

مقایسه عملکرد روش‌های شبکه عصبی مصنوعی پرسپترون چندلایه مبتنی بر الگوریتم wrapper، تحلیل ممیزی و رگرسیون لجستیک در تعیین عوامل خطر دیابت نوع ۲

اقبال زندکریمی^۱ (Ph.D Student)، مجید صادقی فر^{۲*} (Ph.D)، منصور رضایی^۳ (Ph.D)

۱- دانشگاه علوم پزشکی کردستان، سنندج، ایران

۲- گروه آمار، دانشگاه بوعلی سینا همدان، ایران

۳- گروه آمار زیستی و اپیدمیولوژی، مرکز تحقیقات توسعه اجتماعی و ارتقای سلامت، دانشگاه علوم پزشکی کرمانشاه، کرمانشاه، ایران

چکیده

هدف: در این مطالعه عملکرد پیش‌بینی سه مدل آماری جهت تعیین ریسک فاکتورهای دیابت مقایسه گردید. مواد و روش‌ها: شاخص توده بدن (BMI)، قندخون ناشتا (FBS)، فشارخون (HT)، چربی‌های خون (TG, TC)، HDL و LDL (HbA1C)، وزن و سابقه سیگار کشیدن از پرونده درمانی افراد تحت بررسی گردآوری شد. مدل‌های شبکه عصبی مصنوعی پرسپترون چندلایه (MLP) و تحلیل ممیزی (DA) رگرسیون لجستیک (LR) به منظور شناسایی ریسک فاکتورها بر داده‌ها برازش و از منحنی راک جهت مقایسه قدرت پیش‌بینی مدل‌ها استفاده شد. به‌منظور رفع مشکل بیش‌برازش (Overfitting) در مدل MLP از الگوریتم Wrapper استفاده شد.

یافته‌ها: قدرت پیش‌بینی سه مدل MLP، DA و LR بر اساس سطح زیر منحنی راک به ترتیب برابر ۰/۹۸۴ و ۰/۹۸۱ و ۰/۹۸۳ برآورد گردید. متغیرهای FBS ($P < 0/0001$) و HbA1C ($P < 0/0001$)، وزن ($P < 0/001$)، BMI ($P < 0/01$) و LDL ($P < 0/003$) در مدل LR و متغیرهای FBS، HbA1c، SIGAR، HT با $P < 0/0001$ در مدل DA و متغیرهای سن، FBS، TG، BMI، HT و TC با توجه به روش راپر، در مدل MLP معنادار بودند. مدل MLP (۹۷٪) حساسیت بالاتری را نسبت به LR (۹۴٪) و DA (۹۲٪) نشان داد. همچنین مدل MLP (۹۷٪) ویژگی بالاتری نسبت به LR (۹۲٪) و DA (۹۳/۳٪) را داشت.

نتیجه‌گیری: با توجه به یافته‌های این مطالعه سه روش LR، DA و MLP، جهت کشف تفاوت‌ها تقریباً مشابه بودند. بنابراین پیشنهاد می‌شود در مواردی که نیاز به تفسیر ساده اثر متغیرها وجود دارد از روش LR که معیاری مانند OR را فراهم می‌نماید، استفاده شود. این در حالی است که روش MLP مانند یک جعبه سیاه عمل می‌کند که نوع رابطه بین متغیرها در آن نامشخص است. همچنین پیشنهاد می‌شود که عملکرد روش‌های فوق با استفاده مجموعه داده‌های دیگر نیز با یکدیگر مقایسه شود و مطالعات بیشتری انجام شود.

واژه‌های کلیدی: شبکه عصبی مصنوعی پرسپترون چندلایه، رگرسیون لجستیک، تحلیل ممیزی، Wrapper،

منحنی راک

کلاسیک هستند و هیچ یک از این روش‌ها قابلیت مدل‌بندی روابط پیچیده غیرخطی و اثرات متقابل درجه بالا را ندارند. شبکه‌های عصبی مصنوعی برای تشخیص، طبقه‌بندی و پیش‌بینی که در آن‌ها روابط معمولاً به شکل غیر خطی هستند

مقدمه

در نظر گرفتن یک توزیع پیش‌فرض مانند توزیع نرمال برای متغیرهای پاسخ، خطی بودن رابطه‌ی پیشنهادی، یکسان بودن واریانس خطاها از جمله محدودیت‌های مدل‌های

رگرسیون لجستیک برخوردار بود در این مطالعه مدل شبکه عصبی مصنوعی در تفکیک بیماران دیابتی دارای رتینوپاتی از دیابتی بدون رتینوپاتی قدرت بیشتری نشان داد [۱].

در مطالعه کاظم‌نژاد و همکاران در سال ۲۰۰۸ که به منظور مقایسه شبکه عصبی پرسپترون و رگرسیون لجستیک در بررسی متابولیسم قند بیماران دیابتی انجام گرفت، مدل شبکه عصبی پرسپترون بر اساس سطح زیر منحنی راک از قدرت بیشتری در پیش‌بینی مدل برخوردار بود [۴].

هدف از انجام این مطالعه آن است که اولاً با توجه به نقاط قوت و ضعف هر یک از مدل‌های کلاسیک (رگرسیون لجستیک و تحلیل ممیزی) و مدل نوین (شبکه عصبی مصنوعی پرسپترون چندلایه)، مدل بهتری برای تعیین ریسک فاکتورهای بیماری دیابت ارائه شود. ثانیاً معلوم شود که کدامیک از متغیرهای بررسی شده در ابتلا به بیماری دیابت موثرتر است. همچنین با توجه به این که شبکه عصبی مصنوعی به مدل‌های جعبه سیاه معروفند و همانند مدل‌های کلاسیک ضرابی هم‌چون OR و β را جهت تعیین میزان تاثیر هر یک از متغیرهای مستقل در مدل ارائه نمی‌دهد به بررسی روش انتخاب ویژگی راپر (Wrapper) می‌پردازیم.

مواد و روش‌ها

در این مطالعه مقطعی تحلیلی، جامعه مورد مطالعه را بیماران دیابتی (که به‌طور منظم به مرکز تخصصی دیابت کرمانشاه مراجعه می‌کنند) و همچنین بیماران پره دیابتی (افرادی هستند که در غربالگری‌های انجام شده توسط وزارت بهداشت شاخص‌های مربوطه همانند BMI و سطح قند خون ناشتا و... را دارا بوده و اطلاعات آن‌ها در اختیار مراکز بهداشتی است) تشکیل می‌دهند. حداقل نمونه مورد نیاز با سطح اطمینان ۹۵٪ و توان آماری ۹۰٪ و بر اساس اطلاعات به‌دست آمده از مطالعات گذشته [۱] و با استفاده از فرمول حجم نمونه کوکران ۱۲۰ نفر برای هر گروه تعیین شد. ولی به منظور جلوگیری از کاهش توان مطالعه در برخورد با داده‌های گمشده احتمالی این مقدار برای هر گروه ۱۵۰ نفر در نظر

مورد استفاده قرار می‌گیرد. از طرفی مدل شبکه‌های عصبی مصنوعی به دلیل ارابه ندادن ضرابی همانند نسبت بخت‌ها به مدل‌های جعبه سیاه (Black box) معروفند و از این‌رو مدل‌های کلاسیک (رگرسیون لجستیک و تحلیل ممیزی) به دلیل استفاده و تحلیل آسان و همچنین ارائه ضرابی چون نسبت بخت، برای بیان میزان تاثیر هر متغیر مستقل در مدل، کاربرد زیادی در مدل‌های علوم پزشکی دارند [۱]. از طرفی در تحقیقات پزشکی و اپیدمیولوژی به خاطر در میان بودن مسئله سلامت انسان، دقت در پیش‌بینی صحیح رخ داده‌ها اهمیت بیشتری می‌یابد. لذا استفاده از روش‌های مدل‌سازی که پیش‌بینی بر اساس آن‌ها دارای حداقل خطا و بیش‌ترین اطمینان باشد ضروری به نظر می‌رسد [۲].

بیماری دیابت شایع‌ترین بیماری غدد درون‌ریز است که به علت اختلالات متابولیسم قند، چربی و پروتئین رخ می‌دهد و مشخصه آن افزایش قند خون می‌باشد کمیته بین‌المللی تشخیص و طبقه‌بندی دیابت پیشنهاد می‌کند که افراد چاق با بیش‌تر از ۱۲۰٪ وزن ایده‌ال یا شاخص توده بدن (BMI) بیش‌تر یا مساوی ۲۷ کیلوگرم بر متر مربع، افرادی که اقوام درجه یک آنان دیابتی می‌باشند، جمعیت‌های دارای سابقه تولد نوزاد با وزن بیش از ۴/۱ کیلوگرم، کسانی که فشارخون بالای ۹۰/۱۴۰ دارند و کسانی که HDL کم‌تر از ۳۵ Kg/dl دارند از نظر خطر ابتلا به دیابت غربال شوند [۳].

در مطالعه‌ای تحت عنوان مقایسه قدرت شبکه عصبی مصنوعی با رگرسیون لجستیک در تفکیک بیماران دیابتی رتینوپاتی از غیر رتینوپاتی که توسط زندکریمی و همکاران با استفاده از داده‌های بیماران دیابتی مرکز تخصصی دیابت کرمانشاه انجام گرفت ۱۵۰ نفر مورد و ۱۵۰ نفر کنترل وارد مطالعه شدند. اطلاعات دموگرافیک، فشارخون (HT)، کلسترول خون (TC، Hba1c، FBS، BMI) و مدت زمان ابتلا، وضعیت سیگاری بودن و سن بیماران مورد بررسی قرار گرفت با توجه به یافته‌ها قدرت پیش‌بینی رگرسیون لجستیک و MLP به ترتیب ۰/۷۳ و ۰/۸۳ برآورد شد و مدل شبکه عصبی مصنوعی از حساسیت و ویژگی بالاتری نسبت به

چند لایه از روش Wrapper استفاده شد با توجه به این‌که در تکنیک Wrapper جهت انتخاب ویژگی‌ها، کلیه حالت‌هایی که متغیرهای مستقل وارد مدل می‌شود ($n2$ که در آن n تعداد متغیرهای مستقل است) در نظر گرفته می‌شود [۱] و با توجه به این‌که در این مطالعه تعداد یازده متغیر سن، HT، انواع چربی‌های خون (TC, LDL, HDL, TG), BMI, وزن، وضعیت سیگاری بودن، HbA1C و FBS را در نظر گرفته بودیم، تعداد ۲۰۴۸ حالت ممکن کاندید در مدل شبکه عصبی مصنوعی در نظر گرفته شد که از میان آن‌ها مدل با بیش‌ترین قدرت پیش‌بینی (TCR) به عنوان مدل برتر انتخاب شد. ۷۰٪ از مشاهدات به مرحله آموزش و ۳۰٪ به مرحله آزمون اختصاص داده شد. هم‌چنین به منظور مقایسه‌ی قدرت پیش‌بینی سه مدل از سطح زیر منحنی راک و مقادیر حساسیت و ویژگی استفاده شد.

مدل‌های پیشگویی

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

رگرسیون لجستیک دارای شکل کلی زیر است:

$$\text{Log}\left(\frac{\pi}{1-\pi}\right) = \alpha + \sum \beta x$$

حرف X نشان متغیر مستقلی است که در این جا ۱۱ متغیر فوق را شامل می‌شود و β ضریب برآورد شده مدل برای متغیر مستقل است [۵]. هم‌چنین π احتمال ابتلا به دیابت در مدل است که اگر فرد پره دیابتی مبتلا به دیابت شده باشد عدد ۱ می‌گیرد و اگر مبتلا به دیابت نشده باشد عدد صفر می‌گیرد. اگر هر کدام از ریسک فاکتورها در مقایسه دو گروه (مورد و شاهد) معنی‌دار باشد در رابطه بالا با یک ضریب β معنی‌دار وارد می‌شود.

تحلیل ممیزی: تحلیل ممیزی دارای تابعی به صورت زیر است:

$$Y = b_0 + b_1x_1 + b_2x_2 + \dots + b_kx_k$$

که در آن مقادیر b_1, b_2, \dots مقدار ضریب تشخیصی یا وزن هر متغیر را نشان می‌دهد و معادل β در رگرسیون خطی است. تحلیل ممیزی یک تکنیک چندمتغیره است که با جدا کردن

گرفته شد و ۱۵۰ نفر نمونه گروه مورد (افراد دیابتی) با استفاده از روش نمونه‌گیری تصادفی ساده از میان ۱۵۰۰۰ فرد دارای پرونده در مرکز کیلینیک تخصصی دیابت کرمانشاه انتخاب گردید و ۱۵۰ نفر نمونه گروه کنترل (افراد پره دیابتی) نیز از میان افراد شناسایی شده در غربالگری‌های انجام شده توسط مراکز بهداشتی درمانی انتخاب گردید. از این تعداد ۶۰٪ زن و ۴۰٪ مرد بودند. اطلاعات دموگرافیک و اطلاعاتی مانند BMI، قند خون ناشتا (FBS)، فشارخون (HT)، چربی‌های خون (TC, TG, HDL, LDL HbA1C) وزن و سابقه سیگار کشیدن دو گروه در دو فرم جداگانه از روی پرونده‌ها تکمیل گردید. به منظور کاهش خطا در جمع‌آوری اطلاعات، آموزش‌های لازم جهت تکمیل فرم‌ها به تکمیل‌کنندگان فرم داده شد.

جهت انجام تجزیه و تحلیل مدل شبکه عصبی مصنوعی پرسپترون چندلایه از نرم‌افزار MATLAB - ۲۰۰۹ و جهت تجزیه و تحلیل مدل رگرسیون لجستیک و تحلیل ممیزی از نرم‌افزار SPSS - ۱۶ استفاده گردید.

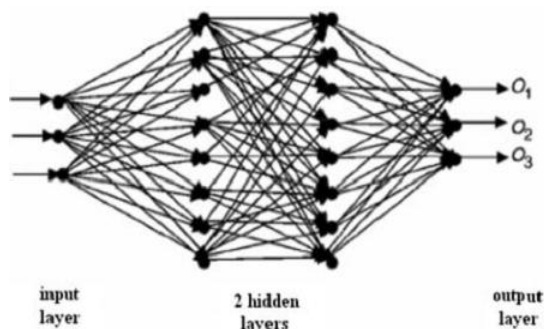
ابتدا به منظور دستیابی به یک دید کلی از مشاهدات از روش‌های آمار توصیفی و نمودارهای آن استفاده گردید. نرمال بودن متغیرهای سن و چربی خون، HDL, LDL, وزن و فشارخون با آزمون کولموگروف-اسمیرنوف بررسی شد. متغیرهای TG و HDL پس از استفاده از تبدیل لگاریتم نرمال شدند.

ابتدا از مدل رگرسیون لجستیک به منظور شناسایی ریسک فاکتورها و ارزیابی دقت مدل در تفکیک بیماران دیابتی از پره دیابتی استفاده شد. در این مدل انتخاب متغیرها به روش Forward LR انجام گرفت. در ادامه از مدل تحلیل ممیزی به منظور کشف عوامل تعیین‌کننده مهم و ارزیابی دقت این روش استفاده شد در این مدل انتخاب متغیرها بر اساس روش لاندای ویلکز انجام گرفت در نهایت از مدل شبکه عصبی مصنوعی پرسپترون چند لایه (MLP) به منظور شناسایی ریسک فاکتورهای دیابتی استفاده شد. جهت رفع مشکل بیش برآزشی مدل شبکه عصبی مصنوعی پرسپترون

یادگیری خاصی صورت می‌گیرد که با تنظیم وزن‌های موجود در ارتباطات بین نورون‌ها، اقدام به آموزش شبکه می‌کنند [۳، ۱].

شبکه عصبی مصنوعی دارای دو جنبه اصلی است:

(۱) معماری (۲) یادگیری



معماری:

شبکه عصبی مصنوعی مورد نظر ما پرسپترون چندلایه

(MLP) است که در مقایسه با سایر روش‌ها بهتر عمل

شکل ۱. ساختار شبکه عصبی مصنوعی

می‌کند [۱]. تصویر شماره ۱ ساختار MLP را با یک لایه ورودی و دو لایه میانی (مخفی) و یک لایه خروجی نشان می‌دهد.

یادگیری: شبکه عصبی MLP از روش یادگیری با ناظر جهت آموزش استفاده می‌کند. در یادگیری با ناظر مجموعه‌ای از زوج داده‌ها به نام نمونه‌های آموزشی به صورت زیر داده می‌شود:

خروجی پرسپترون توسط رابطه زیر مشخص می‌شود:

$$O(x_1, x_2, \dots, x_n) = \begin{cases} 1 & \text{If } (w_0 + w_1x_1 + \dots + w_kx_k) > 0 \\ 0 & \text{If } (w_0 + w_1x_1 + \dots + w_kx_k) \leq 0 \end{cases} - 1$$

که برای سادگی آن را می‌توان به صورت زیر نشان داد:

$$O(X) = \text{sgn}(WX) = \text{Hardlims}(y) = \begin{cases} +1 & \text{if } y > 0 \\ -1 & \text{if } y \leq 0 \end{cases}$$

مجموعه‌های متمایز مشاهده‌ها و با تخصیص مشاهده‌های جدید به دسته‌های از پیش تعریف شده سر و کار دارد. مسأله آماری در مورد به وجود آوردن یک قانون (تابع تشخیص) بر مبنای اندازه‌های حاصل از افراد می‌باشد. با استفاده از این قانون می‌توان افراد جدید را که مشخص نیست به کدام جمعیت تعلق دارند به یکی از جمعیت‌ها منتسب کرد. از معروف‌ترین توابع مورد استفاده در تحلیل ممیزی می‌توان به تابع ممیز فیشر اشاره کرد [۶].

شبکه عصبی مصنوعی پرسپترون چندلایه (MLP): شبکه عصبی مصنوعی برای تشخیص، طبقه‌بندی و پیش‌بینی که در آن‌ها روابط معمولاً به شکل غیرخطی هستند مورد استفاده قرار می‌گیرد [۳]. به طور کلی شبکه‌های عصبی مصنوعی دارای مزایای زیر می‌باشند:

(۱) اجازه وارد نمودن تعداد زیادی از متغیرها در مدل

(۲) عدم نیاز به پیش‌فرض‌هایی چون نرمال بودن و ...

(۳) پیدا کردن الگو علی‌رغم وجود داده‌های گمشده

همچنین این مدل‌ها دارای محدودیت‌هایی نیز هستند:

(۱) بیش برآزش Over fitting [۷]

(۲) ارائه ندادن ضرایبی همچون OR و β همانند مدل‌های

رگرسیون [۸].

روش wrapper نیازمند حجم بالای محاسبات است [۹] مثلاً در این مقاله برای ۱۱ متغیر تعداد $2048 = 2^{11}$ حالت بررسی شد اما علی‌رغم این عیب، این روش در مقابل بیش برآزش Over fitting قدرتمند است. شبکه عصبی مصنوعی پرسپترون نوعی از شبکه عصبی بر مبنای یک واحد محاسباتی به نام پرسپترون ساخته می‌شود. پرسپترون برداری از ورودی‌های با مقادیر حقیقی را گرفته و یک ترکیب خطی از این ورودی‌ها را محاسبه می‌کند. اگر عدد حاصل از یک مقدار آستانه بیش‌تر بود خروجی پرسپترون برابر با ۱ و در غیر این صورت -۱ خواهد بود. شبکه عصبی مصنوعی پرسپترون چندلایه به دلیل قابلیت‌های با ارزش موازی و یادگیری، عمدتاً به منظور حل مسائل پیچیده مورد استفاده قرار می‌گیرد. فرآیند یادگیری در این شبکه‌ها از طریق الگوریتم‌های

$$\text{Logit}(P(y=1)) = 0.026 * \text{LDL} + 3.092 * \text{BMI}(1) + 2.476 * \text{BMI}(2) + 1.17 * \text{weight} + 5.38 * \text{HbA1C} + 3.868 * \text{FBS}(1) + 3.241 * \text{FBS}(2) - 0.171418$$

هر مشاهده‌ی جدید را می‌توان با قرار دادن مقادیر خواسته شده در تابع فوق به یکی از گروه‌های دیابتی یا پره دیابتی منسوب کرد. با برازش مدل رگرسیون لجستیک بر روی مشاهدات، متغیرهای پیشگو طی پنج مرحله وارد مدل شدند. آزمون هاسمر-لمشو با $(p=0/016)$ در مرحله آخر صحت برازش مدل را نشان داد. هم‌چنین آماره $(\text{Nagelkerke } R \text{ square}=0/836)$ نشان داد که مدل تا چه اندازه توانسته تغییرات مربوط به متغیر وابسته (دیابتی-پره دیابتی) را تبیین کند. متغیرهایی که توسط مدل رگرسیون لجستیک به عنوان ریسک فاکتور شناخته شدند شامل، LDL، FBS (با دو سطح)، BMI (با دو سطح) و HbA1C (با یک سطح) بودند (جدول ۲). متغیرهایی که در مدل معنادار نبودند در جدول ۳ آمده است. قدرت مدل در تفکیک بیماران دیابتی از پره دیابتی ۹۳٪ برآورد گردید (جدول ۲).

جدول ۱. مقایسه توزیع فراوانی (درصد) متغیرهای مورد بررسی در دو گروه دیابتی و پره دیابتی

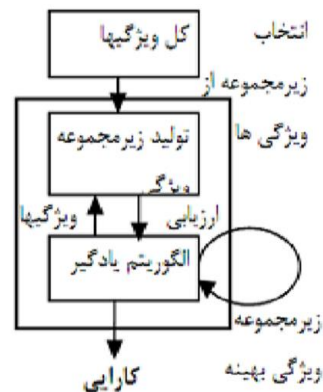
متغیرها	دیابتی تعداد(درصد)	پره دیابتی تعداد(درصد)
جنس		
مرد	(۴۰)۶۰	(۴۰)۶۰
زن	(۶۰)۹۰	(۶۰)۹۰
FBS		
<۱۰۰	(۵)۷	(۲۳)۳۴
۱۰۰-۱۲۵	(۷)۱۰	(۵۷)۸۵
>۱۲۵	(۸۸)۱۳۳	(۲۰)۳۱
HbA1C		
۴-۶	(۸)۱۲	(۹۲)۱۳۹
>۶	(۹۳)۱۳۸	(۷)۱۱
BMI		
<۲۵	(۲۱)۳۲	(۲۳)۳۵
۲۵-۳۰	(۴۴)۶۶	(۵۰)۷۵
>۳۰	(۳۵)۵۲	(۲۷)۴۰
سیگاری		
بله	(۱۵)۲۲	(۲)۴
خیر	(۸۵)۱۲۸	(۹۸)۱۴۶
HT		
<۱۲۰	(۲۹)۴۴	(۴۷)۷۰
۱۲۰-۱۴۰	(۳۳)۴۹	(۳۰)۴۵
>۱۴۰	(۳۸)۵۷	(۲۳)۳۵

برای تحلیل ممیزی آزمون لاندای ویلکز با مقدار $0/223$ و $(p < 0/001)$ معناداری مدل را نشان داد. مقدار همبستگی کانونی مدل $0/882$ به دست آمده است که دلالت بر همبستگی

یادگیری پرسپترون عبارت است از: پیدا کردن مقادیر درستی برای W بنابراین فضای فرضیه H در یادگیری پرسپترون عبارت است از مجموعه تمام مقادیر حقیقی ممکن برای بردارهای وزن [۱].

روش Wrapper: مدل‌های شبکه عصبی مصنوعی قادر به ارایه ضرایب OR و β همانند مدل‌های کلاسیک نیستند. بنابراین نمی‌دانیم میزان تاثیر هر متغیر در مدل چقدر است؟ هم‌چنین یکی دیگر از محدودیت‌های بیش برازش است که برای رفع این محدودیت از روش Wrapper استفاده می‌شود.

در تصویر شماره ۲ روش کار مشخص شده است



شکل ۲. روش انتخاب متغیر با استفاده از روش راپر

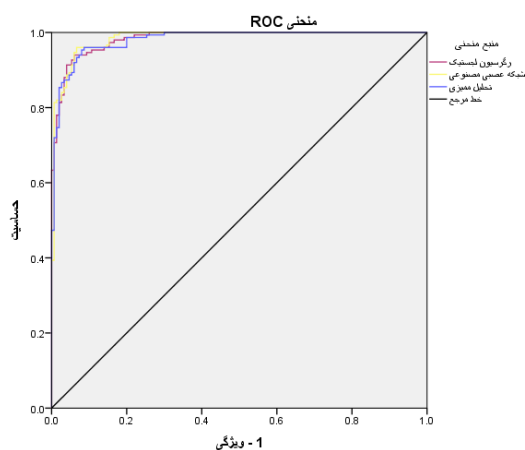
در این روش جهت انتخاب ویژگی از الگوریتم ورودی (induction algorithm) استفاده می‌شود. به این صورت که کلیه حالت‌هایی که متغیرها وارد مدل می‌شوند در نظر گرفته می‌شود. برای n متغیر تعداد 2^n فضای بررسی داریم [۱۰].

نتایج

از ۳۰۰ نمونه تعداد ۱۸۰ نفر (۶۰٪) زن و ۱۲۰ نفر (۴۰٪) مرد بودند. میانگین سنی در گروه پره دیابتی $53/83$ و میانگین سنی در گروه بیماران دیابتی $41/55$ سال بود (جدول ۱). مدل رگرسیون لجستیک برای پنج متغیر پیشگوی وزن، LDL، FBS (با دو سطح)، BMI (با دو سطح) و HbA1C (با یک سطح) به داده‌ها برازش داده شد. سطح اول هر متغیر کیفی به عنوان طبقه مرجع برای تحلیل در نظر گرفته شد:

HbA1C، وزن، TC، LDL (با قدرت ۹۷٪ و مدل ۵،۶،۸،۲)، در حالت شش متغیری سن و HDL و LDL، BMI، HbA1C، و وضعیت سیگاری بودن (با قدرت ۹۷٪ و مدل ۶،۱۰،۶،۲)، در حالت هفت متغیری سن، FBS، TG، HbA1C، وزن، وضعیت سیگاری بودن و HT (با قدرت ۹۷٪ و مدل ۷،۲،۶،۲)، در حالت هشت متغیری سن، FBS، TG، HbA1C، BMI، وزن، HT و TC (با قدرت ۹۷٪ و مدل ۸،۸،۱۰،۲)، در حالت نه متغیری سن، HDL، LDL، FBS، TG، HbA1C، وزن، وضعیت سیگاری بودن و HT (با قدرت ۹۶،۶۷ و مدل ۹،۶،۴،۱)، در حالت ده متغیری سن، HDL، LDL، FBS، TG، HbA1C، BMI، وزن، وضعیت سیگاری بودن و TC (با قدرت ۹۶/۶۷ و مدل ۱۰،۲،۶،۲) و مدل با کل متغیرها قدرت پیشگویی ۹۶/۶۷ و ساختار شبکه ۱۱،۸،۱۰،۱ بود. بنابراین مدل با هشت متغیر از همه قدرت بیش‌تری در پیشگویی دارد (جدول شماره ۶).

هم‌چنین سطح زیر منحنی راک برای مدل شبکه عصبی MLP و رگرسیون لجستیک و تحلیل ممیزی محاسبه شد (شکل ۳). این مقادیر برای سه مدل به ترتیب برابر ۰/۹۸۳، ۰/۹۸۱ و ۰/۹۸۴ به دست آمد. بر اساس این نمودار هر سه مدل از قدرت یکسان پیش‌بینی برخوردار می‌باشند.



شکل ۳. مقایسه سطح زیر منحنی راک در سه مدل

قوی بین نمرات تشخیصی و گروه‌ها (دیابتی-پره دیابتی) دارد و مدل مورد نظر توانسته است ۷۸٪ (۰/۸۸۲۲) از تغییرات متغیر وابسته را تبیین کند.

FBS₁: میزان قند خون ناشتا بین ۱۰۰-۱۲۵

FBS₂: میزان قند خون ناشتا بیش‌تر از ۱۲۵

BMI₂: شاخص توده بدنی بیش از ۳۰

BMI₁: شاخص توده بدنی کم‌تر از ۲۵

Sigar(1): افراد سیگاری

مدل شبکه عصبی MLP بالاترین ارزش تشخیصی را در میان ۲۰۴۸ حالت ممکن برای حالتی که متغیرهای AGE و FBS و TG و HbA1C و BMI و Weight و HT و TC هم‌زمان وارد مدل می‌شوند با مدل ۸،۸،۱۰،۲ (۸ متغیر ورودی با ۸ متغیر در لایه اول میانی و ۱۰ متغیر در لایه دوم میانی و ۲ متغیر در لایه خارجی) و مقدار $TCR=97\%$ را ارایه کرد. سه مدل رگرسیون لجستیک، تحلیل ممیزی و مدل شبکه عصبی مصنوعی پرسپترون چند لایه به ترتیب دارای ۹۲٪، ۹۲٪، ۹۷٪ حساسیت و ۹۴٪، ۹۳/۳٪، ۹۷٪ ویژگی و ۹۳٪، ۹۲/۳٪، ۹۷٪ قدرت پیش‌بینی هستند (جدول شماره ۴).

وقتی تک تک متغیرها را وارد مدل نمودیم قدرت پیش‌بینی هر کدام از متغیرهای سن (۵۷/۳٪)، HDL (۵۸/۶۷٪)، LDL (۶۲/۶۷٪)، FBS (۸۹/۳۳٪)، TG (۶۰٪)، HbA1C (۹۸/۶۷٪)، BMI (۶۰٪)، Weight (۵۶٪)، وضعیت سیگاری بودن (۶۳/۱۶٪)، وضعیت فشار خون (۶۷/۱٪)، TC (۶۱/۸۴٪) قدرت پیش‌بینی مدل را داشتند که در میان آن‌ها HbA1C با قدرت (۹۸/۶۷٪) و مدل (۱،۶،۱۰،۱)، بیش‌ترین قدرت را داشت (جدول شماره ۵). در حالت دو متغیری سن و HbA1C (با قدرت ۹۶،۶۸) و مدل (۲،۱۰،۴،۲) و بر قدرتمندترین مدل بود. در حالت سه متغیری سن، FBS و HbA1C (با قدرت ۹۶،۶۸) و مدل (۳،۱۰،۱۰،۱) در حالت چهار متغیری سن و HbA1C، BMI و وضعیت سیگاری بودن (با قدرت ۹۷٪ و مدل ۴،۲،۶،۲)، در حالت پنج متغیری سن و

جدول ۲. مقایسه ضرایب متغیرهایی که توسط مدل رگرسیون لجستیک به عنوان ریسک فاکتور شناخته شدند

متغیرها	ضرایب مدل	انحراف استاندارد	آماره والد	P-value
LDL	۰/۰۲۶	۰/۰۰۹	۹/۰۱۷	۰/۰۰۳
FBS<۱۰۰	۰	۰	۲۷	<۰/۰۰۰۱
۱۰۰<FBS<۱۲۵	۳/۸۶۸	۰/۹۷۴	۱۵/۷۶۱	<۰/۰۰۰۱
FBS>۱۲۵	۳/۲۴۱	۰/۷۵۶	۱۸/۳۹۳	<۰/۰۰۰۱
BMI<۲۵	۰	۰	۹/۱۳۷	<۰/۰۱
۲۵<BMI<۳۰	۳/۰۹۲	۱/۱۷۷	۶/۹۰۱	<۰/۰۰۹
BMI>۳۰	۲/۴۷۶	۰/۸۳۹	۸/۷۱۸	<۰/۰۰۳
HbA1C	۵/۳۸	۰/۶۶۴	۶۵/۶۵۵	<۰/۰۰۰۱
وزن	۰/۱۱۷	۰/۰۳۷	۱۰/۱۴۳	<۰/۰۰۱
عرض از مبدا	-۱۷/۴۱۸	۳/۸۰۸	۲۰/۹۲۶	<۰/۰۰۰۱

جدول ۳. ضرایب تابع تشخیصی کانونی و میزان همبستگی با تابع تشخیصی در مدل تحلیل ممیزی

متغیر	میزان همبستگی با تابع تشخیصی	نمرات تشخیصی استاندارد نشده	لانداي ويلكز	P-Value
HBA1C	۰/۸۵۲	۳/۴۴۶	۰/۲۸۳	<۰/۰۰۱
FBS	۰/۴۰۱	۰/۷۳۲	۰/۶۴۱	<۰/۰۰۱
Sigar	۰/۱۱۷	۰/۳۷۸	۰/۹۵۵	<۰/۰۰۱
HT	۰/۱۰۶	۰/۰۹۳	۰/۹۶۲	<۰/۰۱
BMI	۰/۰۳۷	۰/۳۴۲	۰/۹۹۵	۰/۲۳۳
Age	۰/۰۳۶	-۰/۰۰۵	۰/۹۹۵	۰/۲۴۱
HDL	۰/۰۳۶	۰/۶۹۲	۰/۹۹۵	۰/۲۴۲
Chol	-۰/۰۲۱	-۰/۰۰۵	۰/۹۹۹	۰/۵۰۸
Weight	-۰/۰۲۰	-۰/۰۳۲	۰/۹۹۹	۰/۵۱
LDL	-۰/۰۲۰	-۰/۰۰۵	۰/۹۹۹	۰/۵۳
TG	۰/۰۰۹	۰/۶۳۰	۱	۰/۷۶۶

جدول ۴. صحت پیش بینی مدلها در تفکیک بیماران دیابتی از پره دیابتی

مدلها	مشاهده شده	پیش بینی شده			درصد
		پره دیابتیک	دیابتی	کل	
رگرسیون لجستیک	پره دیابتیک	۱۴۱	۹	۱۵۰	۹۴
	دیابتی	۱۲	۱۳۸	۱۵۰	۹۲
	کل	۱۵۳	۱۴۷	۳۰۰	۹۳
تحلیل ممیزی	پره دیابتیک	۱۴۰	۱۰	۱۵۰	۹۳/۳
	دیابتی	۱۲	۱۳۸	۱۵۰	۹۲
	کل	۱۵۲	۱۴۸	۳۰۰	۹۲/۳
MLP	پره دیابتیک	۳۶	۱	۳۷	۹۷
	دیابتی	۱	۳۷	۳۸	۹۷
	کل	۳۷	۳۸	۷۵	۹۷

جدول ۵. مقایسه قدرت مدلها در شبکه عصبی مصنوعی پرسپترون چندلایه برای یک منغیر

ساختار شبکه	قدرت پیش بینی	بهترین مدل
۱,۲,۲,۲	۵۷/۳۳	سن
۱,۱۰,۴,۱	۵۸/۶۷	HDL
۱,۴,۸,۲	۶۲/۶۷	LDL
۱,۴,۴,۲	۸۹/۳۳	FBS
۱,۱۰,۲,۲	۶۰	TG
۱,۶,۱۰,۱	۹۸/۶۷	HbA1C

بهترین مدل	قدرت پیش بینی	ساختار شبکه
BMI	۶۰	۱,۴,۶,۲
وزن	۵۶	۱,۶,۴,۲
وضعیت سیگاری بودن	۶۳/۱۶	۱,۶,۲,۲
HT	۶۷/۱	۱,۱۰,۲,۱
TC	۶۱/۸۴	۱,۱۰,۲,۱

جدول ۶. مقایسه قدرت مدلها در شبکه عصبی مصنوعی پرسپترون چندلایه بیش از یک متغیر

تعداد متغیرها	بهترین مدل	قدرت پیش بینی (TCR)	ساختار شبکه
یک متغیره	HbA1C	۹۶/۶۷	۱,۶,۱۰,۱
دو متغیره	سن و HbA1C	۹۶/۶۸	۲,۱۰,۴,۲
سه متغیره	سن، HbA1C و FBS	۹۶/۶۷	۳,۱۰,۱۰,۱
چهار متغیره	سن، HbA1C، BMI و وضعیت سیگاری بودن	۹۷	۴,۲,۶,۲
پنج متغیره	سن، HbA1C، وزن، LDL و TC	۹۷	۵,۶,۸,۲
شش متغیره	سن، HbA1C، LDL، HDL، BMI و وضعیت سیگاری بودن	۹۷	۶,۱۰,۶,۲
هفت متغیره	سن، HbA1C، TG، FBS، وزن، وضعیت سیگاری بودن و HT	۹۷	۷,۲,۶,۲
هشت متغیره	سن، HbA1C، TG، FBS، BMI، وزن، HT و TC	۹۷	۸,۸,۱۰,۲
نه متغیره	سن، HbA1C، TG، FBS، LDL، HDL، BMI، وزن، وضعیت سیگاری بودن، HT	۹۶/۶۷	۹,۶,۴,۱
ده متغیره	سن، HbA1C، TG، FBS، LDL، HDL، BMI، وزن، وضعیت سیگاری بودن و TC	۹۶/۶۷	۱۰,۲,۶,۲
یازده متغیره	کل متغیرها	۹۶/۶۷	۱۱,۸,۱۰,۱

Wrapper، با قدرت پیش بینی ۹۷٪ در شبکه عصبی مصنوعی پرسپترون چندلایه معنادار بودند.

مدل MLP از حساسیت بیش تری برخوردار بود (۹۷٪). هم چنین قدرت تشخیص مدل MLP در شناسایی افراد دیابتی از پره دیابتیک بیش تر از مدل رگرسیون لجستیک بود (۹۷٪). مقایسه سه مدل:

مدل های رگرسیون لجستیک و تحلیل ممیزی به دلیل استفاده و تحلیل آسان و هم چنین ارائه ضرایبی چون نسبت بخت، برای بیان میزان تاثیر هر متغیر مستقل در مدل، کاربرد زیادی در مدل های زیست پزشکی (Biomedicine) دارند، به طوری که در مقالات چاپ شده در مجلات عضو Pubmed تعداد ۴۵۶۴۶ مقاله از رگرسیون لجستیک و ۸۰۱۵ مورد از شبکه عصبی مصنوعی (ANN) استفاده نموده اند [۳، ۱].

با توجه به محدودیت های ذکر شده در شبکه های عصبی مصنوعی که به جعبه سیاه (black box) معروفند و استخراج اطلاعات از آنها (هم چون OR و β) مشکل است، برای رفع محدودیت های ذکر شده برای شبکه عصبی مصنوعی، در این مطالعه از روش Wrapper استفاده شد و کلیه مدل های ممکن

بحث و نتیجه گیری

ریسک فاکتورهای بیماری دیابت مدت هاست در مطالعات و جوامع گوناگون مورد بررسی و تجزیه و تحلیل قرار گرفته است. با این حال تعداد مطالعاتی که این عوامل را در میان افراد پره دیابتیک ارزیابی کند، به دلیل تشابه زیاد میان این گروه از افراد با بیماران دیابتی چندان زیاد و در خور توجه نبوده است. شناسایی این ریسک فاکتورها در جوامعی که رفتاری تقریباً مشابه دارند، نیازمند ابزار آماری دقیق و حساس می باشد. این مطالعه به مقایسه ی کارایی و صحت تشخیص مدل های پرکاربرد و معروف شبکه عصبی MLP و تحلیل ممیزی و رگرسیون لجستیک پرداخت.

در این مطالعه متغیرهای FBS ($p < 0.0001$) و HbA1C ($p < 0.0001$)، وزن ($p < 0.0001$)، BMI ($p < 0.0001$) و LDL ($p < 0.0003$) در مدل رگرسیون لجستیک هم چنین بر اساس آماره لاندا ی و بلکز متغیرهای FBS، HbA1C، SIGAR، HT با $P < 0.0001$ مجوز حضور در مدل را دارند و بقیه متغیرها مجوز حضور در مدل تحلیل ممیزی را نداشتند متغیرهای سن، FBS، TG، HbA1C، BMI، فشارخون، TC با توجه به روش

منحنی راک مدل شبکه عصبی مصنوعی (۰/۷۵۴) کمی بزرگ‌تر از رگرسیون لجستیک (۰/۷۵۲) بود اگرچه از لحاظ آماری تفاوت معناداری بود ولی از نظر پزشکی تفاوت معناداری نداشتند [۱۴].

در این مطالعه مدل رگرسیون لجستیک نشان داد LDL، FBS (با دو سطح)، BMI (با دو سطح) و HbA1c (با یک سطح) می‌تواند ریسک فاکتورهای مهمی برای دیابتی شدن باشند هم‌چنین متغیرهای FBS، HbA1c، SIGAR، HT با $P < 0/0001$ مجوز حضور در مدل تحلیل میزری را یافتند با توجه به یافته‌های این مطالعه مدل MLP، جهت کشف تفاوت‌ها نسبت به دو مدل دیگر برتری معنی‌داری نداشت و هم‌چنین با توجه به آن‌که همانند دو مدل دیگر قادر به ارائه ضرایبی مانند OR جهت تعیین تاثیر هر متغیر معنی‌دار در مدل نیست، جهت انجام مطالعات پزشکی توصیه نمی‌شود بنابراین بایستی به هدف توجه کرد اگر هدف ما برازش بهتر است مدل شبکه عصبی مصنوعی مناسب است ولی اگر مدل با تفسیر ساده مدنظر باشد مدل‌های کلاسیک مانند رگرسیون لجستیک یا تحلیل میزری مناسب است بنابراین توجه به محدودیت‌های بیان شده برای مدل شبکه عصبی مصنوعی پیشنهاد می‌گردد از روش‌های کلاسیک همانند رگرسیون لجستیک یا تحلیل میزری استفاده گردد.

تشکر و قدردانی

مقاله حاضر برگرفته از داده‌های طرح پژوهشی با کد ۹۱۴۰۲ است بدین‌وسیله نویسندگان از معاونت پژوهشی دانشگاه علوم پزشکی کرمانشاه به علت حمایت از این مطالعه قدردانی می‌نمایند.

منابع

- [1] Zandkarimi e ASa. Comparison of artificial neural network predictive power with multiple logistic regressions to determine patients with and without diabetic retinopathy. Razi Med Sci 2014; 21: 79-90. (Persian).
- [2] Mirian N, Sedehi M, Kheiri S, Ahmadi A. Joint prediction of occurrence of heart block and death in patient with myocardial infarction with artificial neural network model. Koomesh 2017; 19: 241-247. (Persian).

در MLP را در نظر گرفته شد و بهترین مدل با بیش‌ترین قدرت پیش‌بینی (۹۹٪) در این مطالعه را که با حضور متغیرهای سن (AGE) و قند خون ناشتا FBS و TG و HbA1c و شاخص توده بدن (BMI) و وزن (Weight) و فشارخون (HT) و TC به‌دست آمد، انتخاب گردید. وقتی تک متغیرها را وارد مدل شد HbA1c با قدرت (۹۸/۶۷٪) و مدل (۱،۶،۱۰،۱)، بیش‌ترین قدرت را داشت. در حالت یک، دو، سه، چهار تا ۱۱ متغیری، قدرت پیشگویی و ساختار شبکه به دست آمد که مدل با هشت متغیر از همه قدرت بیش‌تری در پیشگویی داشت.

شبکه عصبی در مواردی مناسب‌تر است که پیش‌بینی متغیرهای وابسته در اولویت باشد و یا داده‌ها دارای ساختار غیر خطی پیچیده‌ای باشند زمانی که هدف بررسی معنی‌داری تاثیر متغیرهای مستقل بر روی متغیرهای وابسته باشد مدل‌های کلاسیک اولویت دارد [۱۱].

در مطالعه‌ای که در سال ۲۰۱۰ توسط گارسیا و همکاران بر روی ۱۱۵ تصویر شبکه‌ی بیماران دیابتی با هدف ارزیابی چهار مدل شبکه عصبی (MLP، RBF، SVM، MV) و رگرسیون لجستیک در کشف خودکار ضایعه قرمز به منظور تشخیص و غربالگری رتینوپاتی انجام گرفت، مدل RBF با حساسیت ۱۰۰٪، اختصاصیت ۵۶٪ و دقت ۸۳/۰۸٪ بهترین نتایج را به‌دست داد [۱۲].

در مطالعه‌ی لی و همکاران در سال ۲۰۱۲ بر روی ۲۷۴ بیمار دیابتی نوع ۲ به منظور پیش‌بینی نوروپاتی پیش از موعد از مدل‌های درخت تصمیم، پرسپترون چندلایه و رگرسیون لجستیک استفاده شد. مدل شبکه عصبی از دقت تشخیصی بیش‌تری (۸۳٪) برخوردار بود. سطح زیر منحنی راک برای درخت تصمیم (۰/۸۸۶) و رگرسیون لجستیک (۰/۸۸) بیش‌تر از شبکه عصبی (۰/۸۵۴) بود [۱۳].

در مطالعه‌ای تحت عنوان مقایسه رگرسیون لجستیک و شبکه عصبی مصنوعی در پیش‌بینی کم‌رود که توسط پارساییان و همکاران که در سال ۲۰۱۲ بر روی افراد بزرگ‌تر از پانزده سال انجام گرفت نشان دادند که سطح زیر

- [9] Guyon I, Elisseeff A. An introduction to variable and feature selection. *J Machine Learn Res* 2003; 3: 1157-1182.
- [10] Kohavi R, John GH. Wrappers for feature subset selection. *Artific Intell* 1997; 97: 273-324.
- [11] Biglarian A, Hajizadeh E, Kazemnejad A, Zali M. Application of artificial neural network in predicting the survival rate of gastric cancer patients. *Iranian J Public Health* 2011; 40: 80.
- [12] Garcia Ma, Lopez MaI, Alvarez D, Hornero R. Assessment of four neural network based classifiers to automatically detect red lesions in retinal images. *Med Eng Phys* 2010; 32: 1085-1093.
- [13] Li C, Zhi X, Ma J, Cui Z, Zhu Z, Zhang C, Hu L. Performance comparison between Logistic regression, decision trees, and multilayer perceptron in predicting peripheral neuropathy in type 2 diabetes mellitus. *Chin Med J* 2012; 125: 851-857.
- [14] Parsaeian M, Mohammad K, Mahmoudi M, Zeraati H. Comparison of logistic regression and artificial neural network in low back pain prediction: second national health survey. *Iran J Public Health* 2012; 41: 86-92.
- [3] Rezaei M, Zandkarimi E, Hashemian A. Comparison of artificial neural network (rbf), logistic regression and discriminant analysis efficiency in determining risk factors of Type 2 diabetes. *World Appl Sci J* 2013; 11: 1522-1529.
- [4] Kazemnejad A, Batvandi Z, Faradmal J. Comparison of artificial neural network and binary logistic regression for determination of impaired glucose tolerance/diabetes/Comparaison d'un réseau de neurones artificiels et de la régression logistique binaire dans la détermination de l'altération de la tolérance au glucose et du diabète. *East Med Health J* 2010; 16: 615.
- [5] Agresti A, Kateri M. *Categorical data analysis*: Springer; 2011 [book].
- [6] Rencher AC. *Methods of multivariate analysis*: John Wiley & Sons; 2003.
- [7] Haykin SS. *Neural networks: a comprehensive foundation*: Tsinghua University Press; 2001.
- [8] Harrell FE, Lee KL, Mark DB. Tutorial in biostatistics multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat Med* 1996; 15: 361-387.

Comparison of multilayer perceptron neural network based on wrapper algorithm, discriminant analysis and logistic regression in determining risk factors of type 2 diabetes

Eghbal zandkarimi (Ph.D Student)¹, Majid Sadeghifar (Ph.D)^{*2}, Mansour Rezaei (Ph.D)³

1. *Kurdistan University of Medical Sciences, Sanandaj, Iran*

2. *Dept. of Statistics, School of Basic Sciences, Bu-Ali Sina University, Hamadan, Iran*

3. *Dept. of Biostatistics and Epidemiology, Center for Social Development Health Promotion, Kermanshah University of Medical Sciences, Kermanshah, Iran*

(Received: 4 Oct 2016; Accepted: 8 Apr 2017)

Introduction: The present study aimed to evaluate the performance of three statistical models in predicting diabetes type 2 as well as to identify its risk factors.

Materials and Methods: The data related to the potential risk factors of body mass index (BMI), fasting blood sugar (FBS), hypertension (HT), lipids (TC, TG, HDL and LDL), HbA1C and smoking history of 300 patients were extracted from medical records. Artificial neural network multi-layer perceptron (MLP), discriminant analysis (DA) and logistic regression (LR) were applied to identify risk factors. ROC curve was used to compare the performance of the models. To fix the problem "over fitting", the MLP model algorithm was used Wrapper.

Results: The prediction power of the MLP, DA and LR based on the ROC curve were 0.984, 0.981 and 0.983, respectively. In the LR model, variables FBS ($P<0.0001$) and HbA1C ($P<0.0001$), weight ($P<0.001$), BMI ($P<0.01$) and LDL ($P<0.003$) were significant. In the DA model, variables HT, Smoking status, Hba1c, FBS (all, $P<0.0001$) were significant. Age, FBS, TG, HbA1C, BMI, HT and TC were significant in the MLP model. The MLP showed higher sensitivity (97%) compared with the LR (94%) and the DA (92%). Also, the model MLP (97%) exhibited high specificity than the LR (92%) and DA (93.3%).

Conclusion: According to the findings of the present study, the performance of the three used methods of MLP, LR and DA were similar. It is suggested to use the LR where there is a need to simple interpretation as it provides OR for a group relative to the other one while MLP acts like a black box that does not show the relationship between variables. It is also suggested to conduct studies for further investigation of the performance of these methods.

Keywords: Artificial Neural Network MLP, Logistic Regression, Discriminant Analysis, Wrapper, Curve Rock.

* Corresponding author. Tel: +98 9188115417

Sadeghifar@Basu.Ac.Ir