مهندسی صنایع و مدیریت شریف (زمستان ۱۳۹۴)
دوره‌ی ۱ - ۳۱، شماره‌ی ۲/۲، ص. ۱۰۵-۱۱۲، (یادداشت فنی)

Research Note

اجرای الگوریتم خوشه‌بندی بهبود یافته بر روی داده‌های ناباروری

نرگس آقابیگی * **(کارشناسی ارشد)**
سمیه علیزاده (استادیار)
دانشکده‌ی مهندسی صنایع، دانشگاه خواجه نصیرالدین طوسی
ابوطالب صارمی (استاد)
رئیس بیمارستان صارم

داده‌کاوی[1] تلفیقی از روش‌های هوش مصنوعی برای شناسایی اطلاعات یا استخراج دانش از داده‌هاست، به‌نحوی که دانش حاصل در حوزه‌های تصمیم‌گیری، پیش‌بینی، پیش‌گویی و تخمین مورد استفاده قرار گیرد. تحلیل رفتار مشتریان، دسته‌بندی مشتریان، شناخت نیازهای مشتریان و پیش‌بینی در مباحث پزشکی ازجمله کاربردهای داده‌کاوی است. خوشه‌بندی[2] یکی از روش‌های بدون نظارت[3] الگوریتم‌های داده‌کاوی است که به یافتن یک ساختار مشخص درون‌مجموعه‌یی از داده‌های بدون برچسب می‌پردازد. یکی از الگوریتم‌های متداول خوشه‌بندی، الگوریتم k-means است. از معایب این الگوریتم «انتخاب تصادفی خوشه‌های اولیه» در آغاز الگوریتم است که موجب تفاوت نتیجه در هر بار اجرای الگوریتم می‌شود. در این پژوهش، با استفاده از الگوریتم سلسله‌مراتبی[4] مدل جدیدی ارائه شده که می‌کوشد مشکل الگوریتم k-means را برطرف سازد. در ادامه، نتیجه‌ی اجرای این الگوریتم تلفیقی جدید روی داده‌های واقعی مربوط به «ناباروری بیمارستان صارم» ارائه شده است.

واژگان کلیدی: داده کاوی، خوشه‌بندی، ناباروری، روش ICSI، بیمارستان صارم.

n.aghabeigi@sina.kntu.ac.ir
s_alizadeh@kntu.ac.ir
saremat@yahoo.com

۱. مقدمه

اگرچه بیشتر سازمان‌ها به سرعت در حال جمع‌آوری و ذخیره‌سازی داده‌ها هستند، می‌توان ادعا کرد علی‌رغم حجم انبوه این داده‌ها، امروزه سازمان‌ها در تصمیم‌گیری با فقر دانش مواجه‌اند. برای استفاده‌ی بهینه از این داده‌ها در تصمیم‌گیری‌ها، باید آنها را به دانش تبدیل کرد. «داده کاوی» شیوه‌یی است که از آن برای رفع این نیاز استفاده می‌شود. در یک تعریف غیررسمی، داده کاوی فرایندی خودکار برای استخراج الگوهایی است که دانش را بازنمایی می‌کنند. این دانش به‌صورت ضمنی در پایگاه داده‌های عظیم، انبار داده‌ها و دیگرمخازن بزرگ اطلاعات ذخیره شده است. داده کاوی از فنّاوری‌های نوینی است که تحلیل‌های مناسبی برای سیستم‌ها و کسب‌وکارهای مختلف ارائه می‌دهد و با شیوه‌های متفاوت، تفسیر نتایج و تحلیل‌های به‌جا و مناسب را ممکن می‌سازد. این فنّاوری امروزه جای خود را در علوم پزشکی نیز به‌خوبی باز کرده و برای تشخیص و درمان بیماری‌ها به کمک پزشکان آمده است. ناباروری از جمله علوم جدیدی است که امروزه، به خصوص در کشور ما، زمان زیادی از پزشکان و هزینه‌ی بالایی از زوجین نابارور رابه خود اختصاص داده است. با این توصیف، هرگونه پژوهش و تحقیق در زمینه‌ی تسریع و تسهیل درمان‌های ناباروری، بسیار مهم وحیاتی به شمار می‌رود. در پژوهش حاضر ضمن مرور فرایند داده کاوی و الگوریتم‌های مختلف آن، به بهبود الگوریتم k-means پرداخته‌ایم و با رفع نقص این الگوریتم درخصوص انتخاب تصادفی خوشه‌های اولیه، و استفاده از الگوریتم سلسله‌مراتبی یک الگوریتم تلفیقی ارائه شد. در نهایت از این الگوریتم به‌منظور تحلیل داده‌های ناباروری بیمارستان صارم استفاده شده است.

الف) ناباروری

براساس تعاریف سازمان بهداشت جهانی (WHO)[5]، نازایی عبارت است از: «عدم وقوع حاملگی، یک سال پس از ازدواج، یا از زمانی که زوجین تصمیم به بچه‌دار شدن می‌گیرند (بدون استفاده از روش‌های پیشگیری)». به‌طور عام، علل مختلف ناباروری ناشی از عوامل مردانه[6]، عوامل زنانه[7]، یا عوامل نامعین[8] تشکیل می‌دهد. از درمان‌های رایج در ناباروری، تکنیک‌های کمک‌باروری (ART)[9] هستند که نقش مؤثری در درمان ناباروری با علل گوناگون دارند. ازجمله تکنیک‌های کمک‌باروری رایج می‌توان به IUI [10]، IVF [11]، ICSI [12]، GIFT [13] و ZIFT [14]اشاره کرد. در نوشتار حاضر داده‌های مربوط به درمان به روش تزریق داخل سیتوپلاسمی

* نویسنده مسئول
تاریخ: دریافت ۱۳۹۲/۴/۸، اصلاحیه ۱۳۹۲/۱۱/۱۸، پذیرش ۱۳۹۳/۳/۱۱.

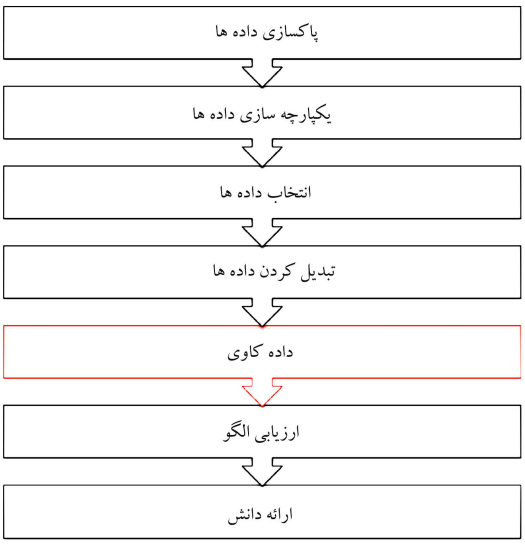
اسپرم (ICSI) بررسی شده است؛ داده‌های منتخب این مطالعه از پایگاه اطلاعات بیمارستان صارم، داده‌های مربوط به فرایند درمانی ICSI است.

ب) کشف دانش و داده کاوی

کشف دانش و داده کاوی، یک حوزه‌ی جدید میان‌رشته‌یی[15] و در حال رشد است که حوزه‌های مختلفی چون پایگاه داده، آمار، یادگیری ماشینی[16] و سایر زمینه‌های مرتبط را با هم تلفیق کرده تا اطلاعات و دانش ارزشمند نهفته در حجم بزرگی از داده‌ها را استخراج کند.[1و2]

کشف دانش در پایگاه داده همانا فرایند شناسایی درست، ساده، و مفید داده‌ها و ایجاد الگوها و مدل‌های قابل فهم در آنهاست که شامل الگوریتم‌های مخصوص داده کاوی است، به‌طوری که تحت محدودیت‌های مؤثر محاسباتی قابل قبول، الگوها یا مدل‌های داده را کشف می‌کند. داده کاوی به دو روش اصلی توصیفی[17] و پیش‌بینانه تقسیم می‌شود. در روش توصیفی هدف یافتن الگوهایی درمورد داده‌هاست به‌گونه‌یی که برای انسان قابل تفسیر باشد.[3] روش پیش‌بینانه نیز برای پیش‌بینی رفتارهای آینده‌ی داده‌ها کاربرد دارد. یکی از رایج‌ترین الگوریتم‌های توصیفی که درمورد داده‌های مختلف مورد استفاده قرار می‌گیرد، تحلیل خوشه‌بندی است که در این پژوهش نیز از این الگوریتم برای تحلیل داده‌ها استفاده شده است. مراحل انجام فرایند کشف دانش در شکل ۱ نمایش داده شده است.[4]

- **پاکسازی داده‌ها**[18]: از بین بردن نویز و ناسازگاری داده‌ها؛
- **یک‌پارچه‌سازی داده‌ها**[19]: ترکیب چند منبع داده؛
- **انتخاب داده‌ها**[20]: بازیافت داده‌های مرتبط با تحلیل از پایگاه داده‌ها؛
- **تبدیل داده‌ها**[21]: تبدیل داده‌ها به فرم مناسب برای داده کاوی، نظیر خلاصه‌سازی و همسان‌سازی؛
- **داده کاوی:** فرایند اصلی که در آن از شیوه‌های هوشمند برای استخراج الگو از داده‌ها استفاده می‌شود. از این دیدگاه، داده کاوی تنها یک مرحله از کل فرایند است؛ مرحله‌ی اساسی که الگوهای مخفی را آشکار می‌سازد؛
- **ارزیابی الگو**[22]: مشخص‌کردن الگوهای صحیح و مورد نظر به‌وسیله‌ی معیارهای اندازه‌گیری؛
- **ارائه‌ی دانش**[23]: نمایش بصری و تکنیک‌های بازنمایی دانش برای ارائه‌ی دانش کشف شده به کاربر.



شکل ۱. فرایند کشف دانش.

ج) مرور ادبیات

ازجمله پژوهش‌های انجام شده درکشور در حوزه‌ی ناباروری می‌توان اشاره کرد به:

- اثر مورفولوژی اسپرم در میزان حاملگی به روش تلقیح داخل رحمی اسپرم: تلقیح داخل رحمی اسپرم (IUI) در درمان زوج‌های نابارور ناشی از فاکتور مردانه کاربرد گسترده‌یی دارد.[5]
- بررسی ارتباط زنان نابارور با موفقیت فنّاوری‌های کمک‌باروری، در مراجعین به مراکز درمان ناباروری منتخب شهر تهران: شواهدی وجود دارد که اضطراب باعث افزایش کورتیزول و پرولاکتین و درنتیجه تشدید ناباروری می‌شود، در حالی که سطح اضطراب پایین‌تر به باروری طبیعی کمک می‌کند.[6]
- نتایج باروری به دنبال IVF و عوامل مؤثر برآن: میزان موفقیت IVF به علل ناباروری و درمان آن بستگی دارد. روش‌های مختلف درمانی و مراکز مختلف، میزان موفقیت متفاوتی را گزارش کرده‌اند.[7]
- ارتباط حاملگی با تعداد جنین منتقل شده در سیکل‌های ART: با توجه به این که در حال حاضر در مراکز ناباروری در ایران محدودیت‌های خاصی در تعداد جنین منتقله اعمال نمی‌شود، در یکی از بخش‌های دانشگاهی و مرجع درمان ناباروری در تهران رابطه‌ی تعداد جنین‌های منتقله با پیامد درمان ناباروری بررسی شد.[8]
- بررسی اپیدمیولوژیک علل ناباروری در بیماران مراجعه‌کننده به پژوهشکده‌ی رویان: شایع‌ترین علت ناباروری گزارش شده در تحقیقات مختلف عبارت‌است از: فاکتور مردانه، فاکتور زنانه، وجود همزمان فاکتور مردانه و زنانه، و در نهایت ناباروری با علت ناشناخته.[9]
- بررسی اثرات سن، مدت و علت ناباروری و تعداد فولیکول‌ها بر میزان موفقیت تلقیح داخل رحمی اسپرم: این تحقیق، با هدف بررسی میزان موفقیت IUI و تعیین عوامل مرتبط با موفقیت آن انجام شد و طی آن ۱۲۰ بیمار با بازه سنی ۲۰ تا ۴۲ سال که طی سال‌های ۱۳۸۴ تا ۱۳۸۷ به مرکز درمان ناباروری کاشان مراجعه کرده‌اند، بررسی شدند.[10]
- بررسی ارتباط فاکتورهای آنالیز اسپرم با پیامد درمان ناباروری به روش تزریق داخل سیتوپلاسمی اسپرم (ICSI): هدف از انجام این پژوهش، بررسی ارتباط فاکتورهای مختلف آنالیز سیمن بر پیامد درمان ناباروری به‌روش تزریق داخل سیتوپلاسمی اسپرم (ICSI) بوده است.[11]

پژوهش‌های زیاد دیگری نیز با استفاده از الگوریتم‌های داده کاوی در حوزه‌ی ناباروری انجام شده که در هریک از آن‌ها بر فاکتور خاصی تمرکز شده است. در مقاله‌ی دسته‌بندی بیزین برای انتخاب جنین‌ها در روش IVF، با استفاده از مشخصات زوجین، انتخاب جنین‌هایی با کیفیت بالاتر برای انجام IVF موجب افزایش احتمال موفقیت IVF می‌شود.[12] در پژوهش دیگری، رابطه‌ی احتمال حاملگی با تعداد جنین منتقل شده در روش IVF، با بررسی تعداد جنین منتقل شده در سیکل‌های ART انجام شد.[13] ترکیب الگوریتم ژنتیک و درخت تصمیم به‌منظور کمک به

پیش‌بینی نتایج IVF با استفاده از الگوریتم ژنتیک در مطالعه‌یی دیگر مورد بررسی قرار گرفته است.[۱۴] همچنین با استفاده از شبکه‌های عصبی برای پیش‌بینی نتایج IVF روشی خاص پیش‌بینی شده است.[۱۵] پژوهش‌های زیادی از این دست در این زمینه وجود دارد که تنها به تعداد اندکی از آن‌ها اشاره شده است.

د) پیش‌پردازش داده‌ها

پس از شناخت محیط بیمارستان و تحلیل فرایندهای درمانی، پایگاه داده بیمارستان به‌طورکامل بررسی شد و ضمن شناسایی فرایندهای درمانی مختلف با کادر بیمارستان در مورد آن‌ها بحث و مذاکره شد. سپس فرایندهای مربوط به پیش‌پردازش[۲۴] داده‌ها به‌منظور سازمان‌دهی آن‌ها به‌شکلی استاندارد که آماده‌ی پردازش توسط برنامه‌های داده کاوی باشند، انجام می‌گیرد. این گام از پروژه، تمامی موارد مربوط به داده‌های ناقص و ناکافی، داده‌های مغشوش و پرت، داده‌های ناسازگار و متضاد و... را در بر می‌گیرد. در نهایت، درگام‌های بعدی روش داده کاوی، تعیین پارامترهای لازم و ساخت و ارزیابی مدل انتخاب می‌شود؛ پس از ساخت مدل، نتایج حاصل از پروژه برای استفاده‌ی بهینه تحلیل و تفسیر خواهد شد. در هریک از مراحل بالا، ممکن است مرور دوباره یا انجام مجدد فرایندهای قبلی لازم باشد؛ بازگشت به عقب در این متدولوژی، مسئله‌یی غیرقابل انکار است که به‌خصوص در قسمت پاکسازی داده‌ها با آن مواجه می‌شویم.

یکی از مراحل مهم و اساسی در این پروژه که مستلزم صرف زمان زیادی بوده است، مرحله‌ی آماده‌سازی و پیش‌پردازش داده‌هاست. پایگاه داده‌های بیمارستان، رکوردهای زیادی از انجام روش‌های درمانی مختلف روی بیماران بیمارستان را شامل می‌شود. در این میان، داده‌های مربوط به بیمارانی که روش درمانی ICSI روی آن‌ها انجام شده، به‌منظور استفاده در پروژه انتخاب شد. فیلدهای مشخصات پرشده به‌ازای رکوردهای بیماران در پایگاه داده، شامل موارد زیاد و گسترده‌یی بود. با بررسی بیشتر و نظر خبرگان تعداد محدودی از این فیلدها که تأثیر بیشتری در فرایند درمانی بیماران داشته‌اند، انتخاب و در فرایند پروژه به کار گرفته شده‌اند. پس از انتخاب فیلدهای مؤثر در فرایند ICSI، بازه معنایی هریک از فیلدها با نظر خبرگان مشخص شد. همچنین با رجوع به پایگاه داده، مواردی که نشان‌گر داده‌های پرت و نادرست بود، حتی‌الامکان از پرونده‌ها مجدداً بازخوانی شد؛ در غیر این‌صورت با استفاده از فیلدهای کمکی دیگر پر شد یا از مجموعه‌ی داده‌ها حذف شد. فرایند دیگر در جهت آماده‌سازی و پیش‌پردازش داده‌ها، ترکیب برخی فیلدها برای ساختن فیلدهای جدید بوده است؛ به‌عنوان مثال می‌توان به فیلد شاخص توده بدنی ــ که از فیلدهای وزن و قد در محاسبه‌ی آن استفاده شد ــ یا فیلد سن ــ که از روی فیلد تاریخ تولد و تاریخ مراجعه به دست آمد ــ اشاره کرد. در نهایت فیلدهای مورد استفاده در پروژه عبارت است از:

ــ سن: بازه سنی مورد قبول درمورد رکوردهای ثبت شده در پایگاه داده بیماران شامل افراد ۲۰ تا ۵۱ سال است.

ــ طول مدت ناباروری: طول مدت ناباروری مورد قبول درمورد رکوردهای ثبت شده در پایگاه داده بیماران، طبق تعریف جهانی ناباروری یک سال به بالا در نظر گرفته شده است و بالاترین میزان ثبت شده در این رکوردها ۳۰ سال بوده است.

ــ وضعیت شاخص توده بدنی (BMI)[۲۵]: این شاخص، سنجشی آماری به‌منظور مقایسه‌ی وزن و قد فرد است و ابزاری مناسب برای تخمین سلامت وزنی فرد نسبت به قد اوست. شاخص توده‌ی بدنی زن با تقسیم وزن برحسب کیلوگرم بر مجذور قد برحسب متر محاسبه شده و نتایج آن در دسته‌های زیر جای گرفته‌اند:

- کم‌تر از ۱۸٫۵: کمبود وزن و لاغری؛
- ۱۸٫۵ تا ۲۴٫۹: وزن طبیعی و محدوده‌ی سلامت وزنی؛
- ۲۵ تا ۲۹٫۹: اضافه وزن؛
- بیشتر از ۳۰: چاق.

ــ علت ناباروری: ناباروری در زوجین با علل گوناگونی رخ می‌دهد که در میان داده‌های پایگاه اطلاعاتی این بیمارستان و پس از انجام پیش‌پردازش‌های لازم، سه علت ناباروری درمورد داده‌ها عبارت است از:

- نقص لوله‌های رحم (TF)[۲۶]: از علل زنانه؛
- نقص تخمدانی (OF)[۲۷]: از علل زنانه؛
- علت مردانه (MF).

ــ نوع ناباروری: ناباروری‌های رایج عبارت‌اند از:

- اولیه: زوج پس از ازدواج هرگز بچه‌دار نشده‌اند.
- ثانویه: شامل زوجینی که یک بار یا بیشتر سابقه‌ی بارداری و تولد فرزند داشته‌اند و سپس نابارور شده‌اند.

ــ نوع پروتکل درمانی: بسته به میزان و نحوه‌ی استفاده‌ی داروهای تحریک تخمدان، پروتکل‌های درمانی مختلفی به کار گرفته می‌شود:

- Short
- Long
- Pure

ــ تعداد جنین منتقل شده: در فرایند درمان ناباروری ICSI، به‌منظور ایجاد باروری فرایند تشکیل جنین در خارج از رحم انجام، و سپس جنین به رحم مادر منتقل می‌شود. تعداد جنین منتقل شده به رحم مادر در داده‌های این بیمارستان بین صفر تا ۷ جنین متغیر بوده که متناسب با شرایط زوج و تصمیم پزشک درمورد آن‌ها عمل شده است.

ــ مقدار نتایج تست هورمونی شامل:

- تست FSH
- تست LH
- تست Estradiol

ــ تعداد فولیکول: تعداد فولیکول بالغ پس از استفاده از داروهای تحریک تخمک‌گذاری؛

ــ تعداد تخمک: تعداد تخمک به دست آمده برای تشکیل جنین؛

ــ ضخامت آندومتر: با افزایش ضخامت آندومتر، احتمال باروری افزایش می‌یابد.

۲. روش بررسی

«خوشه‌بندی» یک تکنیک دسته‌بندی بدون نظارت است که در آن، مجموعه‌ی داده‌ها که معمولاً بردارهایی در فضای چندبعدی هستند، براساس معیار مشابهت یا عدم شباهت به تعداد مشخصی خوشه تقسیم می‌شوند.[۱۶]

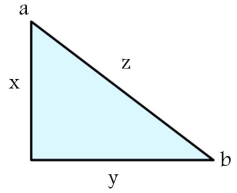
مسئله‌ی اساسی در الگوریتم خوشه‌بندی که در این پروژه به کار گرفته شده، عبارت است از: «توزیع داده‌ها به k گروه مختلف، به‌طوری که داده‌های هر گروه با یکدیگر مشابه و داده‌های گروه‌های مختلف با یکدیگر نامتشابه باشند.»[۱۷]

برای شباهت یا عدم شباهت بین داده‌ها، «فواصل» معیارهای خوبی هستند. برخی از فواصل معروف عبارت‌اند از: مانهاتان[۲۸]، مینکوفسکی[۲۹]، اقلیدسی[۳۰] و... در این پروژه، به‌منظور محاسبه‌ی فواصل بین خوشه‌ها، از معیار فاصله‌ی مانهاتان استفاده شده است. در الگوریتم خوشه‌بندی k-means، برای انتخاب بهترین تعداد خوشه‌ی داده‌ها، پس از خوشه‌بندی‌های مکرر (غالباً بین ۳ تا ۱۰ خوشه)، از یکی از شاخص‌های ارزیابی الگوریتم‌های خوشه‌بندی ــ نظیر شاخص دان[۳۱]، دیویس ـ بولدین[۳۲]، سیلوئت[۳۳] و... ــ استفاده می‌شود.

الف) ارزیابی و مقایسه‌ی مدل جدید

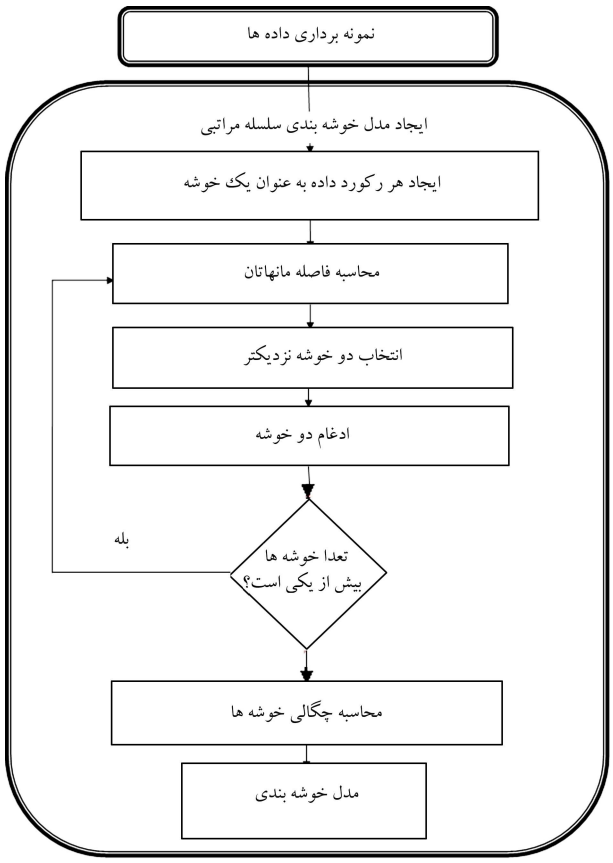
الگوریتم خوشه‌بندی k-means، با انتخاب تصادفی نقاط به‌عنوان مراکز اولیه‌ی خوشه‌ها آغاز شد و با هر بار تکرار خوشه‌بندی، مراکز خوشه‌ها به‌روز و خوشه‌بندی مجدداً انجام می‌شود. این فرایند تا هنگامی که مراکز خوشه‌ها تغییر نکنند ادامه می‌یابد. با توجه به انتخاب تصادفی مراکز اولیه، در آغاز اجرای الگوریتم، بهینه‌ی محلی تولید خواهد شد. بدین‌منظور با استفاده از الگوریتم سلسله‌مراتبی روی پایگاه داده، نقص انتخاب مراکز اولیه‌ی خوشه‌بندی برطرف می‌شود. به‌منظور بهبود الگوریتم k-means، از اجرای الگوریتم سلسله‌مراتبی روی پایگاه داده، به‌عنوان اولین مرحله از خوشه‌بندی داده‌ها استفاده کرده و سپس کار به روش پیشین ادامه می‌یابد. برای بهبود فرایند خوشه‌بندی، از الگوریتم سلسله‌مراتبی برای انتخاب نقاط اولیه و ورود به الگوریتم k-means استفاده شده است. فرایند انجام شده در مرحله‌ی اول فاز مدل‌سازی با الگوریتم سلسله‌مراتبی در شکل ۲ نشان داده شده است.

الگوریتم سلسله‌مراتبی به کار گرفته شده در این پروژه، از نوع تجمیعی یا پایین به بالا[۳۴]ست. مطابق شکل ۲، در اولین قدم هریک از رکوردها به‌عنوان یک خوشه در نظر گرفته می‌شود. در هر مرحله از این الگوریتم، فاصله‌ی مانهاتان همه‌ی خوشه‌ها با یکدیگر محاسبه، و دو خوشه‌ی نزدیک‌تر با یکدیگر ادغام می‌شوند. این روند تا تشکیل خوشه‌یی واحد از کلیه‌ی رکوردها ادامه می‌یابد. پس از اتمام فرایند سلسله‌مراتبی، چگالی تمامی خوشه‌های تشکیل شده در طی فرایند محاسبه می‌شود. مدل خوشه‌بندی که به‌عنوان خروجی این الگوریتم به دست می‌آید، شامل مشخصات تمامی خوشه‌های به دست آمده و چگالی همه‌ی خوشه‌هاست. برای محاسبه‌ی فاصله‌ی خوشه‌ها از یکدیگر و چگالی خوشه‌ها، از معیار فاصله‌ی مانهاتان استفاده شده است. در ادامه، نمایش معیار فاصله مانهاتان شرح داده شده است:



در مثلث نشان داده شده، در صورت نیاز به محاسبه‌ی فاصله‌ی بین a و b، $x+y$ نشان‌گر فاصله‌ی مانهاتان است (در این شکل، a و b نقاطی دوبعدی در نظرگرفته شده‌اند). اگر a و b نقاطی n بعدی باشند و a_i و b_i، نشان‌دهنده‌ی بعد iام این نقاط باشد، آنگاه فاصله‌ی مانهاتان چنین محاسبه می‌شود:

$$d(a,b) = \sum_{1 \le i \le n} |a_i - b_i|$$



شکل ۲. مرحله اول فاز مدل‌سازی (سلسله‌مراتبی).

فاصله‌ی دو خوشه و چگالی خوشه‌ها نیز چنین محاسبه می‌شوند:

$$D(C_i, C_j) : \frac{1}{n_i \times n_j} \sum_{\forall x_i \in c_i, y_j \in c_j} d(x,y)$$

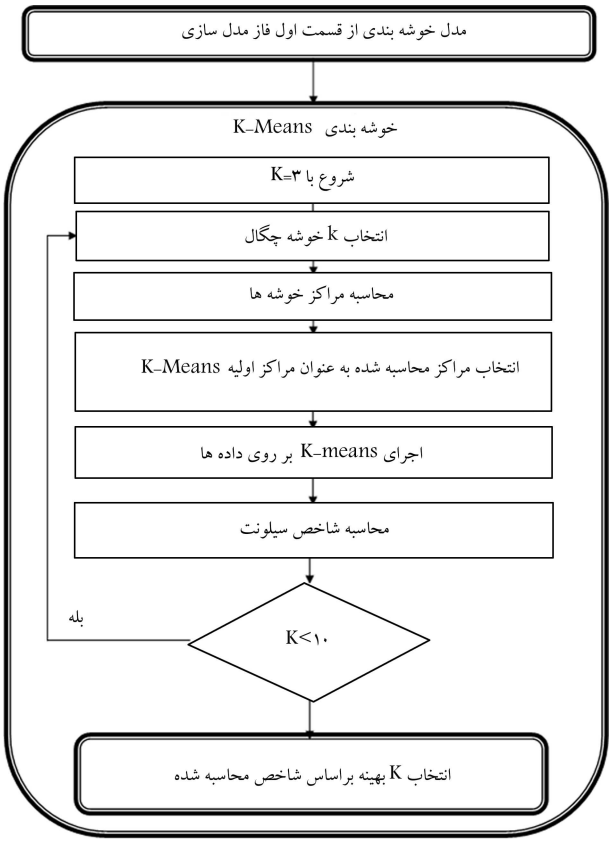
$$Density(C_i) : \frac{1}{n_i} \sum_{\forall x,y \in c_i, x \neq y} d(x,y)$$

n_i تعداد رکوردهای خوشه iام؛ n_j تعداد رکوردهای خوشه jام؛ $D(C_i, C_j)$ فاصله‌ی خوشه‌ی iام تا خوشه jام؛ $Density(C_i)$ چگالی خوشه‌ی iام.

پس از ایجاد مدل خوشه‌بندی با استفاده از الگوریتم سلسله‌مراتبی در مرحله‌ی اول فاز مدل‌سازی، الگوریتم k-means با استفاده از ورودی مدل سلسله‌مراتبی اجرا می‌شود. فرایند صورت گرفته در مرحله‌ی دوم فاز مدل‌سازی در شکل ۳ نشان داده شده است.

در مرحله‌ی دوم از فاز مدل‌سازی پروژه، به‌ازای هر بار اجرای الگوریتم k-means با تعداد k خوشه، با مراجعه به مدل خوشه‌بندی که در مرحله‌ی قبل به دست آمده است، k خوشه‌ی چگال‌تر برای استفاده به‌عنوان نقاط اولیه در الگوریتم k-means انتخاب می‌شود، و فرایند خوشه‌بندی ادامه می‌یابد. روند اجرای الگوریتم k-means به ازای ۳ تا ۱۰ خوشه ادامه می‌یابد و سپس با استفاده از یکی از شاخص‌های ارزیابی خوشه‌بندی، تعداد خوشه‌ی بهینه انتخاب می‌شود. از شاخص‌های مطرح می‌توان به شاخص سیلوئت اشاره کرد که در این پروژه کاربرد داشته است.[۱۸] شاخص سیلوئت از شاخص‌های مطرح در ارزیابی خوشه‌ها، پس از خوشه‌بندی با استفاده از الگوریتم‌های مختلف است.

اگر x را نقطه‌یی از خوشه‌ی C_k و n_k را تعداد نقاط آن خوشه در نظر بگیریم،



شکل ۳. مرحله‌ی دوم فاز مدل‌سازی (k-means).

مقدار الگوریتم سیلوئت در نقطه‌ی x متعلق به خوشه‌ی C_k چنین تعریف خواهد شد:

$$S(x) = \frac{b(x) - a(x)}{\max\left[b(x) - a(x)\right]}$$

که در آن $a(x)$ برابر با میانگین فاصله‌ی نقطه‌ی x با سایر نقاط در خوشه‌ی C_k است.

$$a(x,y) = \frac{1}{n_k - 1} \sum_{y \in C_k, y \neq x} d(x,y)$$

و $b(x)$ برابر است با میانگین فاصله‌ی نقطه‌ی x از نقاط متعلق به نزدیک‌ترین خوشه به C_k.

$$b(x) = \min_{h=1,\ldots,k, h \neq k} \left[\frac{1}{n_h} \sum_{y \in c_h} d(x,y) \right]$$

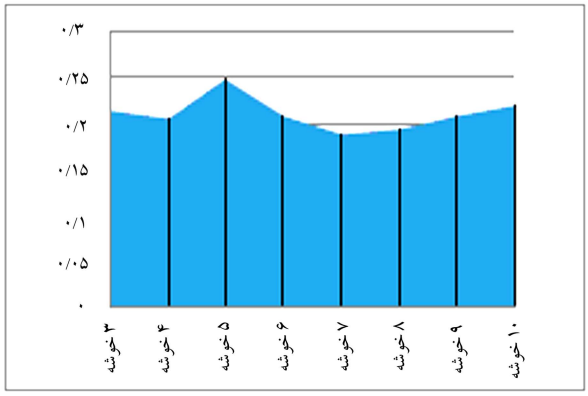
در نهایت، تعریف عام شاخص سیلوئت برای هر خوشه عبارت است از:

$$S = \frac{1}{k} \sum_{k=1}^{k} \left[\frac{1}{n_k} \sum_{x \in C_k} S(x) \right]$$

به‌ازای تمام xها، هرچه بیشتر بودن $b(x)$ و کم‌تر بودن $a(x)$، به‌معنای کم‌بودن قطر خوشه و زیاد بودن کم‌ترین فاصله‌ی بین خوشه‌هاست که منجر به بزرگ‌تر بودن شاخص S کلی خواهد شد و نتیجه‌ی مطلوب به دست می‌آید. مقدار این شاخص به‌ازای ۳ تا ۱۰ خوشه محاسبه شده و مقادیر آن در جدول ۱ و نمودار ۱ نشان داده شده است.

جدول ۱. مقادیر شاخص سیلوئت برای ۳ تا ۱۰ خوشه.

تعداد خوشه	مقدار شاخص
۳	۰٫۲۱۵۱۴۹۹۶۲
۴	۰٫۲۰۶۹۵۳۸۱۱
۵	۰٫۲۴۸۴۴۶۶۴۶
۶	۰٫۲۰۸۶۱۴۹۰۷
۷	۰٫۱۸۷۹۳۵۷۸
۸	۰٫۱۹۵۱۵۴۷۵۹
۹	۰٫۲۰۹۴۱۷۷۳۳
۱۰	۰٫۲۱۹۱۳۴۲۷۳



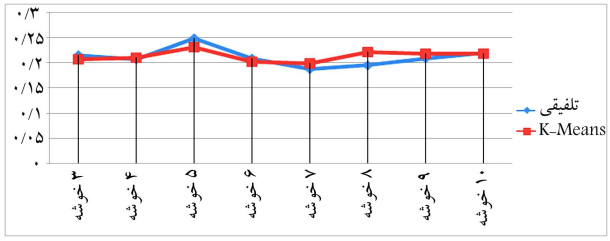
نمودار ۱. مقدار شاخص سیلوئت برای خوشه‌های مختلف.

چنان که در نمودار ۱ مشاهده می‌شود، تعداد بهینه‌ی خوشه برای خوشه‌بندی روی این داده‌ها، پنج خوشه است.

ب) ارزیابی مدل

مدل ارائه‌شده در این پروژه، به دلیل استفاده از الگوریتم سلسله‌مراتبی و انتخاب خوشه‌های چگال برای آغاز خوشه‌بندی، تا حد زیادی مشکل الگوریتم k-means در انتخاب تصادفی مراکز خوشه‌های اولیه را مرتفع کرده است. به‌منظور درک بیشتر بهینگی این مدل تلفیقی نسبت به الگوریتم k-means، خوشه‌بندی حاضر با خوشه‌بندی الگوریتم k-means مقایسه شده است. این مقایسه در شرایطی صورت گرفته که برای غلبه بر انتخاب تصادفی مراکز اولیه در فرایند اجرای الگوریتم k-means (که موجب تفاوت نتایج خوشه‌بندی در هر بار اجرای k-means می‌شود)، به‌منظور بهبود شرایط الگوریتم برای هر خوشه‌بندی با تعداد خوشه مشخص، ۱۰۰ بار الگوریتم k-means اجراشده است. پس از هر بار اجرا، شاخص ارزیابی سیلوئت محاسبه شده و در نهایت بهترین مقدار شاخص، برای هر خوشه‌بندی به‌عنوان مقدار شاخص سیلوئت برای آن تعداد خوشه انتخاب شده است. با این روند، برای هر یک از خوشه‌بندی‌های ۳ تا ۱۰ خوشه‌یی، نقاط تصادفی برای آغاز الگوریتم، ۱۰۰ بار انتخاب شده و مقدار شاخص سیلوئت نیز ۱۰۰ بار محاسبه شده است.

درنتیجه‌ی این فرایند، نمودار ۲ به‌منظور مقایسه‌ی مقادیر شاخص سیلوئت در الگوریتم بهبودیافته‌ی این پروژه و الگوریتم k-means با ۱۰۰ بار تکرار نشان داده شده است.



نمودار ۲. مقادیر شاخص سیلوئت برای الگوریتم k-means با ۱۰۰ بار تکرار و الگوریتم تلفیقی پروژه در ۳ تا ۱۰ خوشه.

چنان‌که در نمودار ۲ مشاهده می‌شود، روند کلی نمودار در هر دو حالت نسبتاً مشابه است و مقدار شاخص سیلوئت در هر دو حالت، تعداد ۵ خوشه را بهینه نشان داده است. اما مقدار شاخص سیلوئت در الگوریتم تلفیقی این پروژه، نسبت به الگوریتم k-means حتی با ۱۰۰ بار اجرای مجدد، بیشتر بوده است.

۳. یافته‌ها

فیلد مؤثر در فرایند درمانی زوجین ناباروری‌که در پرونده‌ی پزشکی آن‌ها در بیمارستان ثبت شده است، نتیجه‌ی درمان و مشخص کردن موفقیت یا عدم موفقیت در فرایند درمانی را شامل می‌شود. برای نتیجه‌گیری بهتر، نیازمند استفاده از داده‌های این فیلد و نگاشت آن به مشخصات ۵ خوشه‌ی بهینه به‌منظور بررسی شرایط بیماران در خوشه‌های مختلف هستیم. در جدول ۲ این فیلد از پرونده‌ی پزشکی زوجین نیز در مورد هر یک از پنج خوشه بیان شده است، و میزان موفقیت یا عدم موفقیت روند درمانی افراد در هر خوشه را نشان می‌دهد. فیلد نتیجه‌ی درمان در داده‌های این بیمارستان به دو نوع تقسیم شده است:

- مرحله‌ی بارداری شیمیایی[۳۵]: در این مرحله اولین تست بارداری را گرفته و نتیجه‌ی آن را ثبت می‌کنند.
- مرحله‌ی بارداری کلینیکی[۳۶]: در این مرحله، مشاهده‌ی کیست حاملگی در رحم نشان‌گر درمان موفقیت‌آمیز است.

در جدول ۲ مشخصات این پنج خوشه و نتیجه‌ی درمان در هر خوشه به تفکیک آمده است.

۴. نتیجه‌گیری

با توجه به نتایجی که در جدول ۲ مشاهده شد، ۵ خوشه به دست آمده با مشخصات مختلف، نتایج درمانی مختلفی به همراه داشته‌اند. از ویژگی‌های بارز هر یک از خوشه‌ها می‌توان به موارد زیر اشاره نمود.

۱. خوشه‌یی با بیشترین تعداد رکورد و کمترین میانگین سنی می‌باشد. (۲۹٫۹۲) نوع ناباروری تمامی نمونه‌های این خوشه، اولیه بوده و پروتکل درمانی به کار گرفته شده در مورد ۹۹٪ آن‌ها، پروتکل short می‌باشد. میانگین مقدار تست هورمونی FSH، کمترین مقدار (۴٫۶۵) و میانکین مقدار تست هورمونی LH، بیشترین مقدار (۹٫۱۸) در بین سایر خوشه‌ها می‌باشد. میزان موفقیت اعضای این خوشه در مرحله بارداری شیمیایی، ۱۷٫۳۵٪ و در مرحله بارداری کلینیکی ۹٫۵۳٪ بوده است.

جدول ۲. مشخصات و نتایج خوشه‌ها.

ویژگی‌ها		خوشه‌ها				
		۱	۲	۳	۴	۵
تعداد رکوردهای خوشه		۴۰۹	۷۷	۶۵	۵۷	۹۴
سن		۲۹٫۹۲	۳۴٫۵۷	۳۲٫۷۱	۳۷٫۰۵	۳۳٫۰۲
طول مدت ناباروری		۷٫۶۷	۵٫۳۵	۹٫۱۲	۹٫۹۹	۶٫۵۱
وضعیت BMI غالب		بین ۱۸٫۵ تا ۲۴٫۹، ۵۹٪	بین ۱۸٫۵ تا ۲۴٫۹، ۶۳٪	بین ۱۸٫۵ تا ۲۴٫۹، ۴۶٪	بین ۱۸٫۵ تا ۲۴٫۹، ۵۹٪	بین ۲۵ تا ۲۹٫۹، ۷۳٪
نوع ناباروری غالب		اولیه: ۱۰۰٪	ثانویه: ۱۰۰٪	اولیه: ۸۴٪	اولیه: ۹۴٪	ثانویه: ۵۷٪
علت ناباروری	MF	۹۵٪	۹۳٪	۹۲٪	۹۲٪	۷۹٪
	OF	۲۷٪	۴۹٪	۳۳٪	۲۴٪	۲۲٪
	TF	۸٪	۱۰٪	۲۱٪	۱۲٪	۷۶٪
پروتکل به کارگرفته شده		Short: ۹۹٪	Short: ۷۰٪	long: ۹۵٪	pure: ۴۷٪	Short: ۸۰٪
تعداد فولیکول		۱۰٫۲۵	۱۰٫۹۹	۷٫۶۶	۵٫۴۴	۶٫۱۸
تعداد تخمک		۱۱٫۹۶	۱۲٫۴۲	۸٫۲۸	۵٫۵۲	۶٫۹
ضخامت آندومتر		۹٫۳۳	۸٫۸۱	۸٫۶۸	۹٫۲	۹٫۴۶
تست‌های هورمونی	FSH	۴٫۶۵	۴٫۶۵	۵٫۴۱	۶٫۱۷	۴٫۶۵
	LH	۹٫۱۸	۸٫۱۷	۶٫۱۵	۷٫۱۶	۶٫۱۵
	Estradiol	۸۸٫۰۱	۵۴٫۷۳	۶۷٫۸۴	۷۴٫۵۶	۷۴٫۵۶
غالب تعداد جنین منتقل شده		۴ جنین: ۴۹٪	۴ جنین: ۶۴٪	۴ جنین: ۳۴٪	۳ جنین: ۶۴٪	۳ جنین: ۳۷٪
نتیجه درمان	مرحله بارداری شیمیایی	۱۷٫۳۵٪	۱۵٫۵۸٪	۲۳٫۰۷٪	۱۲٫۲۸٪	۱۴٫۸۹٪
	مرحله بارداری کلینیکی	۹٫۵۳٪	۱۲٫۹۸٪	۱۰٫۷۶٪	۵٫۲۶٪	۵٫۳۱٪

۲. اعضای این خوشه، دارای کمترین میانگین طول مدت ناباروری بوده (۵/۳۵) و تمامی اعضای آن دارای ناباروری ثانویه هستند. وضعیت BMI در ۶۳٪ اعضای این گروه، در حالت نرمال و محدوده سلامت قرار دارد و میانگین تعداد فولیکول (۱۰/۹۹) و تخمک (۱۲/۴۲) در اعضای این خوشه در مقایسه با سایر خوشه‌ها در سطح بالاتری قرارگرفته است. در مورد ۹۰٪ رکوردهای این خوشه، از پروتکل short استفاده کرده‌اند میزان موفقیت نمونه‌های این خوشه در مرحله بارداری شیمیایی، ۱۵/۵۸٪، و در مرحله بارداری کلینیکی ۱۲/۹۸٪ بوده است.

۳. ویژگی بارز این خوشه در نوع پروتکل درمانی استفاده شده در مورد اعضای آن می‌باشد. ۸۴٪ از اعضای این خوشه، دارای ناباروی اولیه بوده و ۹۵٪ از نمونه‌های این خوشه، پروتکل درمانی long را پشت سر نهاده‌اند. میانگین میزان تست LH(۶/۱۵) و Estradiol(۶۷/۸۴) در مورد اعضای این خوشه، کمترین سطح را در برگرفته است. میزان موفقیت اعضای این خوشه، نسبت به سایر خوشه‌ها در سطح بالاتری قرارگرفته است و موفق‌تر عمل نموده است. میزان موفقیت نمونه‌های این خوشه در مرحله بارداری شیمیایی، ۲۳/۰۷٪، و در مرحله بارداری کلینیکی ۱۰/۷۶٪ بوده است.

۴. افراد این خوشه را، نمونه‌هایی با میانگین سنی بالا (حدود ۳۷ سال) در برگرفته‌اند. میانگین طول مدت ناباروری در مورد رکوردهای این خوشه در حدود ۱۰ سال می‌باشد. نوع پروتکل درمانی به کارگرفته شده در مورد ۴۷٪ از نمونه‌های این خوشه، پروتکل long بوده است که در مقایسه با سایر خوشه‌ها که تعداد محدودی پروتکل long را اجرا نموده‌اند، نکته قابل تاملی به شمار می‌آید. کمترین میانگین تعداد فولیکول (۵/۴۴) و تعداد تخمک (۵/۵۲) به اعضای این گروه متعلق است و در ۶۳٪ نمونه از رکوردهای این خوشه، ۳ جنین به رحم مادر منتقل شده است. این خوشه دارای کمترین میزان موفقیت در بین همه خوشه‌ها بوده است. میزان موفقیت نمونه‌های این خوشه در مرحله بارداری شیمیایی، ۱۲/۲۸٪ و در مرحله بارداری کیلنیکی ۵/۲۶٪ بوده است.

۵. ۷۳٪ از اعضای این خوشه در وضعیت اضافه وزن BMI قرارگرفته‌اند. ۷۶٪ از نمونه‌های این گروه دارای مشکل لوله‌های رحمی هستند و بیشترین میانگین ضخامت آندومتر مربوط به اعضای این خوشه می‌باشد. (۹/۴۶) میزان موفقیت در مرحله بارداری کلینیکی در این خوشه با افت بسیار همراه بوده است و کمتر از نصف افراد موفق در مرحله بارداری شیمیایی، در مرحله بعدی موفق بوده‌اند. میزان موفقیت نمونه‌های این خوشه در مرحله اول، ۱۴/۸۹٪ و در مرحله بعد ۵/۳۱٪ بوده است.

۵. پیشنهادها

پس از انجام مطالعات توصیفی داده کاوی، در مطالعات بعدی به‌منظور استفاده بیشتر از داده‌های موجود می‌توان به الگوریتم‌های پیش‌بینی داده کاوی پرداخت و از آن‌ها برای پیش‌بینی احتمال موفقیت درمورد زوجین نابارور و نیز پیشنهاد راه درمان بهتر برای هریک استفاده کرد.

تقدیر و تشکر

این پروژه بدون یاری کادر متخصص و متعهد بیمارستان تخصصی صارم، به‌ویژه جناب آقای دکتر صارمی امکان‌پذیر نبود. در اینجا بر خود لازم می‌دانیم، مراتب قدردانی و تشکر خود را از ایشان و تمامی همکاران محترم‌شان اعلام نماییم.

پانوشت‌ها

1. data mining
2. clustering
3. unsupervised
4. hierarchical
5. world health organization (WHO)
6. male factor
7. female factor
8. unexplained
9. assisted reproductive technology (ART)
10. intra uterine insemination
11. in vitro fertilization
12. intra cytoplasmic sperm injection
13. Gamete intra fallopian transfer
14. Zygote intra fallopian transfer
15. interdisciplinary
16. machine learning
17. descriptive
18. data cleaning
19. data integration
20. data selection
21. data transformation
22. pattern evaluation
23. knowledge presentation
24. preprocessing
25. body mass index (BMI)
26. tubal factor
27. ovarian factor
28. Manhattan
29. Minkowski
30. Euclidean
31. dunn index
32. Davies-Bouldin's index
33. Silhouette
34. bottom - up
35. chemical pregnancy
36. clinical pregnancy

منابع (References)

1. Chakrabarti, S., Ester, M., Fayyad, U., Gehrke, J., Han, J., Morishita, S., Piatetsky-Shapiro, G. and Wang, W. "Data mining curriculum: A proposal (version 1.0)", Intensive Working Group of ACM SIGKDD Curriculum Committee (2006).

2. Berry, M.J.A. and Linoff, G.S., *Data Mining Thechniques: For Marketing, Sale, and Customer Relationship Managment*, 2nd Edition, Wiley (2004).
3. Han, J. and Kamber, M., *Chapter 1: Intruduction, Data Mining Concept: and Techniques*, 2nd Edition, Morgan kufman (2006).
4. Edelstein, H.A., *Introduction to Data Mining and Knowledge Discovery*, 3nd Edition, Two Crows Corporation (1999).
5. Esmaeilzadeh S., Farsi M., Bizhani A., "Effects of sperm morphology on pregnancy rate in IUI cycles" (in Persain) (2007).
6. Simbar M., Hashemi S., Shams J., AlaviMajd H., "Association between infertile womens anxiety with art success rates" (in Persain) (2009).
7. Vahid Roudsari F., Ayati S., Mirzaeeyan S., Shakeri M., Akhtardel H., "Fertility outcome after IVF and related factors" (in Persain) (2009).
8. Sohrabvand F., Shariat M., Fotouhi Ghiam N., Hashemi M., "The relationship between number of transferred embryos and pregnancy rate in ART cycles" (in persain) (2009).
9. Kamali M., Baghestani A., Kashfi F., Kashani H., Tavajohi SH., Amir Chaghmaghi E., "A survey on interfility in royan institute" (in Persain) (2007).
10. Hasani Baferani H., Abedzadeh M., Fruzanfard F., Tabasi Z., "Effects of patient age, duration and cause of infertility and number of pre-ovulatory follicles on intrauterine insemination outcomes" (in Persain) (2010).
11. Dadkhah F., Kashanian M., Agahi A., "Evaluation of the relationship between semen parameters and the outcome of the intra-cytoplasmic sperm injection (icsi)" (in Persain) (2010).
12. Morales, D.A., Bengoetxea, E., Larrañaga, P., García, M., Franco, Y., Fresnada, M. and Merino, M. "Bayesian classification for the selection of in vitro human embryos using morphological and clinical data", *Computer Methods and Programs in Biomedicine*, **90**(2), pp. 104-116 (May 2008).
13. Sohrabvand, F., Shariat, M., Fotouhi Ghiam, N. and Hashemi, M. "The relationship between number of transferred embryos and pregnancy rate in ART cycles", *Tehran University Medical Journal (TUMJ)*, **67**(2), pp. 132-136 (May 2009).
14. Guh, R.Sh., Wu, T.-Ch.J. and Weng, Sh.P. "Integrating genetic algorithm and decision tree learning for assistance in predictingin vitro fertilization outcomes", *Expert System with Application: An International Journal*, **38**(4), pp. 4437-4449 (April 2011).
15. Kaufmann, S.J., Eastaugh, J.L., Snowden, S., Smye, S.W. and Sharma, V. "The application of neural networks in predicting the outcome of in-vitro fertilization", *Human Reproduction*, **12**(7), pp, 1454-1457 (1997).
16. Han, J., Kamber, M. and Tung, A.K.H. "Spatial clustering methods in data mining: A survey", Geographic Data Mining and Knowledge Discovery, pp. 1-29 (2001).
17. Han, J. and Kamber, M., *Chapter 7: Cluster Analysis: Data Mining Concept and Techniques*, 2nd Edition, Morgan Kufman (2006).
18. Rousseeuw, P.J. "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis", *Journal Computational and Applied Mathematics*, **20**, pp. 53-65 (1987).