorithms for medical decision making. Health Policy. 2005; 71: 315-31.
24. Hagan M. Neural network design, pws, USA; 1995.
25. Hosseini SM, Tazhibi M, Amini M, Zaree A, Hashemi HJ. Using Classification Tree for prediction of Diabetic Retinopathy on Type II Diabetes. Journal of Isfahan Medical School. 2010; 28: 15-24.
26. Kurt I, Ture M, Kurum AT. Comparing performances of logistic regression, classification and regression tree, and neural networks for predicting coronary artery disease. Expert systems with applications. 2008; 34: 366-74.

Iranian Journal of Epidemiology Winter 2019; 15 (4): 362-370.
Original Article

Comparison of Gestational Diabetes Prediction Between Logistic Regression, Discriminant Analysis, Decision Tree and Artificial Neural Network Models

Rezaei M[1], Fakhri N[2], Shahsavari S[3], Rajati F[4]
1- Professor of Biostatistics, Fertility and Infertility Research Center, Kermanshah University of Medical Sciences, Kermanshah, Iran
2- MSc of Biostatistics, Faculty of Public Health, Kermanshah University of Medical Sciences, Kermanshah, Iran
3- Assistant Professor of Biostatistics, Faculty of Par Medicine, Kermanshah University of Medical Sciences, Kermanshah, Iran
4- Associate Professor of Health Education , Research Center for Environmental Determinants of Health, Kermanshah University of Medical Sciences, Kermanshah, Iran
Corresponding author: Fakhri N, n.fakhri94@yahoo.com
(Received 11 April 2019**; Accepted** 31 August 2019**)**

Background and Objectives: Gestational Diabetes Mellitus (GDM) is the most common metabolic disorder in pregnancy. In case of early detection, some of its complications can be prevented. The aim of this study was to investigate early prediction of GDM by logistic regression (LR), discriminant analysis (DA), decision tree (DT) and perceptron artificial neural network (ANN) and to compare these models.

Methods: The medical files of 420 pregnant women (2010-12) in Kermanshah health centers were evaluated using convenience sampling. Demographic data, pregnancy-related variables, lab tests results, and a diagnosis of GDM according to a fasting blood sugar level of 92 or more were collected from their files. After fitting the four models, the performance of the models was compared and according to the criteria of accuracy, sensitivity and specificity (based on the ROC curve), the superior model was introduced.

Results: Following the fitting of LR, DA, DT and perceptron ANN models, the following results were obtained. The accuracy of the above models was 0.81, 0.83, 0.78 and 0.83, respectively, the sensitivity of the models was 0.50, 0.63, 0.58 and 0.58, the specificity of the models was 0.96, 0.93, 0.87 and 0.94, and the area under the ROC curve was 0.86, 0.78, 0.73 and 0.87, respectively.

Conclusion: In predicting and categorizing the presence of GDM, the ANN model had a lower error rate and a higher area under the ROC curve compared to other models. It can be concluded that this model offers better predictions and is closer to reality than other models.

Keywords: GDM, Accuracy, Sensitivity, Specificity, ROC curve