ارائه مدل پیش‌بینی تشخیص عوامل ناباروری؛ با استفاده از الگوریتم‌های داده کاوی

سمیرا درمحمدی[1]/ سمیه علیزاده[2]/ محسن اصغری[3]/ مریم شامی[4]

چکیده

مقدمه: حدود ۱۵-۱۰ درصد از زوجین نابارور هستند. ناباروری علل متفاوتی دارد و تشخیص روش درمان بیماران بر اساس نوع عامل ناباروری آن‌ها انجام می‌شود. در این تحقیق مدلی ارائه شده است که بر اساس ویژگی‌های اولیه و نتایج آزمایشات ساده علل ناباروری افراد را پیش‌بینی می‌کند که می تواند به پزشکان در تشخیص زودهنگام علت ناباروری و تصمیم‌گیری بهینه کمک کند.

روش کار: داده‌های این تحقیق برگرفته از داده‌های ناباروری بیمارستان صارم تهران می‌باشد. در این تحقیق از روش‌های داده‌کاوی استفاده شده است. ابتدا روش خوشه‌بندی k-means و سپس روش‌های دسته‌بندی ماشین بردار پشتیبان (SVM: Support Vector Machine) و شبکه‌های عصبی مصنوعی به منظور پیش‌بینی نوع علل ناباروری، اجرا و نتایج دو الگوریتم دسته‌بندی با هم مقایسه شدند. همچنین برای تحلیل داده‌ها و اجرای الگوریتم‌های بخش مدل، از نرم‌افزار SPSS Clementine 12.0 استفاده شده است.

یافته‌ها: در بخش خوشه‌بندی بر اساس الگوریتم K-means داده‌ها به پنج خوشه تقسیم شدند. در هر گروه یک یا چند علت ناباروری مشاهده شد. در ادامه و با اجرای الگوریتم‌های دسته‌بندی SVM و شبکه عصبی مصنوعی، مشخص شد که الگوریتم SVM با نوع کرنل چندجمله‌ای بالاترین کارایی را به دست آورد.

نتیجه گیری: انجام این تحقیق علاوه بر اینکه منجر به شناخت بهتر ویژگی های بیماران ناباروری شد، می تواند زمینه ای برای انجام تحقیقات آتی باشد. از آنجائی که با تشخیص علل ناباروری افراد قبل از مراحل ثانویه و آزمایشات سنگین، به مقدار قابل توجهی در هزینه و زمان صرفه‌جویی و از اثرات جسمی که بر بیماران می‌گذارد کاسته خواهد شد، می‌توان در مطالعات آینده با استفاده از نتایج این تحقیق سیستمی را جهت اجرای این مدل پیاده‌سازی نمود.

کلید واژه‌ها: ناباروری، مدل، داده کاوی، k-means، ماشین بردار پشتیبان، شبکه‌های عصبی مصنوعی

•وصول مقاله: ۹۲/۱۱/۲۱ •اصلاح نهایی:۹۳/۲/۲٤ •پذیرش نهایی:۹۳/٤/٤

۱.دانشجوی کارشناسی ارشد فناوری اطلاعات، دانشکده مهندسی صنایع، دانشگاه خواجه نصیر الدین طوسی، تهران، ایران، نویسنده مسئول (samira.dormohammadi@ymail.com)

۲. استادیار گروه مهندسی فناوری اطلاعات، دانشکده مهندسی صنایع، دانشگاه خواجه نصیر الدین طوسی، تهران، ایران

۳.کارشناسی ارشد، فناوری اطلاعات، دانشکده مهندسی صنایع، دانشگاه خواجه نصیر الدین طوسی، تهران، ایران

۴. کلینیک، ریاست، بیمارستان صارم، تهران، ایران

مقدمه

دستیابی به اینترنت و تکنیک‌های پیشرفته ذخیره اطلاعات، منابع و سازمان‌ها را با حجم زیادی از اطلاعات روبه‌رو کرده است، بدون اینکه دانش خاصی از آن‌ها دریافت شود. با رشد نمایی حجم اطلاعات تغییر ماهیت داده‌ها و زیاد شدن ابعاد داده (ویژگی‌های هر رکورد) ابزارهای تحلیل داده سنتی قادر به تحلیل داده‌ها نیستند. از طرفی سؤالاتی که در هر سازمان باید به آن‌ها پاسخ داده شود، نیاز به توسعه ابزارهای جدید را بیشتر کرده است. به عنوان راه حل این مسئله، تکنیک‌ها و ابزارهای داده‌کاوی توسعه پیدا کردند [۱].

«داده کاوی فرآیند کشف دانش مطلوب از مقدار بزرگی از داده است که در پایگاه داده، انبار داده و دیگر مخازن اطلاعات ذخیره شده است»[۲]. الگوریتم‌های داده‌کاوی به دو دسته کلی نظارتی و غیرنظارتی و یا پیش‌بینی و توصیفی تقسیم شده است. در الگوریتم‌های پیش‌بینی، هدف پیش‌بینی یک ویژگی خاص بر مبنای ویژگی‌های دیگر است. ویژگی پیش‌بینی شونده متغیر وابسته و بقیه متغیرها متغیر مستقل نامیده می‌شوند، اما در الگوریتم‌های توصیفی هدف استخراج الگو از داده‌ها است که نیاز به تحلیل نتایج دارد. الگوریتم خوشه بندی یکی از زیرمجموعه‌های الگوریتم‌های توصیفی است. در خوشه بندی، داده‌ها به گروه‌هایی به نام خوشه تقسیم می‌شوند به طوری که اعضاء یک خوشه بیشترین شباهت را به هم داشته باشند و اعضاء خوشه‌های مختلف کمترین شباهت را به هم داشته باشند [۱,۲].

از داده‌کاوی در زمینه‌های مختلفی از جمله مدیریت ارتباط با مشتری و تحلیل رفتار مشتریان، پیش‌بینی ارزش سهام، دسته بندی سبد محصول، پزشکی و هر جایی که حجم زیادی از داده موجود است استفاده می‌شود [۱]. یکی از شاخه‌های پزشکی که توجه محققان داده‌کاوی را به خود جلب کرده است، ناباروری می‌باشد. ناباروری عبارت است از «عدم وقوع بارداری تا یک سال پس از اینکه زوجین تصمیم به بچه‌دار شدن می‌گیرند، بدون اینکه از روش‌های پیش‌گیری از بارداری استفاده کنند» [۳]. حدود ۱۵-۱۰ درصد درصد از زوجین نابارور هستند [۴]. دلایل گوناگونی برای ناباروری ذکر شده است از جمله علایمی مانند: فاکتور مردانه، فاکتور لوله‌ای، فاکتور دهانه رحم، علائم تخمک گذاری و علائم نامشخص. بر اساس تشخیص علت ناباروری که طی مراحل نسبتاً طولانی صورت می‌گیرد، یکی از روش‌های کمک باروری برای درمان انتخاب می‌شود. این روش‌ها شامل تلقیح داخل رحمی (IUI: Intra Uterine Insemination)، تلقیح اسپرم داخل سیتوپلاسم تخمک (ICSI: Intra Injection Cytoplasmic Sperm)، لقاح آزمایشگاهی (IVF: In Vitro Fertilization) و غیره است [۳].

در فرایند تشخیص علت ناباروری جلسه اول شامل گرفتن شرح حال در مورد تاریخچه قاعدگی، سابقه حاملگی قبلی، عمل جراحی، ابتلا به بیماری‌های زنان، مصرف داروها و درمان‌های قبلی نازایی می‌باشد. سایر بررسی‌ها شامل آزمایشات هورمونی به منظور بررسی عملکرد تخمدان‌ها است. این آزمایشات شامل تست‌های تیروئیدی، AMH، تستسترون، پرولاکتین، استروژن، LH و FSH می‌باشد. آزمایشات سرمی شامل HBSA، HBSAb، HBCAb، HCVAb، HIV، Rubella، Toxoplasma، CMV و قندخون و انسولین است. در اکثر زوج‌ها تست بعد از مقارمت (PCT: Post Coital Test) به منظور بررسی سلامت نطفه مرد و یا اثر ترشحات دهانه رحم زن بر نطفه‌های مرد انجام می‌شود. انجام یک سونوگرافی واژینال نیز در روزهای ۱۴- ۱۲ جهت بررسی تخمک‌گذاری غالباً هم‌زمان با PCT توصیه می‌گردد. هیستروسالپنگوگرافی (HSG) نیز جهت بررسی باز بودن لوله‌ها و اطلاع از وضعیت حفره رحم توصیه می‌شود. نهایتاً پس از تکمیل مراحل فوق‌الذکر با توجه به نتیجه بررسی‌ها، درمان مورد نظر انتخاب و آماده‌سازی بیمار جهت پروتکل انتخابی آغاز می‌گردد.

با توجه به مطالعات بررسی شده در زمینه داده‌کاوی در ناباروری اشرفی کاخکی و همکاران [۵]، شیانگ گو و همکاران [۶]، اویار و همکاران [۸،۷]، والد [۹]، ونکات و

همکـاران [۱۰]، جوریسـیکا و همکـاران [۱۱] و کـافمن و همکاران [۱۲] به پیش‌بینی نتیجه عمـل کمـک بـاروری IVF پرداختنـد. در بـین ایـن مطالعـات مـدل پیشـنهادی در اشـرفی کـاخکی و همکـاران [۵] بیشـترین صـحت (۹۳ درصـد) را به‌دست آورده است. میلسکی و همکاران بـه دسـته بنـدی داده های بیماران IVF و ICSI پرداختند [۱۳]. مورالز و همکـاران نیز با استفاده از روش دسته بندی بیزین بـه پـیش‌بینـی بهتـرین جنین آزمایشگاهی پرداختند[۱۴]. در مطالعات دیگر، میکاس و همکـاران بـه بررسـی یکـی از علـل نابـاروری (مشـکل آزوسپرمیا که یکی از انواع مشکل در علت فـاکتور مردانـه در ناباروری است) [۱۵] و زروسکی و همکاران به مشکل حـذف کروموزومی Y پرداخته اسـت [۱۶] و دیگرعوامـل نابـاروری در هیچ کدام از مطالعاتی که در زمینه داده‌کاوی در نابـاروری مورد مطالعه قرار گرفته، بررسی نشده است. بنابر وجود چنـین خلائی، در این تحقیق قرار شد تـا بـا اسـتفاده از ویژگـی‌هـای اولیه و نتایج آزمایشات سـاده بیمـاران نابـاروری، مـدلی ارائـه شود که ابتدا این بیماران را به گروه‌های مختلف تقسیم نموده و سپس بر اساس یک روش پیش‌بینـی، دسـته بیمـار جدیـد را تشخیص داده و علل ناباروری احتمالی را برای آن پـیش‌بینـی نماید. در صورتی که بتوان پیش‌بینی نمود کـه علـت نابـاروری افراد چیست، می‌توان با حذف یک مرحله و یا ترتیب درست انجام آزمایشات، در هزینه و زمان برای بیمـار صـرفه‌جـویی و فرایند درمان بیماران را بهینه نمود. همچنین نتایج ایـن تحقیـق می‌تواند در فرایند تصمیم‌گیـری در مـورد نـوع روش درمـانی که بر اساس نوع علت ناباروری صورت می گیـرد بـه پزشـک کمک نماید.

روش کار

مطالعه حاضر داده‌محور می باشد که در آن تحقیقی با استفاده از روش‌های داده‌کاوی بـر روی داده‌هـای نابـاروری صـورت گرفته است. مراحل مختلف اجرای تحقیـق حاضـر در شـکل (۱) نشان داده شده است.

شکل ۱: مراحل اجرای تحقیق حاضر

الف) شناخت سیستم

در این مرحله بـه شـناخت زمینـه مـورد مطالعـه پرداختـه شـده است. مواردی همچون شـناخت داده‌هـا، هـدف از انجـام ایـن تحقیق، نیازمنـدی‌هـای مـورد نیـاز محـیط تحقیـق، مشـکلات موجود در سیستم و موارد دیگر در این مرحله جای می گیرد.

ب) شناخت و جمع‌آوری داده‌ها

مجموعـه داده ایـن تحقیـق شـامل ۶۴۶ رکـورد از اطلاعـات بیماران بیمارستان صارم تهران می باشد که بین سال‌های ۱۳۸۵ تـا ۱۳۹۰ جمـع‌آوری شـده اسـت. در ایـن مرحلـه عملیـات پیش‌پردازش و پاکسازی داده بر روی آن‌ها جهت اسـتفاده در ابـزار تحلیـل صـورت گرفتـه اسـت [۱۸,۱۷]. مرحلـه پـیش پردازش به منظور بهبود داده‌ها انجام مـی شـود [۱] کـه شـامل انجام فرایندهایی از قبیل: تصحیح و یا حذف داده‌هـای بـدون مقدار، تعیین محدوده مجاز و تصحیح مقادیر غیرمجاز، انجـام محاسبات مجدد برای برخی از ویژگـی‌هـا و تبـدیل آن‌هـا بـه ویژگی‌های دیگر است [۱۸,۱۷]. هر رکـورد از پایگـاه داده، ۱۵ ویژگـی دارد. مشخصـات مربـوط بـه هـر متغیـر و شـرح مختصری از آن در جدول (۱) نشان داده شده است.

جدول ۱: متغیرهای تحقیق

نام ویژگی	شرح	نوع	بازه مقادیر
BMI Status	شاخص توده بدنی(تناسب قد و وزن زن)	اسمی	لاغر، نرمال، اضافه وزن و چاق
OF	وجود مشکل تخمک گذاری	اسمی	مثبت، منفی
TF	وجود مشکل لوله‌ای	اسمی	مثبت، منفی
MF	وجود مشکل مردانه	اسمی	مثبت، منفی
Infertility	نوع ناباروری	اسمی	اولیه،ثانویه
Kind Of Protocol	نوع پروتکل	اسمی	Poor ،Long،Short
Embryo No Transform	تعداد جنین منتقل شده	عددی	بازه [۰-۷]
No Of Ovum	تعداد تخمک‌ها	عددی	بازه [۰-۴۶] میلیمتر
Duration	طول مدت ناباروری	عددی	بازه [۱-۳۰] میلیمتر
Female Hormonal Test FSH	میزان هورمون FSH زن	عددی	بازه [۰/۱-۷۶]
Female Hormonal Test LH	میزان هورمون LH زن	عددی	بازه [۰/۱-۱۰۱]
Female Hormonal Test Estradiol	میزان هورمون استروژن زن	عددی	بازه [۰/۶-۶۷۳]
Age	سن زن	عددی	بازه [۲۰-۵۰]
No Follicle	تعداد فولیکول	عددی	بازه [۱-۳۸] میلیمتر
Thickness	ضخامت آندومتر(پوشش داخلی رحم)	عددی	بازه [۴-۱۷] میلیمتر

همانطور که در جدول (۱) ملاحظه نمودید برخی از ویژگی‌ها که اغلب آن‌ها اسمی هستند مقادیر مشخصی دارند.

ج) اجرای روش k-means و محاسبه شاخص دیویس-بولدین

از معروف‌ترین الگوریتم‌های خوشه بندی k-means است. در این الگوریتم ابتدا k مقدار از اشیاء به عنوان مراکز اولیه خوشه‌ها انتخاب می‌شود و سپس فاصله هر شیء با این مراکز خوشه محاسبه می‌شود و هر شیء به خوشه‌ای اختصاص می‌یابد که فاصله آن با مرکز خوشه کمترین باشد. سپس مقدار مرکز جدید خوشه محاسبه می‌شود. این فرآیند تا زمانی که تابع معیار همگرا شود ادامه می‌یابد [۲].

در این تحقیق از نرم‌افزار داده‌کاوی SPSS Clementine 12.0 استفاده شده است. یکی از مهم‌ترین معایب الگوریتم k-means، انتخاب k است که از ابتدا مشخص نیست. ابتدا باید این الگوریتم را برای kهای مختلف محاسبه کنیم و سپس با استفاده از یک شاخص بررسی کنیم که کدام k خوشه بندی بهینه را به ما می‌دهد[۲]. در منابع مختلف شاخص‌های متفاوتی از جمله شاخص دان (Dunn)، دیویس-بولدین (Davies-Bouldin)، سیلهوئت (Silhouette) و غیره تعریف شده است. انتخاب k بهینه در شاخص دیویس-بولدین بر مبنای اصل «کمترین شباهت بین خوشه‌ای و بیشترین شباهت درون خوشه‌ای» استوار است. در این تحقیق شاخص دیویس-بولدین به دلیل گستردگی استفاده به کار گرفته شده است.

در معادلات زیر V مرکز خوشه، DB مقدار نهایی شاخص دیویس بولدین، d نشان دهنده فاصله خوشه‌ها از یکدیگر، S نشان دهنده پراکندگی داخل خوشه، q و t یک عدد صحیح، A_iمجموعه‌ای از رکوردهایی است که در هر مرحله در خوشه i قرار می‌گیرد، $|A_i|$ تعداد عناصر مجموعه A_i و c نشان دهنده تعداد خوشه‌ها در هر مرحله از محاسبه شاخص است[۱۹].

بنابراین داریم:

(معادله ۱) [۱۹]

$$S_{i,q} = \left(\frac{1}{|A_i|}\sum_{x\in A_i} ||x - V_i||_2^q\right)^{1/q}$$

(معادله ۲) [۱۹]

$$d_{ij,s} = \left\{\sum_{s=1}^{p}|v_{si} - v_{sj}|^t\right\}^{1/t} = \left\|\underline{v_i} - \underline{v_j}\right\|_t$$

(معادله ۳) [۱۹]

$$R_{i,qt} = \max_{j\in c, j=1}\left\{\frac{S_{i,q}+S_{j,q}}{d_{ij,s}}\right\}$$

و معادله نهایی شاخص دیویس-بولدین برای تعداد c خوشه به صورت زیر است:

(معادله ۴) [۱۹] $DB(c) = \frac{1}{c}\sum_{i=1}^{c} R_{i,q}$

تعداد خوشه بهینه برابر کمترین مقدار DB (c) است. الگوریتم k-means با مقادیر k، از۳ =k تا ۸=k بر روی مجموعه داده حاصل اجرا شد. بقیه مقادیر الگوریتم، همان مقادیر پیش‌فرض است و تمام ویژگی‌های نمونه‌های مجموعه داده در خوشه‌بندی به عنوان ورودی انتخاب شدند.

د) اجرای روش‌های دسته بندی SVM و شبکه عصبی مصنوعی الگوریتم SVM یکی از الگوریتم‌هایی است که در زیرگروه الگوریتم‌های پیش‌بینی قرار می‌گیرد و مبنای آن از تئوری یادگیری آماری است. SVM با استفاده از یک خط به نام مرز تصمیم،ن مونه‌های کلاس‌های مختلف را از هم جدا می‌کند. این مرز تصمیم بردار پشتیبان نامیده می‌شود. هر مرز تصمیم به دو ابرصفحه (hyper plan) محدود می‌شود که فاصله آن‌ها از مرز تصمیم یکسان است و فاصله بین این دو ابرصفحه حاشیه دسته‌کننده است. هدف دسته‌کننده SVM پیدا کردن یک مرز تصمیم با حاشیه دسته‌کننده حداکثر است. مرز تصمیم می‌تواند خطی و یا غیرخطی باشد. این الگوریتم برای داده‌هایی با ابعاد بالا خوب عمل می‌کند[۱].

یکی دیگر از الگوریتم‌های پیش‌بینی شبکه‌های عصبی مصنوعی است. این الگوریتم قصد دارد برخی از عملکردهای ساده مغز انسان را شبیه‌سازی کند. هرگره در شبکه عصبی مانند یک نرون در مغز انسان است که شبکه اتصال این نرون‌ها وظایف یادگیری پیچیده‌ای را انجام می‌دهند. فرایند تحلیل داده‌ها در شبکه عصبی مانند یک جعبه سیاه است. شبکه‌های عصبی در تخمین و پیش‌بینی مثلاً در تخمین قیمت سهام در ماه بعد، بسیار کاربرد دارد [۲].

پس از تعیین گروه‌های مختلف، روش دسته بندی SVM و شبکه‌های عصبی برای پیش‌بینی گروه مربوط به نمونه جدید اجرا گردید و نتیجه آن‌ها با هم مقایسه شد. در روش‌های دسته بندی، یک فیلد باید به عنوان فیلد خروجی انتخاب شود که در اینجا، خوشه‌ها به عنوان فیلد خروجی انتخاب شد و فیلدهای وضعیت BMI، نوع ناباروری، ضخامت آندومتر، سن، تعداد فولیکول، تست هورمونی LH زن، تست هورمونی FSH زن، تست هورمونی استروژن زن، طول مدت ناباروری و تعداد تخمک به عنوان فیلدهای ورودی در نظر گرفته شدند.

در الگوریتم SVM، در نرم‌افزار SPSS Clementine،کرنل، نوع خط مرز تصمیم را نشان می‌دهد و چهار مقدار مختلف می‌پذیرد و با تغییر مقدار این ویژگی نتایج متفاوت حاصل می‌گردد.این مقادیر شامل خطی (Linear)، چندجمله‌ای (Polynomial)، حلقوی (Sigmoid) و RBF است که همه این چهار روش در بخش دسته بندی این تحقیق اجرا شد و با هم مقایسه گردید. پارامترهای دیگر الگوریتم SVM شامل Regularization Parameter، Precision Regression (epsilon)، RBF gamma، Bias و Degree است که در این تحقیق به ترتیب دارای مقادیر ۱/۲۷،۲۵، ۱،۰/۰،۵/۰،۱ و ۴ هستند. الگوریتم شبکه عصبی مصنوعی نیز با مقادیر پیش‌فرض اجرا گردید.

ه) ارزیابی روش‌های دسته بندی و انتخاب روش بهینه

در این مرحله ارزیابی مدل صورت می‌گیرد. صحت، یکی از معیارهای ارزیابی مدل‌های دسته بندی است که مقدار آن برابر درصد مشاهدات مجموعه آموزشی است که توسط روش مورد استفاده، به درستی دسته بندی شده است. ماتریس اغتشاش (Confusion) یکی از ابزارهای مفید برای ارزیابی عملکرد روش‌های دسته‌بندی است. اگر تعداد دسته‌های موجود m باشد ماتریس، اغتشاش جدولی با اندازه m*m است. اگر i شماره سطر باشد و j شماره ستون باشد عنصر C_{ij} تعداد مشاهداتی از دسته i است که توسط الگوریتم دسته‌بندی j تشخیص داده شده است[۲۰]. معیارهای دیگر برای ارزیابی عملکرد الگوریتم دسته بندی، حساسیت (Sensitivity)، شفافیت (Specificity)، دقت (Precision) و صحت است[۲۱] که این مقادیر در ذیل تعریف شده اند:

$$\text{حساسیت} = \frac{\text{تعداد داده های کلاس مثبت که درست دسته بندی شدند}}{\text{تعداد کل داده های کلاس مثبت}} \quad \text{(معادله ۵)}$$

$$\text{شفافیت} = \frac{\text{تعداد داده های کلاس منفی که درست دسته بندی شدند}}{\text{تعداد کل داده های کلاس منفی}} \quad \text{(معادله ۶)}$$

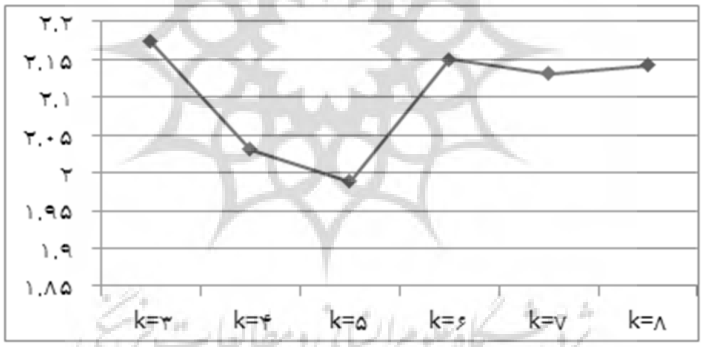
$$\text{دقت} = \frac{\text{تعداد داده های کلاس مثبت که درست دسته بندی شدند}}{\text{تعداد داده های کلاس مثبت که درست دسته بندی شدند+تعداد داده های کلاس منفی که به اشتباه مثبت دسته بندی شدند}} \quad \text{(معادله ۷)}$$

$$\text{صحت} = \text{حساسیت} * \frac{\text{تعداد داده های کلاس مثبت}}{\text{تعداد کل داده ها}} + \text{شفافیت} * \frac{\text{تعداد داده های کلاس منفی}}{\text{تعداد کل داده ها}} \quad \text{(معادله ۸)}$$

در فرمول‌های ۵، ۶، ۷ و ۸ منظور از کلاس مثبت کلاسی است که شاخص برای آن محاسبه می‌شود (یکی از پنج خوشه) و کلاس منفی مابقی کلاس‌ها به جز کلاس انتخاب شده است [۲]. قبل از اجرای دسته‌بندی، داده‌ها به سه دسته داده‌های آموزش (۶۵ درصد) برای ساخت روش دسته بندی و داده‌های تست (۲۵ درصد) برای آزمایش دسته بندی و داده‌های اعتبارسنجی (۱۰ درصد) برای بررسی ارزش و اعتبار تقسیم شدند.

یافته ها

پس از اجرای الگوریتم k-means، تعداد خوشه بهینه محاسبه شد که برابر مینیمُم شاخص دیویس بولدین (شکل ۲) یعنی k=۵ است.



شکل ۲: مقدیر شاخص دیویس بولدین برای مجموعه داده با نتایج منفی

نتایج حاصل از خوشه‌بندی k-means (خوشه‌ها) در جدول (۲) نشان داده شده است. برای متغیرهایی که مقادیر آن‌ها پیوسته و در بازه‌ی مشخصی قرار دارد، میانگین مقادیر و انحراف معیار که نشان دهنده مقدار پراکندگی مقادیر یک متغیر حول مقدار میانگین است، محاسبه شده است.

خوشه یک: تعداد بیماران (۱۱۰ نفر) در این خوشه نسبت به سایر خوشه‌ها کمتر است و بیماران این خوشه پایین‌ترین میانگین سنی (۳۰/۳) و کمترین طول مدت ناباروری (۶/۲) را دارند. ۹۸/۷ درصد از بیماران این خوشه فقط علت ناباروری مردانه دارند و ۱/۳ درصد از آنان هیچ علت شناخته شده‌ای ندارند و جز بیماران با علل ناشناخته هستند.

خوشه دو: بیماران این خوشه بالاترین میانگین سنی (۳۳/۶) را دارند و افرادی که مشکل عامل مردانه نداشتند (حدود ۲۱ درصد) در این خوشه نسبت به همه خوشه‌ها بیشتر است. بیشترین میزان (۲۲/۶ درصد) استفاده از پروتکل Long در این خوشه است. حدود ۶۰ درصد وزن غیرنرمال دارند. همه افراد آن مشکل لوله‌ای (TF) دارند، همچنین حدود ۲۳ درصد علت تخمک‌گذاری دارند. نابراین وجود علت تخمک‌گذاری نسبت به علل دیگر با ابهام بیشتری روبرو است.

جدول ۲: خوشه‌های حاصل از اجرای الگوریتم k-means بر روی داده‌های بیماران با نتایج بارداری منفی

نام ویژگی			خوشه ۱	خوشه ۲	خوشه ۳	خوشه ٤	خوشه٥
تعداد رکورد			۱۱۰	۱۲۴	۱۵۸	۱۲۰	۱۳۴
سن		میانگین	۳۰/۳	۳۳/۶	۳۱/۶	۳۲/۲	۳۱/۸
		انحراف معیار	۵	۵/۱	۶/۶	۵/۶	۶/۱
مدت ناباروری		میانگین	۶/۲	۷/۸	۸	۸/۸	۷/۵
		انحراف معیار	۵	۵/۵	۴/۹	۵/۴	۵/۳
هورمون FSH		میانگین	۴/۵	۵	۵/۹	۴/۷	۵/۸
		انحراف معیار	۲/۶	۳/۵	۶/۶	۴/۷	۴/۲
هورمون LH		میانگین	۷/۵	۷/۲	۱۰/۴	۷/۸	۹/۹
		انحراف معیار	۴/۸	۷/۴	۱۱/۴	۸/۱	۱۲/۶
هورمون Estradiol		میانگین	۸۴/۱	۸۴/۹	۷۷/۹	۱۰۳/۱	۸۶
		انحراف معیار	۶۱/۷	۸۱/۶	۷۷/۶	۹۲/۸	۶۷/۱
تعداد تخمک		میانگین	۱۲/۹	۹/۱	۱۲/۷	۱۰/۶۷	۸/۷
		انحراف معیار	۶/۵	۶/۳	۸/۴۶	۷	۶/۵
تعداد فولیکول		میانگین	۱۰/۶	۷/۷	۱۱/۳	۹/۲	۸/۱
		انحراف معیار	۵	۴/۶	۶	۵/۵	۴/۶
ضخامت آندومتر		میانگین	۹/۱	۹/۲	۹/۰۹	۹/۶۵	۹/۳۲
		انحراف معیار	۱/۶	۱/۸	۱/۵	۱/۸	۱/۹
وضعیت BMI		BMI< ۱۸/۹ لاغر	٪۰/۹	٪۱/۶	٪۰/۶	۰	٪۱/۵
		۱۹<BMI<۲۴/۹ نرمال	٪۸۸/۲	٪۳۸/۷	٪۵۳/۸	۰	٪۸۶/۶
		۲۵<BMI<۲۹/۹ اضافه وزن	۰	٪۴۶	٪۲۹/۱	٪۱۰۰	۰
		۳۰<BMI چاق	٪۱۰/۹	٪۱۳/۷	٪۱۶/۴	۰	٪۱۱/۹
نوع ناباروری		اولیه	٪۸۲/۷	٪۶۲/۹	٪۷۵/۳	٪۸۷/۵	٪۸۵/۸
		ثانویه	٪۱۷/۳	٪۳۷/۱	٪۲۴/۷	٪۱۲/۵	٪۱۴/۲
نوع پروتکل		Short	٪۹۰/۹	٪۷۱	٪۸۵/۴	٪۸۷/۵	٪۷۷/۶
		Long	٪۶/۴	٪۲۲/۶	٪۱۱/۴	٪۶/۷	٪۱۴/۲
		Poor	٪۲/۷	٪۶/۵	٪۳/۲	٪۵/۸	٪۸/۲
تعداد جنین منتقل شده		۰	۰	۰	۰	۰	٪۰/۸
		۱	۰	٪۱۴/۵	٪۳/۸	٪۶/۷	٪۱۴/۹
		۲	۰	٪۱۷/۷	٪۱۵/۸	٪۱۸/۳	٪۳۰/۶
		۳	۰	٪۲۱/۸	٪۲۰/۳	٪۲۰	٪۳۵.۰۷
		۴	٪۱۰۰	٪۳۸/۷	٪۴۶/۸	٪۳۹/۲	۰
		۵	۰	٪۷/۳	٪۱۱/۴	٪۱۴/۲	٪۱۶/۴
		۶	۰	۰	٪۱/۹	٪۱/۷	٪۱/۵
		۷	۰	۰	۰	۰	٪۰/۸
علت ناباروری	OF	مثبت	۰	٪۲۳/۴	٪۱۰۰	۰	۰
	MF	مثبت	٪۹۸/۲	٪۷۹	٪۹۰/۵	٪۹۹/۲	٪۹۹/۲
	TF	مثبت	۰	٪۱۰۰	۰	۰	۰

خوشه سه: تعداد افراد این خوشه (۱۵۸ نفر) نسبت به سایر خوشه‌ها بیشتر است. حدود ۴۷ درصد از بیماران این خوشه وزن غیرنرمال دارند. این تنها خوشه‌ای است که همه بیماران آن مشکل تخمک‌گذاری دارند، اما هیچ یک از آن‌ها مشکل لوله‌های فالوپ ندارند.

خوشه چهار: همه افراد این خوشه اضافه وزن دارند. حدود ۹۹ درصد از آن‌ها فقط مشکل عامل مردانه دارند و بقیه جز بیماران با علل ناباروری ناشناخته هستند و مشکل لوله‌های فالوپ و تخمک‌گذاری ندارند. این بیماران بالاترین میانگین هورمون استروژن را دارند.

خوشه پنج: حدود ۸۶ درصد از افراد این خوشه وزن نرمال دارند و حدود ۹۹ درصد از آن‌ها نیز همانند خوشه (چهار) فقط مشکل مردانه دارند و حدود یک درصد جز افراد با علل ناشناخته هستند و هیچ‌کدام از این افراد مشکل لوله‌ای و تخمک‌گذاری ندارند.

بعد از اجرای الگوریتم‌های دسته بندی صحت الگوریتم SVM برای انواع کرنل و شبکه عصبی در جدول (۳) نشان داده شده است.

جدول ۳: مقدار صحت الگوریتم‌های متفاوت

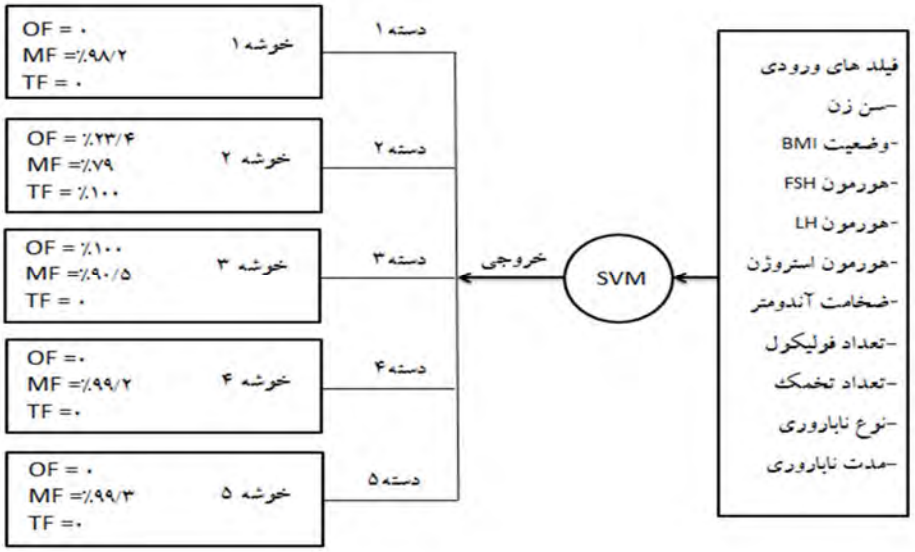
	شبکه عصبی (درصد)	SVM (کرنل چندجمله‌ای) (درصد)
داده‌های آموزشی	۴۱/۸	۷۵/۷
داده‌های تست	۴۸	۷۶/۷
داده‌های اعتبارسنجی	۴۲/۵	۷۴

همانطور که ملاحظه می‌شود، الگوریتم SVM با کرنل چندجمله‌ای به عنوان الگوریتم دسته بندی با بیشترین صحت انتخاب شد. مقادیر معیارهای مختلف میزان صحت برای برچسب کلاس‌های مختلف برای روش بهینه در جدول (۴) نشان داده شده است.

جدول ٤: مقادیر شاخص‌های مختلف برای الگوریتم SVM با تابع کرنل چندجمله‌ای

نام خوشه	حساسیت (درصد)	شفافیت (درصد)	دقت (درصد)	صحت (درصد)
خوشه ۱	۷۴/۵	۷۵/۹	۶۷/۲	۷۵/۷
خوشه ۲	۶۲/۹	۷۸/۷	۸۱/۳	۷۶/۴
خوشه ۳	۷۵/۳	۷۵/۸	۷۴/۸	۷۴/۶
خوشه ۴	۸۹/۲	۷۲/۶	۸۱	۷۶
خوشه ۵	۷۶/۹	۷۵/۴	۷۵/۲	۵۷/۷

برای درک بهتر الگوریتم SVM، عملکرد آن به صورت شماتیک در شکل (۳) نشان داده شده است. با توجه به شکل، الگوریتم SVM بر اساس ویژگی‌های اولیه، بیمار را در یکی از دسته ها که همان خوشه‌های بدست آمده در بخش الگوریتم K-Means هستند قرار می دهد و در هر یک از آن‌ها احتمال وجود عوامل ناباروری مشخص شده است.



شکل ۳: عملکرد الگوریتم SVM

تأثیر ویژگی‌ها در دسته‌بندی متفاوت است. SVM با کرنل چندجمله‌ای ویژگی‌های وضعیت BMI، تعداد تخمک، سن، تست استرادیول، تعداد فولیکول، نوع ناباروری، تست FSH را به ترتیب به عنوان مهم‌ترین ویژگی‌ها در دسته‌بندی معرفی کرد.

بحث و نتیجه گیری

در این تحقیق با استفاده از ترکیب الگوریتم‌های داده‌کاوی، علت ناباروری افراد با استفاده از ویژگی‌های اولیه آن‌ها پیش‌بینی شد. در بخش خوشه بندی، داده‌ها به پنج خوشه تقسیم شدند. علت ناباروری (ویژگی هدف) در هر یک از خوشه‌ها متفاوت بود. لازم به ذکر است که خوشه‌های (یک، چهار و پنج) نتایج مشابهی در علل ناباروری داشتند. در بخش پیش‌بینی، الگوریتم SVM با کرنل چندجمله‌ای بالاترین صحت را به دست آورد. برای داده‌های آموزشی میزان صحت ۷۵/۷ درصد، برای داده‌های تست ۷۶/۷ درصد و برای داده‌های اعتبارسنجی ۷۴ درصد حاصل شد. درصورتی که الگوریتم SVM بیماری با ویژگی خاص را در خوشه‌های (یک، چهار و پنج) قرار دهد، این بیمار فقط مشکل مردانه دارد. اگر در خوشه (دو) قرار دهد با احتمال ۱۰۰ درصد مشکل لوله‌ای دارد، با احتمال ۷۹ درصد مشکل مردانه دارد و با احتمال کم (۲۳ درصد) علت ناباروری بیمار عامل تخمک‌گذاری است. بنابراین در این حالت ابتدا می توان فاکتور لوله‌ای که احتمال وجود آن ۱۰۰ درصد است را بررسی کرد. نهایتاً اگر الگوریتم دسته‌بندی خوشه (سه) را برای نمونه بیمار جدید پیش‌بینی نمود،آنگاه می‌توانیم بگوئیم که این بیمار با احتمال ۱۰۰ درصد مشکل تخمک‌گذاری دارد و با احتمال ۹۷/۲ درصد مشکل مردانه دارد و مشکل لوله‌ای ندارد، بنابراین بررسی عامل تخمک‌گذاری در ابتدا منطقی‌تر به نظر می‌رسد.

با تشخیص زودهنگام علت ناباروری با استفاده از نتایج آزمایشات ساده اولیه می‌توان به روند درمان جهت داده و آن را بهینه نمود. چرا که روش‌هایی که برای تشخیص علل ناباروری بکار گرفته می شود، عوارض زیادی به همراه دارد، بنابراین انجام بیهوده آن برای بیمار عذاب آور است. به عنوان مثال، هیستروسالپنگوگرافی که می تواند برای زن عوارضی از قبیل درد، خونریزی و عوارض ناشی از تاباندن اشعه را داشته باشد [۲۱]. تصور اینکه حذف این مرحله چه اثرات جسمی و مادی مثبتی برای بیمار به همراه خواهد داشت، به دور از ذهن نیست.

طبق بررسی‌های انجام شده در زمینه داده‌کاوی در ناباروری، تحقیق میکاس و همکاران و همچنین زروسکی و همکاران در زمینه عوامل ناباروری است،که به گروه بندی یک نوع از بیماران با عوامل مردانه می‌پردازد. ولی به نظر می‌رسد تاکنون در هیچ مطالعه‌ای مدلی که علل ناباروری افراد را با استفاده از ابزار داده کاوی، پیش‌بینی کند ارائه نشده است و تحقیق حاضر اولین مطالعه در این زمینه است. نتایج حاصل از پیش‌بینی انجام شده در این تحقیق فقط می‌تواند به عنوان کمکی در تصمیم‌گیری استفاده شود و به هیچ عنوان نمی تواند جایگزین پزشک شود. همچنین نتایج این تحقیق وابسته به داده‌های بیمارستان صارم تهران می باشد. برای بررسی بیشتر در این زمینه می‌توان در مطالعات بعدی از داده‌های مراکز درمانی دیگر و الگوریتم‌های دیگر نیز استفاده کرده و نتایج را با هم مقایسه نمود.

تشکر و قدردانی

از زحمات جناب آقای دکتر صارمی و پرسنل محترم بیمارستان که در مراحل انجام این تحقیق همکاری لازم را داشتند،کمال تشکر و قدردانی را داریم.

References

1. Ghazanfari M, Alizadeh S, Teymourpour B. editors. [Data mining and knowledge discovery]. 2nd ed. Tehran: publication of university of science and technology ;2011.[Persian]

2. Han J, Kamber M . Data Mining: Concepts and Techniques. 2nd ed.San Francisco: Morgan Kufman Publisher; 2006.

3. Saremi A. [Introduction to Infertility]. 2nd ed. Tehran: Author; 2004. [Persian]

4. Esmailzadeh S, Farsi M, Bijani A. [Impact of morphology of sperm on the rate of fertility in Intra Uterine Insemination method]. Journal of Reproduction and Infertility 2007; 8(32): 205-212.[Persian]

5. Ashrafi Kakhki S, Malekara B, Rahati Quchani S, Khadem N . A model a Bayesian network for prediction of IVF success rate. 7th international symposium on advances in science and technology. Bandar-Abbas Iran. 2013 March.

6. Shiang Guh R, Chieh Jackson wu T, Ping Weng S. Integrating genetic algorithm and decision tree learning for assistance in predicting in vitro fertilization outcomes . Expert Systems with Applications 2011; 38: 4437-4449.

7. Uyar A, Bener A, Ciray H N, Bahceci M. ROC Based Evaluation and Comparison of Classifiers for IVF Implantation Prediction. Electronic Healthcare 2010; 27: 108-111.

8. Uyar A, Bener A, Ciray H N, Bahceci M. Handling the imbalance problem of IVF

9. implantation prediction. IAENG International Journal of Computer Science 2010 May; 37(2): 164-170.

10.Wald M. Computational models for prediction of IVF/ICSI outcomes with surgically retrieved spermatozoa. Reproductive Bio Medicine Online 2005 July; 11(3): 325-331.

11.Venkat G, Al-Nasser R, Jercovic S, Craft I. Prediction of success in IVF treatments using neural networks. Fertility and Sterility 2004 Sep; 82: s215.

12.Jurisica I, Mylopoulos J, Glasgow J, Shapiro H, Casper R. Case-based reasoning in IVF: prediction and knowledge mining. Artificial Intelligence in Medicine 1998; 12: 1-24.

13.Kaufman S J, Eastaugh JL , Snowden S, Smye SW , Sharma V. The application of neural networks in predicting the outcome of in-vitro fertilization .Human Reproduction 1997; 12(7): 1454–1457

14.Milewski R, Malinowski P, Milewski AJ , Ziniewicz P, Wolczynski S. Classification issue in IVF ICSI/ET data analysis. Studies in logic, grammar and rhetoric 2012; 29(42): 75-85.

15.Morales D A, Bengoetxea E, Larranaga P , Garcia M, Fresnada Y, Merino M. Bayesian classification for the selection of in vitro human embryos using morphological and clinical data .computer methods and programs in biomedicine 2008; 9 0: 104–116.

16.Mikos T, Pantazis K, Goulis D G ,Maglaveras N, Bontis J N , papadimas J. The use of Data Mining in the categorization of patients with Azoospermia. HORMONES 2005 October; 4(4): 214-218.

17.Dzeroski S, Hristovski D, Peterlin B. Using Data Mining and OLAP to Discover Patterns in a Database of Patients with Y-Chromosome Deletions. Journal of the American Medical Informatics Association(AMIA) 2000; 7(Suppl): 215-219.

18.Hoseini M. [Developing a predictive model based on the Sarem hospital infertility data][MSc thesis]. Tehran: K.N. Toosi University of technology; 2012. [Persian]

19.Aghabeigi N. [Providing a hybrid model for clustering on infertility data] [MSc thesis]. Tehran: K.N. Toosi University of technology; 2012. [Persian]

20.Abolmasum F, Alizadeh S, Asghari M. [Utilizing Data Mining Techniques for Investigating Factors Influencing the Failure of Intrauterine Insemination Infertility Treatment]. Journal of Tehran university of medical science 2013; 16(54) : 46-55.[Persian]

21.Ameri H, Alizade S, Barzegari A. [Knowledge Extraction of Diabetics' Data by Decision Tree Method]. Journal of Tehran university of medical science 2013 ; 16(53) : 58-72.[Persian]

22.Saghafi N, Farzaneh S. [Comparison of two methods of hystrosalpyngography and hydrosonohy sterosalpingopraphy in evaluation of the uterine cavity and fallopian tubes in infertile women]. Journal of Iran's women, midwifery and infertility 2003; 6: 35-39.[Persian]

Proposing a prediction model for diagnosing Causes of Infertility by Data Mining Algorithms

Dormohammadi S[1]/ Alizadeh S[2]/ Asghari M[3]/ Shami M[4]

Abstract

Introduction: About 10-15 percent of Iranian couples are infertile which is due to different causes determining particular diagnostic and treatment methods. In this study, the model presented is based on basic features and simple tests, helping physicians predict the causes of infertility

Methods: The data were taken from Sarem hospital infertility data bank by using data mining methods. First, K-means clustering was run; then, support vector machine and artificial neural network classification methods were used to predict the type of infertility, and finally, the results of two classification algorithms were compared. In addition, SPSS Clementine 12.0 was used to analyze the data and implement the algorithm in modeling part.

Results: In k-means clustering, the data were divided into five clusters. In each cluster, one or more causes of infertility were observed. Then, by applying SVM and artificial neural network classification algorithms, the SVM algorithm with a polynomial kernel appeared to have the maximum accuracy.

Conclusion: The findings of this study, could contribute to the understanding of the factors responsible for infertility and pave the way for future investigations. These findings can be used in future studies to develop a system for applying this model since by diagnosing the causes of infertility prior to secondary stages and before performing heavy tests, a considerable amount of time and cost will be saved, and physical burden on patient will be decreased,

Keywords: Infertility, Model, Data Mining, k-means, Support Vector Machine, Artificial Neural Network.

● **Received:** 10/Feb/2014 ● **Modified:** 14/May/2014 ● **Accepted:** 25/June/2014

1. MSc Student of Information Technology, Faculty of Industrial Engineering, K.N Toosi University of Technology, Tehran, Iran; Corresponding Author (samira.dormohammadi@ymail.com)
2. Assistant Professor of Information Technology Department, Faculty of Industrial Engineering, KN Toosi University Of Technology, Tehran, Iran
3. MSc of Information Technology, Faculty of Industrial Engineering, K.N Toosi University of Technology, Tehran, Iran
4. Head of Sarem hospital clinic, Tehran, Iran.