علوم رایانش و فناوری اطلاعات
نشریه علمی انجمن کامپیوتر ایران
مجلد ۲۲، شماره ۱، ۱۴۰۳
صفحات ۳۱-۳۹
مقاله پژوهشی

تحلیل آماری و کاوش روی داده‌های پزشکی به‌منظور پیش‌بینی و تشخیص بیماری

علی نقاش اسدی[1*]، فاطمه عاصی آتشکاهی[2]

*نویسنده مسئول، دریافت: ۰۳/۰۷/۱۹، بازنگری: ۰۳/۰۸/۲۳، پذیرش: ۰۳/۰۹/۱۷

[1] استادیار، گروه مهندسی کامپیوتر، دانشکده فنی فومن، دانشکدگان فنی دانشگاه تهران، فومن، ایران
[2] دانشجوی کارشناسی ارشد، دانشکده مهندسی کامپیوتر، دانشگاه صنعتی خواجه‌نصیرالدین طوسی، تهران، ایران

چکیده

استفاده از هوش مصنوعی در پیش‌بینی و تشخیص بیماری‌ها، موضوع بسیار مهمی است. برای دستیابی به هوش مصنوعی در این حوزه، اولین کاری که باید انجام شود، تحلیل داده آماری و انجام فرآیند داده‌کاوی بر روی داده‌های موجود است. تحلیل داده آماری، اطلاعات مهمی از وضعیت داده‌های موجود در اختیار محققان قرار می‌دهد و فرآیند داده‌کاوی، الگوهای میان داده‌ها را استخراج می‌کند. از این اطلاعات و الگوها می‌توان برای ساخت ابزارهایی برای کمک به پزشکان برای تشخیص بیماری‌ها استفاده کرد. متأسفانه بیماری‌های مختلفی در جهان وجود دارد و علائم آن‌ها در بیماران مختلف، متفاوت است. انتخاب یک بیماری و شناسایی علائم آن و سپس استخراج الگویی که بتوان از طریق علائم قابل‌مشاهده، بیماری را تشخیص داد، از اهمیت بالایی برخوردار است. در این مقاله، داده‌های جمع‌آوری‌شده از طریق پرسش‌نامه از وضعیت مادران پس از زایمان موردبررسی قرارگرفته است. ابتدا تحلیل آماری از پاسخ مادران به عوامل مختلف ارائه می‌شود. سپس با استفاده از الگوریتم بدون نظارت، مادران بر اساس پاسخ‌هایشان به دو خوشه تقسیم می‌شوند. در ادامه عوامل مهمی که در خوشه‌بندی تأثیرگذار بودند، شناسایی‌شده و با استفاده از درخت تصمیم، نحوه انتساب پاسخ‌ها به هر خوشه، مشخص می‌شود. با استفاده از نتایج به‌دست‌آمده می‌توان با مشاهده یک یا چند عامل در یک مادر جدید، خطراتی که او در آینده با آن‌ها مواجه خواهد شد را شناسایی کرد. نتایج به‌دست‌آمده منجر به بهبود مراقبت‌های بهداشتی و درمانی برای مادران خواهد شد.

کلمات کلیدی: افسردگی، تحلیل داده آماری، تشخیص بیماری، داده‌کاوی، درخت تصمیم.

۱- مقدمه

امروزه در بیمارستان‌ها، داده‌های زیادی از وضعیت بیماران و نتیجه فرآیند درمانی آن‌ها وجود دارد، ولی این داده‌ها در بیشتر مواقع مورد تحلیل و ارزیابی قرار نمی‌گیرند [1]. از دلایل این موضوع می‌توان به عدم ذخیره‌سازی ساختارمند این داده‌ها در بیمارستان‌ها اشاره کرد. علاوه بر این، سیاست‌های بیمارستان در خصوص رعایت حریم خصوصی بیماران و همچنین عدم انتشار اطلاعات سبب شده است که داده‌های بیمارستانی کمی برای بررسی در اختیار محققان قرار گیرد. درصورتی‌که این داده‌ها می‌توانند در فرآیند داده‌کاوی مورداستفاده قرارگرفته و امکان ساخت ابزارهایی مبتنی بر هوش مصنوعی برای کمک به پزشکان برای پیش‌بینی و تشخیص بیماری‌ها را فراهم کنند. بنابراین این داده‌ها می‌توانند تأثیر بسیار زیادی در بهبود مراقبت‌های بهداشتی و درمانی بیماران داشته باشند.

محققان با تحلیل داده آماری داده‌های بیمارستانی می‌توانند فراوانی و احتمال وقوع هر یک از علائم را در هر بیماری شناسایی کنند. همچنین با انجام فرآیند داده‌کاوی، الگوهای پنهان میان داده‌ها کشف می‌شود [2]. به‌عبارت‌دیگر، با کشف این الگوها می‌توان تشخیص داد که با مشاهده یک یا چند عامل در یک فرد، احتمال وقوع چه بیماری‌هایی در او وجود دارد. متأسفانه بیماری‌های مختلفی در جهان وجود دارد و علائم آن‌ها در بیماران مختلف، متفاوت است. بنابراین اگر پزشکان بخواهند صرفاً بر اساس دانش و تجربه خود، اقدام به درمان بیماری‌ها کنند، احتمال خطای آن‌ها بالا خواهد رفت و هزینه‌های زیادی برای بیماران، شرکت‌های بیمه و دولت‌ها به همراه خواهد داشت [3]؛ درصورتی‌که اگر ابزارهایی مبتنی بر هوش مصنوعی در اختیار داشته باشند، که بر اساس داده‌کاوی روی داده‌های بیمارستانی ساخته‌شده‌اند، احتمال خطای آن‌ها تا حد زیادی کاهش پیدا خواهد کرد.

یکی از بیماری‌های مهمی که در تحقیقات کمی موردتوجه قرارگرفته است، افسردگی پس از زایمان مادران است. این بیماری یک اختلال روانی و جسمی بوده و می‌تواند بر رفتار و سلامت مادران تأثیرگذار باشد [4]. بر اساس اطلاعات منتشرشده توسط سازمان بهداشت زنان آمریکا، تجربه کوتاه‌مدت غم و اندوه پس از زایمان برای مادران جدید امری طبیعی است. در بیشتر موارد، این احساسات طی دو هفته کاهش می‌یابند. بااین‌حال، از هر نه مادر جدید، یک مادر با احساسات غمگینی، اختلال در خواب و تغذیه، و ناتوانی در تمرکز برای بیش از دو هفته مواجه می‌شود. مادرانی با حمایت اجتماعی کم و سابقه افسردگی قبلی درخطر ابتلا به این بیماری هستند. حتی مادرانی که این مشکلات را نداشته باشند نیز ممکن است به این بیماری مبتلا شوند. عدم درمان افسردگی پس از زایمان می‌تواند تأثیر منفی بر توانایی مادر در ارائه مراقبت‌های لازم برای تربیت فرزند سالم داشته باشد. بنابراین، شناسایی زودهنگام این اختلال و دریافت کمک‌های مناسب می‌تواند به بهبود سلامت جسمی و روانی مادر و کودک کمک کند.

در ادامه این مقاله و در بخش ۲، ابتدا کارهای مرتبط معرفی‌شده و با کار انجام‌شده در این مقاله مقایسه می‌شود. در ادامه و در بخش ۳، تحلیل داده آماری و فرآیند داده‌کاوی انجام‌شده با استفاده از الگوریتم k-means و درخت تصمیم ارائه می‌شود. همچنین در این بخش، تفسیری از نتایج به‌دست‌آمده ارائه می‌شود. در انتها و در بخش ۴، نتیجه‌گیری و کارهای آینده مقاله معرفی می‌شود.

۲- کارهای مرتبط

از داده‌کاوی برای استخراج الگوها در موضوعات مختلفی استفاده‌شده است [5] [6]. همچنین فرآیند داده‌کاوی در حوزه پزشکی و تشخیص بیماری‌ها در مقالات بسیاری انجام‌شده است؛ بااین‌حال در بسیاری از این مقالات، بیماری‌های روانی مانند افسردگی، بررسی نشده است. در جدول ۱، خلاصه‌ای از مقایسه بین کارهای مرتبط ارائه‌شده است. برای مثال، در [7]، مقالات منتشرشده بین سال‌های ۲۰۰۰ تا ۲۰۲۲ که از روش‌های داده‌کاوی یا فرآیندکاوی[1] در پیش‌بینی و تشخیص بیماری‌های مزمن استفاده کرده‌اند، موردبررسی قرارگرفته است. از نتایج این مقاله، شناسایی روند رو به رشد استفاده از روش‌های داده‌کاوی در تحقیقات مربوط به تشخیص بیماری دیابت و سرطان بود. در [8]، تحقیقات انجام‌شده از سال ۲۰۰۴ تا ۲۰۲۰ مربوط به تشخیص بیماری پارکینسون (به‌عنوان دومین بیماری مهم عصبی بعد از آلزایمر) و مراحل آن با استفاده از روش‌های داده‌کاوی موردبررسی قرارگرفته است. در [9]، از الگوریتم k نزدیک‌ترین همسایه‌ها[2] (KNN) برای تشخیص بیماری کووید۱۹ بر پایه اینترنت اشیا استفاده‌شده است. در [1]، از الگوریتم‌های داده‌کاوی مانند درخت تصمیم[3] و بیز ساده[4] برای بررسی مجموعه داده‌ای از آزمایش‌های خونی برای پیش‌بینی بیماری‌های خونی در مراحل اولیه استفاده‌شده است. در [3]، از چهار الگوریتم درخت تصمیم، جنگل تصادفی[5]، بیز ساده و k نزدیک‌ترین همسایه‌ها، برای پیش‌بینی بیماری‌ها و دسته‌بندی آن‌ها استفاده شد، به‌طوری‌که بتوان تشخیص داد که هر بیمار از ۱۴۷ دسته مختلف از چه نوع بیماری رنج می‌برد و هر بیماری به کدام‌یک از ۲۲ دسته جداگانه تعلق دارد. در [10]، باهدف ارائه مدلی برای پیش‌بینی بیماری‌های قلبی عروقی و شناسایی عوامل کلیدی مؤثر، از روش‌های مختلف داده‌کاوی مبتنی بر یادگیری گروهی استفاده‌شده است. در [11]، طرح کلی از راهبردهای داده‌کاوی قابل‌استفاده در زمینه‌های سلامت شامل بیماری‌های قلبی، دیابت، سرطان سینه، اختلال کبدی و بیماری کلیوی ارائه‌شده است. در [12]، با بررسی داده‌های دریافتی از حسگرها و انجام فرآیند داده‌کاوی، از ویژگی‌ها و الگوهای حرکتی، برای تشخیص بیماری‌های عصبی استفاده شد. در [13]، از روش‌های داده‌کاوی بدون نظارت ازجمله خوشه‌بندی و استخراج قوانین انجمنی برای به حداقل رساندن نیاز به آزمایش‌های اضافی خون در تشخیص بیماری‌ها استفاده‌شده است. استفاده از الگوهای به‌دست‌آمده از این مقاله می‌تواند منجر به صرفه‌جویی قابل‌توجه در زمان و هزینه و درعین‌حال کاهش بار کاری برای کارکنان آزمایشگاه شود.

جدول ۱- مقایسه بین کارهای مرتبط

مقاله	عوامل بررسی‌شده
[1]	استفاده از درخت تصمیم و بیز ساده برای پیش‌بینی بیماری‌های خونی
[3]	استفاده از درخت تصمیم، جنگل تصادفی، بیز ساده و KNN برای پیش‌بینی بیماری‌ها
[7]	مرور مطالعات انجام‌شده در پیش‌بینی و تشخیص بیماری‌های مزمن
[8]	مرور مطالعات انجام‌شده در تشخیص بیماری پارکینسون
[9]	استفاده از الگوریتم k نزدیک‌ترین همسایه‌ها برای تشخیص بیماری کووید۱۹ بر پایه اینترنت اشیا
[10]	استفاده از الگوریتم‌های یادگیری گروهی برای پیش‌بینی بیماری‌های قلبی
[11]	ارائه یک طرح کلی از راهبردهای داده‌کاوی برای تشخیص بیماری‌های قلبی، دیابت، سرطان سینه، اختلال کبدی و بیماری کلیوی
[12]	استفاده از فرآیند داده‌کاوی، برای تشخیص بیماری‌های عصبی
[13]	استفاده از روش‌های داده‌کاوی بدون نظارت برای به حداقل رساندن نیاز به آزمایش‌های اضافی خون در تشخیص بیماری‌ها
[14]	استفاده از الگوریتم داده‌کاوی برای پیش‌بینی افسردگی شدید بر اساس اطلاعات بالینی
[15]	استفاده از روش‌های طبقه‌بندی برای پیش‌بینی خطر افسردگی دانش‌آموزان بر اساس جمعیت‌شناختی اجتماعی، اعتیاد به اینترنت، اعتیاد به مصرف الکل و سطح استرس
[16]	تشخیص افسردگی از متون موجود در شبکه‌های اجتماعی با مدل‌های یادگیری ماشین
[17]	ارائه یک چارچوب برای تشخیص افسردگی کاربران شبکه‌های اجتماعی
[18]	پیش‌بینی افسردگی پس از زایمان با بررسی فشارخون
مقاله ما	استفاده از درخت تصمیم و k-means برای پیش‌بینی بیماری افسردگی پس از زایمان

در معدود کارهایی که از روش‌های داده‌کاوی در پیش‌بینی و تشخیص بیماری افسردگی استفاده کرده‌اند، موضوع افسردگی آن‌ها متفاوت بوده است. در [14]، عملکرد یک الگوریتم داده‌کاوی برای پیش‌بینی افسردگی شدید بر اساس اطلاعات بالینی، بررسی‌شده است. در [15]، برای پیش‌بینی خطر افسردگی دانش‌آموزان بر اساس جمعیت‌شناختی اجتماعی، اعتیاد به اینترنت، اعتیاد به مصرف الکل و سطح استرس، از پنج روش طبقه‌بندی مختلف استفاده‌شده است. همچنین در این مقاله، یک روش نمونه‌گیری ترکیبی برای بهبود عملکرد طبقه‌بندی ارائه‌شده است. در [16]، متونی که حاوی مطالبی ناشی از افسردگی هستند، موردبررسی قرارگرفته است. برای این منظور، ابتدا ویژگی‌ها استخراج‌شده و با استفاده از تحلیل مؤلفه‌های اصلی، رویکرد تحلیل احساسات دسته‌بندی‌شده و با استفاده از اعتبارسنجی متقابل با مدل‌های یادگیری ماشین (مانند چندجمله‌ای ساده بیز، k نزدیک‌ترین همسایه‌ها و SVM) یک پیش‌بینی کننده ساخته‌شده است. در [17]، یک چارچوب برای تجزیه‌وتحلیل کلان داده‌ها[6] برای تشخیص افسردگی کاربران شبکه‌های اجتماعی ارائه‌شده است. برای این منظور علاوه بر ویژگی‌های نوشتاری، ویژگی‌های عمل‌گرایانه نیز موردبررسی قرارگرفته است. ازآنجایی‌که رفتار دوستان هر کاربر در شبکه‌های اجتماعی بر خود کاربر نیز تأثیرگذار است، این چارچوب نیز تأثیر دوستان را بر حالات ذهنی کاربر، مدل می‌کند.

در تحقیقات محدودی، از روش‌های داده‌کاوی برای تشخیص بیماری افسردگی پس از زایمان استفاده‌شده است ولی بااین‌حال، معیارها و عوامل در نظر گرفته‌شده توسط

[1] Process Mining
[2] K Nearest Neighbor
[3] Decision Tree
[4] Naïve Bayes
[5] Random Forest
[6] Big Data

آن‌ها، متفاوت با مقاله ما بوده است. برای مثال، در [18]، یک الگوریتم بهبودیافته برای سامانه‌های هوشمند آگاه از احساسات ارائه‌شده است که قادر است، خطر افسردگی پس از زایمان در زنان مبتلابه اختلالات فشارخون بالا در دوران بارداری را از طریق تجزیه‌وتحلیل داده‌های زیست پزشکی و جمعیت‌شناختی پیش‌بینی کند.

۳- تحلیل داده آماری و داده‌کاوی

۳-۱- مجموعه داده و تحلیل داده آماری

برای انجام این تحقیق، داده‌های جمع‌آوری‌شده از یک بیمارستان از تاریخ ۱۴ الی ۱۵ ژوئن ۲۰۲۲ موردبررسی قرار گرفت. داده‌های جمع‌آوری‌شده در [19] قابل‌دسترس است. در این بیمارستان، پرسش‌نامه‌ای تهیه‌شده و از مادران جدید در مورد علائم افسردگی پس از زایمان پرسیده شده است. ۱۵۰۳ نفر از زنان بین ۲۵ تا ۵۰ سال در این پرسش‌نامه شرکت کردند که بزرگ‌ترین گروه بین ۴۰ تا ۴۵ سال و کوچک‌ترین گروه بین ۲۵ تا ۳۰ سال بوده است. در جدول ۲، اطلاعاتی در مورد ۱۰ عامل در نظر گرفته‌شده در پرسش‌نامه، که مؤثر بر بیماری افسردگی پس از زایمان هستند، ارائه‌شده است. این عوامل شامل «سن»، «احساس غمگینی[1]»، «تحریک‌پذیری نسبت به نوزاد یا شریک زندگی[2]»، «مشکلات خواب[3]»، «مشکلات تمرکز یا تصمیم‌گیری[4]»، «پرخوری یا کم‌اشتهایی[5]»، «احساس اضطراب[6]»، «احساس گناه[7]»، «مشکلات ارتباط با کودک[8]»، و «اقدام به خودکشی[9]» هستند. در بررسی اولیه مجموعه داده مشخص شد که به ترتیب ۱۲ و ۶ رکورد از مجموعه داده برای عوامل «مشکلات تمرکز یا تصمیم‌گیری» و «تحریک‌پذیری نسبت به نوزاد یا شریک زندگی» بدون مقدار بودند. بنابراین رکوردها حذف‌شده و تحلیل‌ها روی ۱۴۹۱ رکورد انجام‌شده است. در شکل ۱، نمودار فراوانی پاسخ مادران به هر عامل نشان داده‌شده است.

همان‌طور که از نمودارهای شکل ۱ برداشت می‌شود، بیشتر مادران فقط به عواملی مانند «پرخوری یا کم‌اشتهایی» و «احساس گناه»، و «اقدام به خودکشی» پاسخ خیر داده‌اند و به سایر عوامل در بیشتر موارد پاسخ بله داده‌اند. برای آنکه متوجه شویم، احتمال ارائه پاسخ بله به هر سؤال در کدام گروه سنی بیشتر است، شکل ۲ ارائه‌شده است. ازآنجایی‌که تعداد مادران در هر گروه سنی متفاوت است، برای محاسبه احتمال پاسخ، تعداد پاسخ بله مادران در یک گروه سنی تقسیم‌بر تعداد کل مادران در آن گروه سنی خواهد شد. مطابق با شکل ۲، اطلاعات زیر قابل دریافت است:

- بیشترین احتمال احساس غمگینی با مقدار ۰/۷ مربوط به مادرانی با سن ۴۵ الی ۵۰ سال است؛ و کمترین احتمال با مقدار ۰/۵۸ مربوط به مادرانی با سن ۴۰ الی ۴۵ سال است.
- بیشترین احتمال تحریک‌پذیری نسبت به نوزاد یا شریک زندگی با مقدار ۰/۷۵ مربوط به مادرانی با سن ۴۰ الی ۴۵ سال است؛ و کمترین احتمال با مقدار ۰/۵۹ مربوط به مادرانی با سن ۲۵ الی ۳۰ سال و ۴۵ الی ۵۰ سال است.
- بیشترین احتمال مشکلات خواب با مقدار ۰/۷۹ مربوط به مادرانی با سن ۲۵ الی ۳۰ سال است؛ و کمترین احتمال با مقدار ۰/۶۴ مربوط به مادرانی با سن ۴۵ الی ۵۰ سال است.
- بیشترین احتمال مشکلات تمرکز یا تصمیم‌گیری با مقدار ۰/۷۱ مربوط به مادرانی با سن ۳۵ الی ۴۰ سال است؛ و کمترین احتمال با مقدار ۰/۵۵ مربوط به مادرانی با سن ۴۵ الی ۵۰ سال است.
- بیشترین احتمال پرخوری یا کم‌اشتهایی با مقدار ۰/۳۱ مربوط به مادرانی با سن ۲۵ الی ۳۰ سال است؛ و کمترین احتمال با مقدار ۰/۱۳ مربوط به مادرانی با سن ۴۵ الی ۵۰ سال است. مطابق با این نمودار، هرچه سن مادران بالاتر می‌رود، احتمال مشکلات پرخوری یا کم‌اشتهایی آن‌ها کمتر می‌شود.
- بیشترین احتمال احساس اضطراب با مقدار ۰/۷۲ مربوط به مادرانی با سن ۴۵ الی ۵۰ سال است؛ و کمترین احتمال با مقدار ۰/۵۷ مربوط به مادرانی با سن ۲۵ الی ۳۰ سال است.
- بیشترین احتمال احساس گناه با مقدار ۰/۳۳ مربوط به مادرانی با سن ۲۵ الی ۳۰ سال است؛ و کمترین احتمال با مقدار ۰/۱۶ مربوط به مادرانی با سن ۴۵ الی ۵۰ سال است.
- بیشترین احتمال مشکلات ارتباط با کودک با مقدار ۰/۶۸ مربوط به مادرانی با سن ۳۰ الی ۳۵ سال است؛ و کمترین احتمال با مقدار ۰/۵۵ مربوط به مادرانی با سن ۴۰ الی ۴۵ سال است.
- بیشترین احتمال اقدام به خودکشی با مقدار ۰/۳۴ مربوط به مادرانی با سن ۲۵ الی ۳۰ سال و ۳۵ الی ۴۰ سال است؛ و کمترین احتمال با مقدار ۰/۲۲ مربوط به مادرانی با سن ۴۵ الی ۵۰ سال است.

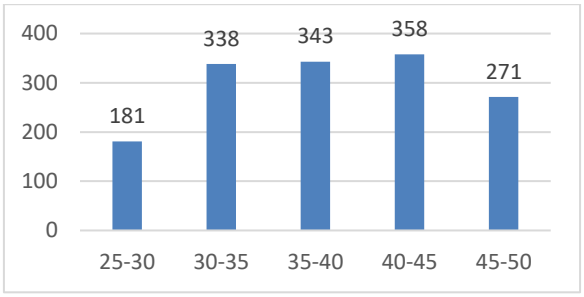
از اطلاعات فوق می‌توان برداشت کرد، مادران که در گروه سنی ۲۵ الی ۳۰ سال قرار دارند در ۴ عامل شامل مشکلات خواب، پرخوری یا کم‌اشتهایی، احساس گناه، و اقدام به خودکشی بیشترین احتمال مشکل را دارند.

جدول ۲- عوامل در نظر گرفته‌شده برای بیماری افسردگی پس از زایمان

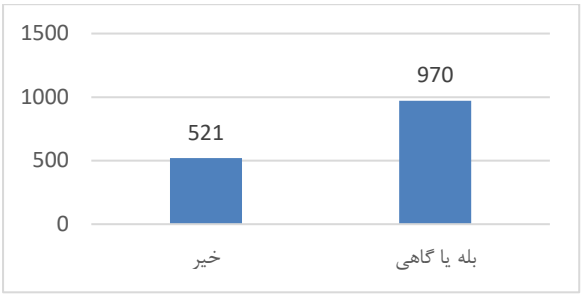
ردیف	عامل	مقادیر کیفی و کمی جمع‌آوری‌شده					
۱	سن	کیفی	۲۵-۳۰	۳۰-۳۵	۳۵-۴۰	۴۰-۴۵	۴۵-۵۰
		کمی	۱	۲	۳	۴	۵
۲	احساس غمگینی	کیفی	بله	گاهی	خیر		
		کمی	۱	۱	۰		
۳	تحریک‌پذیری نسبت به نوزاد یا شریک زندگی	کیفی	بله	گاهی	خیر		
		کمی	۱	۱	۰		
۴	مشکلات خواب	کیفی	بله	زیاد	خیر		
		کمی	۱	۱	۰		
۵	مشکلات تمرکز یا تصمیم‌گیری	کیفی	بله	گاهی	خیر		
		کمی	۱	۱	۰		
	پرخوری یا کم‌اشتهایی	کیفی	بله	خیر	اصلاً		

[1] Feeling Sad or Tearful
[2] Irritable Towards Baby & Partner
[3] Trouble Sleeping at Night
[4] Problems Concentrating or Making Decision
[5] Overeating or Loss of Appetite
[6] Feeling Anxious
[7] Feeling of Guilt
[8] Problems of Bonding with Baby
[9] Suicide Attempt

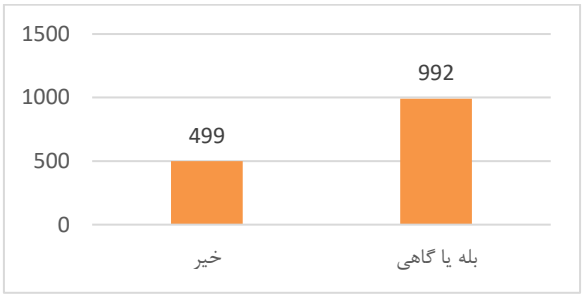
۶		کمی	۱	۰	۰		
۷	احساس اضطراب	کیفی	بله	خیر			
		کمی	۱	۰			
۸	احساس گناه	کیفی	بله	شاید	خیر		
		کمی	۱	۰/۵	۰		
۹	مشکلات رابطه با کودک	کیفی	بله	گاهی	خیر		
		کمی	۱	۱	۰		
۱۰	اقدام به خودکشی	کیفی	بله	بی‌پاسخ	خیر		
		کمی	۱	۰/۵	۰		



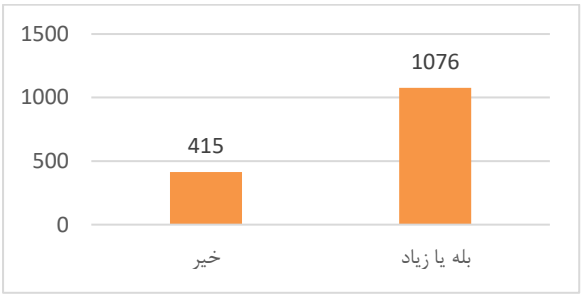
الف) تعداد مادران در هر گروه سنی



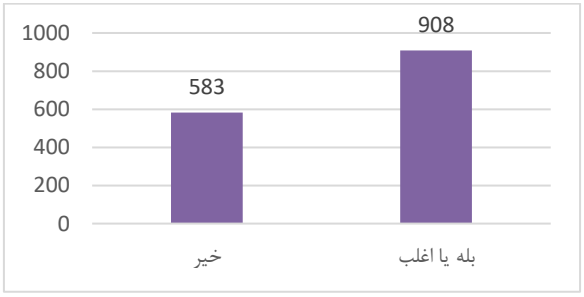
ب) احساس غمگینی



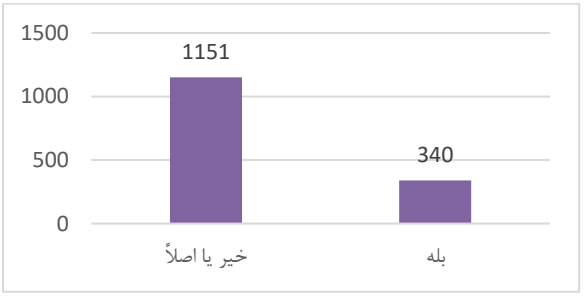
ج) تحریک‌پذیری نسبت به نوزاد یا شریک زندگی



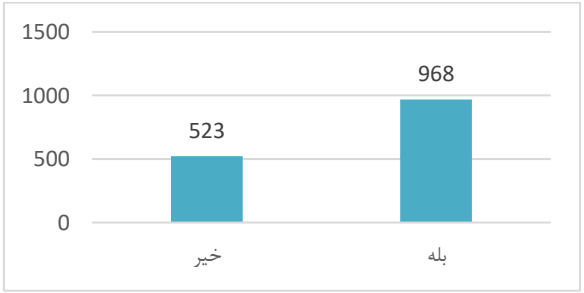
د) مشکلات خواب

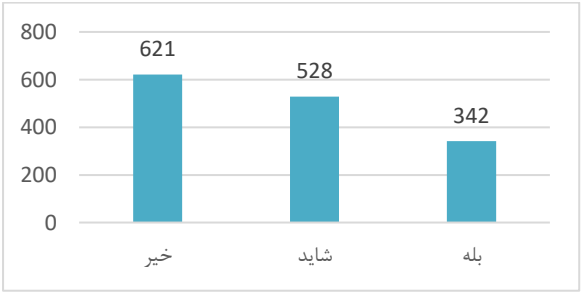


ه) مشکلات تمرکز یا تصمیم‌گیری

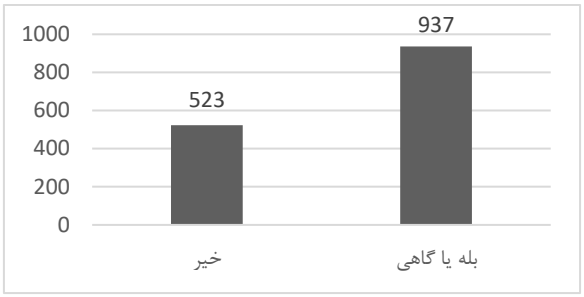


و) پرخوری یا کم‌اشتهایی

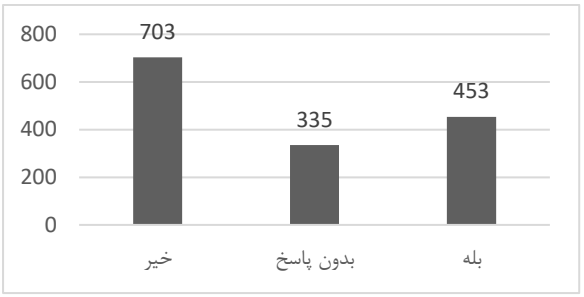




ز) احساس اضطراب



و) احساس گناه



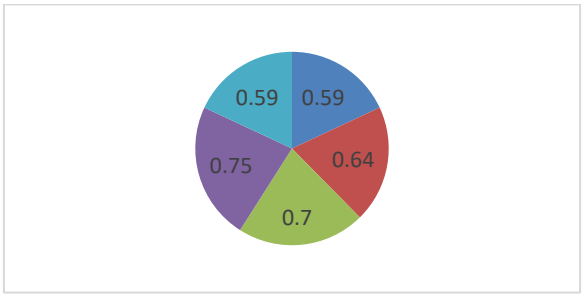
ح) مشکلات ارتباط با کودک

ط) اقدام به خودکشی

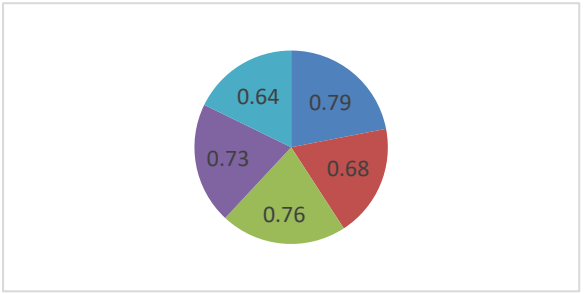
شکل ۱- فراوانی پاسخ مادران به هر عامل

الف) نمودار راداری احتمال پاسخ بله به همه عوامل

ب) احتمال پاسخ بله یا گاهی به احساس غمگینی

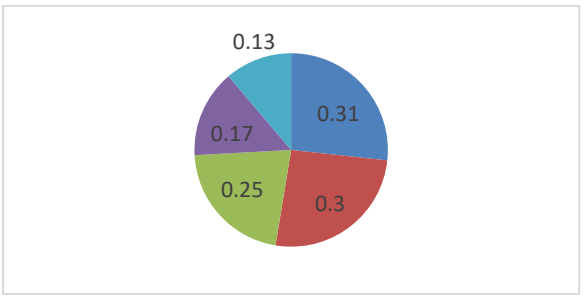


ج) احتمال پاسخ بله یا گاهی به تحریک‌پذیری نسبت به نوزاد یا شریک زندگی

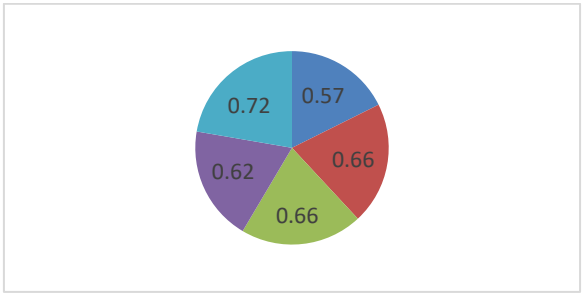


د) احتمال پاسخ بله یا زیاد به مشکلات خواب

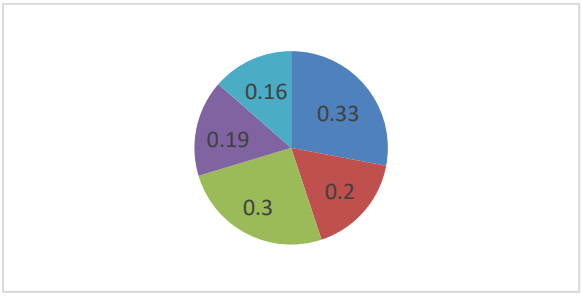
ه) احتمال پاسخ بله یا اغلب به مشکلات تمرکز یا تصمیم‌گیری



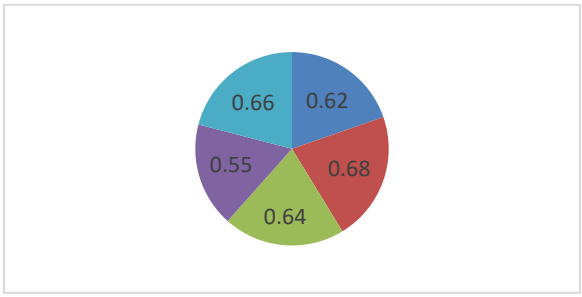
و) احتمال پاسخ بله به پرخوری یا کم‌اشتهایی



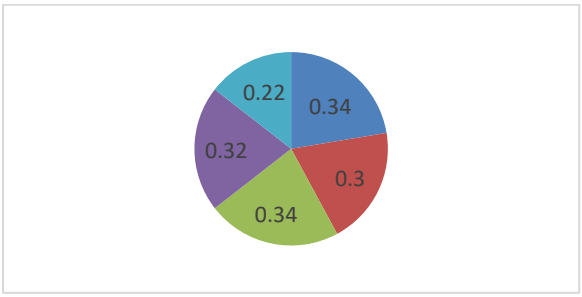
ز) احتمال پاسخ بله به احساس اضطراب



و) احتمال پاسخ بله به احساس گناه



ح) احتمال پاسخ بله یا گاهی به مشکلات ارتباط با کودک



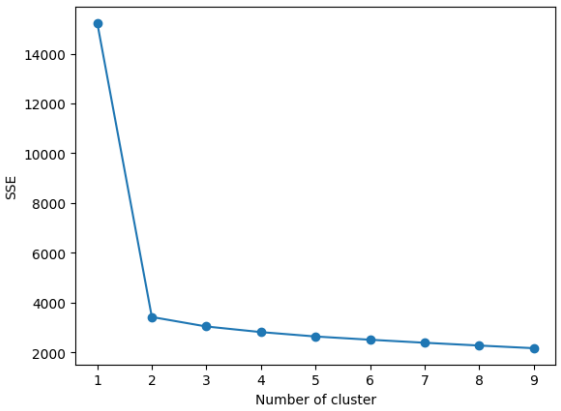
ط) احتمال پاسخ بله به اقدام به خودکشی

شکل ۲- احتمال ارائه پاسخ مثبت به هر سؤال در هر گروه سنی

۳-۲- خوشه‌بندی

به‌منظور خوشه‌بندی داده‌ها، ابتدا باید تعداد خوشه بهینه مشخص شود. برای این منظور از نمودار آرنج[1] استفاده می‌شود. در این نمودار، محور افقی تعداد خوشه‌ها و محور عمودی، خطای مجموع مربعات[2] (SSE) است. همان‌طور که در شکل ۳ مشاهده می‌شود، تعداد خوشه بهینه برای این مجموع داده برابر با دو است؛ جایی که مقدار SSE به حداقل مقدار خود رسیده و با افزایش تعداد خوشه‌ها، SSE به‌طور چشم‌گیری کاهش پیدا نمی‌کند.

در ادامه با استفاده از الگوریتم خوشه‌بندی k-means، پاسخ‌ها (رکوردها) به دو خوشه، خوشه‌بندی می‌شوند. در این خوشه‌بندی، ۶۱۳ پاسخ به خوشه یک و ۸۷۸ پاسخ به خوشه دو تعلق گرفتند. در شکل ۴ قابل‌مشاهده است که به‌طور میانگین، مقادیر عوامل در هر خوشه چه مقدار است. همان‌طور که در شکل ۴ قابل‌مشاهده است، اختلاف مقادیر میانگین عواملی مانند «سن»، «پرخوری یا کم‌اشتهایی»، «احساس گناه» و «خودکشی» کمتر از ۰/۱ است و بنابراین تقریباً تأثیری بر خوشه‌بندی ندارند. سایر ویژگی‌های هر خوشه به شرح جدول ۳ است.



شکل ۳- نمودار آرنج برای پیدا کردن تعداد خوشه بهینه

بنابراین مطابق ویژگی‌های خوشه یک می‌توان نتیجه گرفت، مادرانی که تحریک‌پذیری بیشتری نسبت به نوزاد یا شریک زندگی خود دارند، مشکلات خواب بیشتری دارند. همچنین مطابق ویژگی‌های خوشه دو می‌توان نتیجه گرفت، مادرانی که احساس غمگینی بیشتری دارند، مشکلات تمرکز یا تصمیم‌گیری بیشتری داشته و احساس اضطراب بیشتری دارند و همچنین مشکلات ارتباط با کودک بیشتری دارند.

مطابق نتایج خوشه‌بندی، ازآنجایی‌که عواملی مانند «سن»، «پرخوری یا کم‌اشتهایی»، «احساس گناه» و «خودکشی»، تأثیری بر خوشه‌بندی ندارند، این عوامل از مجموعه داده حذف شدند. پس از حذف این عوامل، الگوریتم درخت تصمیم روی مجموعه داده اجرا شد. درخت تصمیم به‌دست‌آمده که در ۳ سطح خلاصه‌شده است، مطابق با شکل ۵ است. در این درخت در هر گره، حداکثر چهار اطلاعات ارائه‌شده است. اطلاعات اول، شرط گره است و درصورتی‌که برای نمونه‌های موجود در گره، این شرط صادق باشد، آن نمونه‌ها به سمت چپ رفته و در غیر این صورت، به سمت راست خواهند رفت. اطلاعات دوم، متغیر gini است [20] که از رابطه (۱) محاسبه می‌شود و کیفیت تقسیم نمونه‌ها را مشخص کرده و مقداری بین صفر الی ۰/۵ دارد. در رابطه (۱)، متغیر x تعداد نمونه‌هایی است که در این گره در خوشه یک قرار می‌گیرند و متغیر y تعداد نمونه‌هایی است که در این گره در خوشه دو قرار می‌گیرند. همچنین متغیر n تعداد نمونه‌های موجود در گره را نشان می‌دهد.

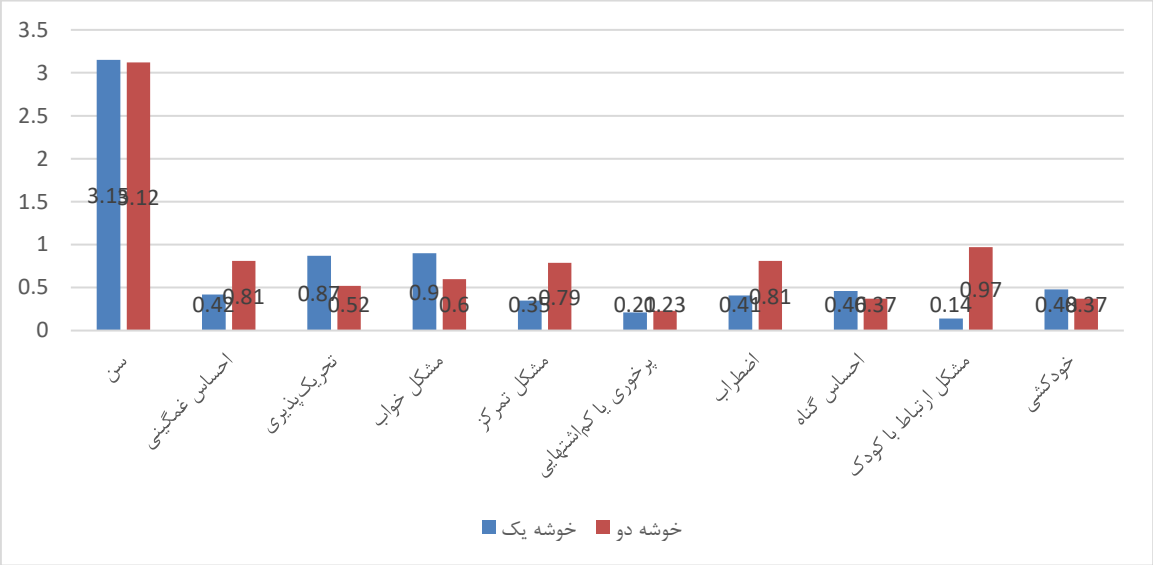
$$gini = 1 - \left(\frac{x}{n}\right)^2 - \left(\frac{y}{n}\right)^2 \qquad (۱)$$

اگر مقدار متغیر gini برابر با صفر باشد، به این معنی است که همه نمونه‌های موجود در این گره در یک خوشه قرار دارند؛ همچنین اگر این مقدار برابر با ۰/۵ باشد، به این معنی است که نصف نمونه‌ها به خوشه یک و نیمی دیگر به خوشه دو تعلق دارند. در شکل ۵، متغیر samples (در رابطه (۱) برابر با n است) تعداد

[1] Elbow

[2] Sum Squared Error

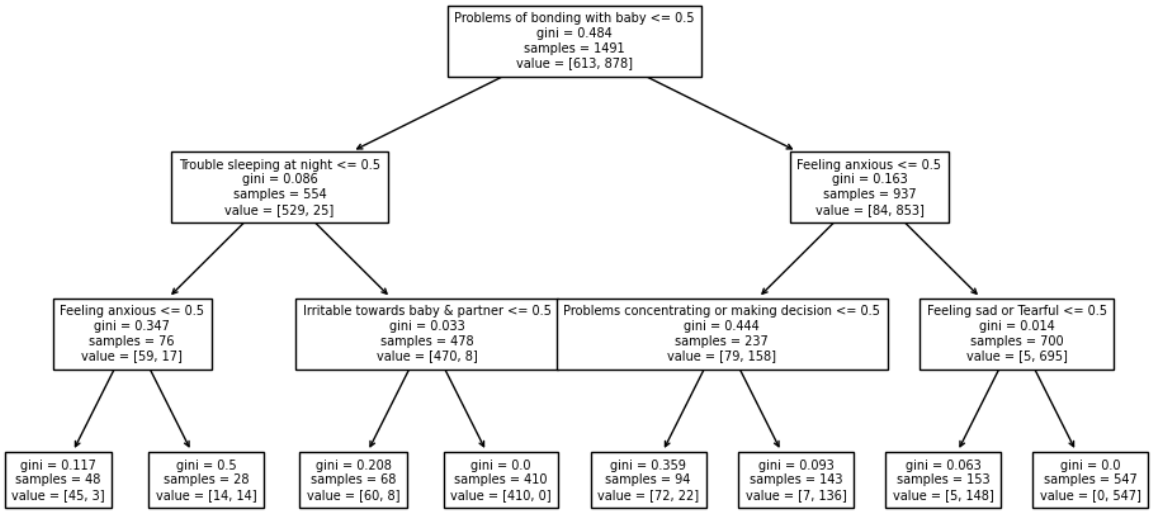
نمونه‌های موجود در گره را نشان می‌دهد. همچنین عدد اول متغیر value (در رابطه (۱) برابر با x است) نشان‌دهنده تعداد نمونه‌هایی است که در این گره در خوشه یک قرار می‌گیرند و به‌طور مشابه، عدد دوم متغیر value (در رابطه (۱) برابر با y است) نشان‌دهنده تعداد نمونه‌هایی است که در این گره در خوشه دو قرار می‌گیرند. برای مثال در گره ریشه در شکل ۵، شرط گره این است که آیا مقدار مشکلات ارتباط با کودک کمتر از ۰/۵ است یا خیر. همان‌طور که در جدول ۲ اشاره شد، اگر مادران به عامل «مشکلات ارتباط با کودک» پاسخ بله یا گاهی می‌دادند، مقدار یک و در غیر این صورت، مقدار صفر ثبت می‌شد. بنابراین ۵۵۴ مادری که این مشکل را ندارند به سمت چپ و ۹۳۷ مادری که این مشکل را دارند، به سمت راست درخت منتقل می‌شوند. تعداد نمونه‌های موجود در این گره (samples) برابر با کل پاسخ‌های جمع‌آوری‌شده (۱۴۹۱ پاسخ) است که ۶۱۳ مورد از آن‌ها به خوشه یک و ۷۸۷ مورد از آن‌ها به خوشه دو تعلق دارند. بنابراین مقدار gini برابر با $0.484 = \left(\frac{613}{1491}\right)^2 - \left(\frac{878}{1491}\right)^2$ محاسبه‌شده است.



شکل ۴- میانگین مقادیر عوامل در هر خوشه

جدول ۳- ویژگی‌های پاسخ‌های موجود در خوشه یک و دو

خوشه یک	خوشه دو
احساس غمگینی کمتری دارد.	احساس غمگینی بیشتری دارد.
تحریک‌پذیری بیشتری نسبت به نوزاد یا شریک زندگی دارد.	تحریک‌پذیری کمتری نسبت به نوزاد یا شریک زندگی دارد.
مشکلات خواب بیشتری دارد.	مشکلات خواب کمتری دارد.
مشکلات تمرکز یا تصمیم‌گیری کمتری دارد.	مشکلات تمرکز یا تصمیم‌گیری بیشتری دارد.
احساس اضطراب کمتری دارد.	احساس اضطراب بیشتری دارد.
مشکلات ارتباط با کودک کمتری دارد.	مشکلات ارتباط با کودک بیشتری دارد.



شکل ۵- درخت تصمیم ۳ سطحی روی مجموعه داده

از شکل ۵ اطلاعات بسیاری می‌توان برداشت کرد که در زیر به بعضی از آن‌ها اشاره‌شده است:

- از ۱۴۹۱ مادر، ۸۷۸ مادر به خوشه دو تعلق دارند که وجه مشترک ۵۴۷ نفر از آن‌ها، پاسخ بله به سه عامل مشکلات ارتباط با کودک، احساس اضطراب و احساس غمگینی است. به‌عبارت‌دیگر، مادرانی که مشکلات ارتباط با کودک دارند، احساس اضطراب و غمگینی دارند.
- از ۱۴۹۱ مادر، ۶۱۳ مادر به خوشه یک تعلق دارند که وجه مشترک ۴۱۰ نفر از آن‌ها، پاسخ خیر به عامل مشکلات ارتباط با کودک، و پاسخ بله به دو عامل مشکلات خواب و تحریک‌پذیری نسبت به نوزاد یا شریک زندگی است. به‌عبارت‌دیگر، مادرانی که مشکلات ارتباط با کودک ندارند، مشکلات خواب و تحریک‌پذیری نسبت به نوزاد یا شریک زندگی دارند.

۳-۳- تفسیر و ارائه نتایج

در این بخش، نتایج به‌دست‌آمده از فرآیند داده‌کاوی مورد تفسیر و ارزیابی قرار می‌گیرند. همان‌طور که پیش‌تر توضیح داده شد، از تاریخ ۱۴ الی ۱۵ ژوئن ۲۰۲۲، از ۱۵۰۳ مادری که به‌تازگی در یک بیمارستان زایمان کرده بودند، ۱۰ سؤال مطرح شد تا در مورد علائم افسردگی پس از زایمان اطلاعاتی به دست آید. این سؤالات شامل «سن»، «احساس غمگینی»، «تحریک‌پذیری نسبت به نوزاد یا شریک زندگی»، «مشکلات خواب»، «مشکلات تمرکز یا تصمیم‌گیری»، «پرخوری یا کم‌اشتهایی»، «احساس اضطراب»، «احساس گناه»، «مشکلات ارتباط با کودک»، و «اقدام به خودکشی» بودند. پس از حذف پاسخ‌نامه‌های غیرکامل، ۱۴۹۱ پاسخ باقی ماند و در ادامه، تحلیل‌ها روی این رکوردها انجام شد. از تحلیل داده آماری پاسخ‌های دریافت‌شده، این اطلاعات به دست آمد که بیشتر مادران فقط به عواملی مانند «پرخوری یا کم‌اشتهایی» و «احساس گناه»، و «اقدام به خودکشی» پاسخ خیر داده‌اند و به سایر عوامل در بیشتر موارد پاسخ بله داده‌اند. بنابراین از ۹ مشکلی که مادران می‌توانستند پس از زایمان داشته باشند، در بیشتر مادران، ۶ مشکل مشاهده‌شده است. همچنین دریافت شد، مادران که در گروه سنی ۲۵ الی ۳۰ سال قرار دارند در ۴ عامل شامل مشکلات خواب، پرخوری یا کم‌اشتهایی، احساس گناه، و اقدام به خودکشی بیشترین احتمال مشکل را دارند. بنابراین برخلاف تصور عموم، زایمان در سنین پایین می‌تواند مشکلات بیشتری را برای مادران ایجاد کند. در ادامه با توجه به اینکه داده‌ها برچسب‌گذاری نشده بودند، از الگوریتم بدون نظارت k-means برای خوشه‌بندی پاسخ‌های دریافتی استفاده شد. قبل از آن طبق نمودار آرنج، تعداد خوشه بهینه برابر با دو محاسبه شد. در زمان فرآیند خوشه‌بندی مشخص شد که عواملی مانند «سن»، «پرخوری یا کم‌اشتهایی»، «احساس گناه» و «خودکشی»، تأثیری بر خوشه‌بندی ندارند، و بنابراین این عوامل از مجموعه داده حذف شدند. پس از فرآیند خوشه‌بندی، مادرانی که تحریک‌پذیری بیشتری نسبت به نوزاد یا شریک زندگی خود داشتند، و مشکلات خواب بیشتری داشتند، در خوشه یک قرار گرفتند. همچنین مادرانی که احساس غمگینی، مشکلات تمرکز یا تصمیم‌گیری، احساس اضطراب، و مشکلات ارتباط با کودک بیشتری داشتند، در خوشه دو قرار گرفتند. در این مرحله، شماره خوشه هر رکورد داده مشخص شد و بنابراین رکوردها دارای برچسب شدند. در ادامه با استفاده از الگوریتم بانظارت درخت تصمیم، عوامل مشخص‌کننده خوشه رکوردها، تعیین شدند. مطابق با درخت تصمیم به وجود آمده، از ۱۴۹۱ مادر، ۸۷۸ مادر به خوشه دو تعلق داشتند که وجه مشترک ۵۴۷ نفر از آن‌ها، پاسخ بله به سه عامل مشکلات ارتباط با کودک، احساس اضطراب و احساس غمگینی بود. به‌عبارت‌دیگر، مادرانی که مشکلات ارتباط با کودک دارند، احساس اضطراب و غمگینی دارند. همچنین از ۱۴۹۱ مادر، ۶۱۳ مادر به خوشه یک تعلق داشتند که وجه مشترک ۴۱۰ نفر از آن‌ها، پاسخ خیر به عامل مشکلات ارتباط با کودک، و پاسخ بله به دو عامل مشکلات خواب و تحریک‌پذیری نسبت به نوزاد یا شریک زندگی بود. به‌عبارت‌دیگر، مادرانی که مشکلات ارتباط با کودک ندارند، مشکلات خواب و تحریک‌پذیری نسبت به نوزاد یا شریک زندگی دارند.

۴- نتیجه‌گیری و کارهای آینده

فرآیند داده‌کاوی در حوزه پزشکی و تشخیص بیماری‌ها در مقالات بسیاری انجام‌شده است ولی بسیاری از آن‌ها، بیماری‌های روانی مانند افسردگی را موردبررسی قرار نداده‌اند. در معدود کارهایی که از روش‌های داده‌کاوی در پیش‌بینی و تشخیص بیماری افسردگی استفاده کرده‌اند، افسردگی پس از زایمان را موردمطالعه قرار نداده‌اند. در این مقاله، داده‌های جمع‌آوری‌شده از تاریخ ۱۴ الی ۱۵ ژوئن ۲۰۲۲ از وضعیت ۱۵۰۳ مادر بین ۲۵ تا ۵۰ سال پس از زایمان موردبررسی قرار گرفت. ابتدا تحلیل آماری از پاسخ مادران به ۱۰ سؤال شامل «سن»، «احساس غمگینی»، «تحریک‌پذیری نسبت به نوزاد یا شریک زندگی»، «مشکلات خواب»، «مشکلات تمرکز یا تصمیم‌گیری»، «پرخوری یا کم‌اشتهایی»، «احساس اضطراب»، «احساس گناه»، «مشکلات ارتباط با کودک»، و «اقدام به خودکشی» ارائه شد. سپس با استفاده از الگوریتم k-means، مادران بر اساس پاسخ‌هایشان به دو خوشه تقسیم شدند. سپس با استفاده از الگوریتم درخت تصمیم، نحوه رسیدن به هر خوشه از طریق سوال‌ها مشخص شد. با نتایج به‌دست‌آمده از این مقاله می‌توان با مشاهده یک یا چند عامل در یک مادر جدید، خطراتی که او در آینده با آن‌ها مواجه خواهد شد را شناسایی کرد. نتایج به‌دست‌آمده منجر به بهبود مراقبت‌های بهداشتی و درمانی برای مادران خواهد شد.

از عوامل مهم دیگری که می‌تواند بر افسردگی پس از زایمان مادران تأثیرگذار باشد، وضعیت اقتصادی، اجتماعی و غیره است. همچنین وضعیت جسمی و روحی مادران قبل از بارداری نیز می‌تواند برشدت یا ضعف این بیماری موثر باشد. به‌عنوان کار آینده می‌توان این عوامل را نیز در بررسی‌ها در نظر گرفت. همچنین می‌توان از الگوریتم‌های جدیدتری که مبتنی بر یادگیری گروهی هستند، برای پیش‌بینی و تشخیص بیماری‌ها استفاده کرد.

۵- مآخذ

[1] A. H. Shurrab and A. Y. A. Maghari, "Blood diseases detection using data mining techniques," in *8th International Conference on Information Technology (ICIT)*, Amman, Jordan, 17-18 May 2017, pp. 625-631.

[2] G. Schuh, G. Reinhart, J.P. Prote, F. Sauermann, J. Horsthofer, F. Oppolzer and D. Knoll, "Data Mining Definitions and Applications for the Management of Production Complexity," *Procedia CIRP,* vol. 81, pp. 874-879, 2019.

[3] S. N. Nezhad, M. H. Zahedi and E. Farahani, "Detecting diseases in medical prescriptions using data mining methods," *BioData Min,* vol. 15, no. 29, pp. 1-19, 2022.

[4] S. Shorey, C. Y. I. Chee, E. D. Ng, Y. H. Chan, W. W. S. Tam and Y. S. Chong, "Prevalence and incidence of postpartum depression among healthy mothers: A systematic review

and meta-analysis," *Journal of Psychiatric Research,* vol. 104, pp. 235-248, 2018.

[5] M. R. Keyvanpour, Z. Karimi Zandian and N. Mottaghi, "BRTSRDM: Bi-Criteria Regression Test Suite Reduction based on Data Mining," *Journal of AI and Data Mining,* vol. 11, no. 2, pp. 161-186, 2023.

[6] M. Z. Abedin, P. Hajek, T. Sharif, M. S. Satu and M. Imran Khan, "Modelling bank customer behaviour using feature engineering and classification techniques," *Research in International Business and Finance,* vol. 65, p. 101913, 2023.

[7] K. Chen, F. Abtahi, J.J. Carrero, C. Fernandez-Llatas and F. Seoane, "Process mining and data mining applications in the domain of chronic diseases: A systematic review," *Artificial Intelligence in Medicine,* vol. 144, p. 102645, 2023.

[8] A. K. Srivastava, K. Jeberson and W. Jeberson, "A systematic review on Data Mining Application in Parkinson's disease," *Neuroscience Informatics,* vol. 2, no. 4, p. 100064, 2022.

[9] S. Z. Hosseini, R. Radfar, A. Nasiripour and A. Rajabzadeh Ghatary, "Designing an optimal diagnosis algorithm based on IoT for Covid-19," *Signal and Data Processing,* vol. 20, no. 3, pp. 87-102, 2023.

[10] M. Nazari, H. Emami, R. Rabiei, A. Hosseini and S. Rahmatizadeh, "Detection of Cardiovascular Diseases Using Data Mining Approaches: Application of an Ensemble-Based Model," *Cognitive Computation,* vol. 16, pp. 2264-2278, 2024.

[11] P. Chhabra and R. Madaan, "Data mining concepts in healthcare with discussion on prediction of diseases," in *International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COM-IT-CON)*, Faridabad, India, 26-27 May 2022, pp. 71-77.

[12] S. Zolfaghari, S. Suravee, D. Riboni and K. Yordanova, "Sensor-Based Locomotion Data Mining for Supporting the Diagnosis of Neurodegenerative Disorders: A Survey," *ACM Computing Surveys,* vol. 56, no. 1, pp. 1-36, 2023.

[13] M. Roostaee and R. Meidanshahi, "Hidden Pattern Discovery on Clinical Data: an Approach based on Data Mining Techniques," *Journal of AI and Data Mining,* vol. 11, no. 3, pp. 343-355, 2023.

[14] H. M. van Loo, T. B. Bigdeli, Y. Milaneschi, S. H. Aggen and K. S. Kendler, "Data mining algorithm predicts a range of adverse outcomes in major depression," *Journal of Affective Disorders,* vol. 276, pp. 945-953, 2020.

[15] W. Narkbunnum and K. Wisaeng, "Prediction of Depression for Undergraduate Students Based on Imbalanced Data by Using Data Mining Techniques," *Applied System Innovation,* vol. 5, no. 6, p. 120, 2022.

[16] S. Arora, S. Bindra, M. Ahmad and T. Ahmad, "An Analysis of Depression Detection Model Applying Data Mining Approaches Using Social Network Data," in *Fourth International Conference on Electrical, Computer and Communication Technologies (ICECCT)*, Erode, India, 15-17 Sep. 2021, pp. 1-7.

[17] X. Yang, R. McEwen, L. R. Ong and M. Zihayat, "A big data analytics framework for detecting user-level depression from social networks," *International Journal of Information Management,* vol. 54, p. 102141, 2020.

[18] M. W. Moreira, J. J. Rodrigues, N. Kumar, K. Saleem and I. V. Illin, "Postpartum depression prediction through pregnancy data analysis for emotion-aware smart systems," *Information Fusion,* vol. 47, pp. 23-31, 2019.

[19] "PostPartum Depression," Kaggle, 2022. [Online]. Available: https://www.kaggle.com/datasets/parvezalmuqtadir2348/postpartum-depression. [Accessed 08 08 2024].

[20] T. Daniya, M. Geetha and S. Kumar K Dr, "Classification and regression trees with gini index," *Advances in Mathematics Scientific Journal,* vol. 9, no. 10, pp. 1857-8438, 2020.

معرفی نویسندگان

علی نقاش اسدی استادیار گروه مهندسی کامپیوتر دانشکده فنی فومن، دانشکدگان فنی دانشگاه تهران است. موضوعات پژوهشی موردعلاقه ایشان، مدل‌سازی و شبیه‌سازی، شبکه‌های پتری، هوش مصنوعی، داده‌کاوی و یادگیری ماشین است. نشانی رایانامه ایشان عبارت است از:

naghashasadi@ut.ac.ir

فاطمه عاصی آتشکاهی فارغ‌التحصیل رشته مهندسی کامپیوتر از دانشکده فنی فومن، دانشکدگان فنی دانشگاه تهران است. ایشان در حال حاضر دانشجوی کارشناسی ارشد رشته مهندسی

کامپیوتر گرایش شبکه‌های کامپیوتری در دانشکده مهندسی کامپیوتر دانشگاه صنعتی خواجه‌نصیرالدین طوسی است. موضوعات پژوهشی موردعلاقه ایشان، هوش مصنوعی، یادگیری ماشین، داده‌کاوی، فناوری‌های وب و IOT است. نشانی رایانامه ایشان عبارت است از:

fatemeh.asiatashkahi@email.kntu.ac.ir