مجله‌ی بررسی‌های آمار رسمی ایران
سال ۲۷، شماره‌ی ۲، پاییز و زمستان ۱۳۹۵، صص ۱۲۱-۱۴۳

داده‌کاوی اطلاعات طرح نیروی کار به‌منظور استخراج قاعده‌های پیوند در جمعیت غیر فعال اقتصادی

سمیه آهنگر[†]، مجید خلیلیان[‡,*] و نریمان یوسفی[†]

[†] مرکز آمار ایران
[‡] دانشگاه آزاد اسلامی

چکیده : امروزه با افزایش سرعت در رشد فناوری و روش‌های گرداوری داده، اغلب با حجم زیادی از داده در سازمان‌ها روبه‌رو هستیم. با توجه به نقش مهم فرایندهای تحلیل و استخراج دانش از داده، به‌کارگیری علم داده‌کاوی به‌عنوان مهم‌ترین فناوری برای بهره‌برداری مؤثر، صحیح و سریع از داده‌ها رو به افزایش است. یکی از روش‌های کاربردی داده‌کاوی، فن قاعده‌های پیوند است. این فن، رابطه‌های اقلام داده‌ای را مورد تحلیل قرار می‌دهد و وابستگی‌های متقابل بین مجموعه‌ی بزرگی از اقلام داده‌ای را در قالب الگوها و قاعده‌های استخراجی نشان می‌دهد. در این مقاله با به کارگیری این فن بر روی داده‌های طرح آمارگیری نیروی کار سال‌های ۱۳۸۸-۱۳۹۳، دانش تولیدشده حول خصیصه‌های جمعیت غیر فعال اقتصادی ارایه خواهد شد. طرح آمارگیری نیروی کار یکی از مهم‌ترین و جامع‌ترین منبع‌های اطلاعات بازار کار کشور است که بر اساس آن نرخ مشارکت اقتصادی و نرخ بیکاری کشور محاسبه می‌شود. بر اساس نتیجه‌های این طرح در سال‌های اخیر جمعیت غیر فعال کشور افزایش و نرخ مشارکت اقتصادی کاهش یافته است. فهم چرایی این موضوع دغدغه‌ی بسیاری از متخصصین، کارشناسان و برنامه‌ریزان بازار کار است. این تحقیق به‌منظور بررسی و تحلیل این مسئله، به شناسایی رابطه‌های میان ویژگی‌های افراد و وضعیت فعالیت اقتصادی آنان می‌پردازد و در این زمینه الگوریتم Apriori به‌عنوان یکی از روش‌های مطرح در فن قاعده‌های پیوند استفاده شده است.

واژگان کلیدی: داده‌کاوی؛ فن قاعده‌های پیوند؛ جمعیت غیر فعال اقتصادی؛ نرخ مشارکت اقتصادی.

---

* نویسنده‌ی عهده‌دار مکاتبات
دریافت:۱۳۹۴/۹/۲۸، پذیرش: ۱۳۹۷/۴/۱۶.

۱- مقدمه

پیشرفت فناوری اطلاعات و سامانه‌های یکپارچه‌ی اطلاعاتی منجر به ایجاد داده‌انبارها شده است. استفاده از فناوری و به‌کارگیری فن داده‌کاوی به‌عنوان یکی از راهکارهای برتر، تحول بزرگی را در تحلیل و بررسی داده‌ها ایجاد کرده است. به‌عبارتی، تحلیل داده و تصمیم‌گیری بر مبنای فن‌های داده‌کاوی نقش مهمی در رشد و ارتقای فرایندها دارد. در این‌باره می‌توان الگوها و ساختارهای جالب و با ارزش را از مجموعه‌ی وسیعی از داده‌ها استخراج کرد.

با این اوصاف، به‌کارگیری فن‌های داده‌کاوی در سازمان‌ها و مراکز آماری، به‌منظور تولید دانش در راستای تحلیل اطلاعات، مهم است. هر چند در مرکز آمار ایران سابقه‌ی استفاده از الگوریتم‌های داده‌کاوی بر روی بانک اطلاعاتی طرح هزینه و درامد خانوارهای شهری و روستایی وجود دارد، اما اهمیت و حساسیت شاخص‌های بازار کار از جمله نرخ مشارکت اقتصادی، به‌کارگیری این فن بر روی داده‌های طرح نیروی کار را نیز ضروری می‌سازد. طرح آمارگیری نیروی کار منبع رسمی تولید و انتشار شاخص‌های بازار کار ایران است. این طرح منطبق بر آخرین توصیه‌های سازمان بین‌المللی کار است و هدف آن محاسبه‌ی شاخص‌های بازار کار، از جمله نرخ مشارکت اقتصادی و نرخ بیکاری است [۳]. بر اساس نتایج این طرح، در سال‌های اخیر نرخ مشارکت اقتصادی کشور کاهش یافته است و علاوه بر اثرات زیان‌بار آن در حوزه‌های اقتصادی، اجتماعی و فرهنگی، به‌عنوان چالش اساسی در حوزه‌ی آمارهای بازار کار مورد بررسی است.

پژوهش‌های انجام‌شده در این زمینه به‌طور معمول مبتنی بر روش‌های سنتی و نظریه محور تحلیل داده‌های بازار کار است. علاوه برآن می‌توان در رویکرد نوین، امکان انجام پژوهش‌های مبتنی بر استخراج دانش و تحت مفهوم داده‌کاوی را نیز در اختیار داشت.

با وجود تنوع در راهکارهای داده‌کاوی، یکی از فن‌های مؤثر، یافتن اقلام متناوب به‌منظور کشف قاعده‌های پیوند است. الگوریتم Apriori یکی از مهم‌ترین روش‌ها برای تولید مجموعه اقلام متناوب در یک بانک داده‌ی تراکنشی است. از این‌رو، در این تحقیق اجرای این الگوریتم بر روی مجموعه داده‌های نیروی کار، به‌عنوان استخراج قاعده‌های پیوند و الگوهای پنهان مورد استفاده قرار گرفته است.

۲- مبانی نظری و فنی پژوهش

۲-۱- مفاهیم طرح نیروی کار

آمارهای مربوط به ساختار نیروی کار، مبنایی برای ارزیابی و تحلیل سیاست‌های اقتصادی، تحلیل وضعیت اجتماعی و فرهنگی یک کشور محسوب می‌شود. همچنین شاخص‌های بازار کار، تعیین‌کننده‌ی جایگاه کشور در مقایسه با سایر کشورها است. از این‌رو، در این طرح از مفاهیم و شاخص‌های استانداردشده توسط سازمان بین‌المللی کار استفاده شده است. مفاهیم مورد استفاده ارجاع داده‌شده در این تحقیق به شرح زیر است [۳]:

- **شاغل:** همه‌ی افراد ۱۰ ساله و بیش‌تر که در طول هفته‌ی مرجع، طبق تعریف کار، حداقل یک ساعت کار کرده و یا بنا به دلایلی موقتاً کار خود را ترک کرده باشند، شاغل محسوب می‌شوند. به‌طور عمده شاغلان شامل دو گروه مزدبگیران و خوداشتغالان هستند.
- **کار:** آن دسته از فعالیت‌های اقتصادی (فکری و یا بدنی) که به‌منظور کسب درآمد (نقدی و یا غیر نقدی) صورت پذیرد و هدف آن تولید کالا یا ارایه‌ی خدمت باشد، کار محسوب می‌شود.
- **بیکار:** بیکار به همه‌ی افراد ۱۰ ساله و بیش‌تر اطلاق می‌شود که:

۱- در هفته‌ی مرجع فاقد کار باشند (دارای اشتغال مزدبگیری یا خوداشتغالی نباشند).

۲- در هفته‌ی مرجع یا هفته‌ی بعد از آن آماده برای کار باشند (برای اشتغال مزدبگیری یا خوداشتغالی آماده باشند).

۳- در هفته‌ی مرجع و سه هفته قبل از آن جویای کار باشند (اقدامات مشخصی را به‌منظور جستجوی اشتغال مزدبگیری و یا خوداشتغالی به‌عمل آورده باشند).

٤- افرادی که به‌دلیل آغاز به کار در آینده یا انتظار بازگشت به شغل قبلی جویای کار نبوده، ولی فاقد کار و آماده برای کار بوده‌اند نیز بیکار محسوب می‌شوند.

- **جمعیت فعال اقتصادی:** همه‌ی افراد ۱۰ ساله و بیش‌تر که در هفته‌ی تقویمی قبل از هفته‌ی آمارگیری (هفته‌ی مرجع) طبق تعریف کار، در تولید کالا و خدمات مشارکت داشته (شاغل) و یا از قابلیت مشارکت برخوردار بوده‌اند (بیکار)، جمعیت فعال از نظر اقتصادی محسوب می‌شوند.

- **جمعیت غیر فعال اقتصادی:** همه‌ی افراد ۱۰ ساله و بیش‌تر که در طی هفته‌ی مرجع طبق تعریف در هیچ‌یک از دو گروه شاغلان و بیکاران قرار نمی‌گیرند.
- **اشتغال ناقص زمانی:** افراد دارای اشتغال ناقص زمانی شامل همه‌ی شاغلین مزدبگیر یا خوداشتغالی است که در هفته‌ی مرجع، حاضر در سر کار یا غایب موقت بوده، به‌دلایل اقتصادی کم‌تر از ۴۴ ساعت کار کرده، خواهان و آماده برای انجام کار اضافی در هفته‌ی مرجع هستند.
- **باسواد:** کسانی که می‌توانند به زبان فارسی یا هر زبان دیگری متن ساده‌ای را بخوانند و بنویسند، خواه مدرک رسمی داشته یا نداشته باشند، باسواد تلقی می‌شوند.
- **بی‌سواد:** کسانی که نمی‌توانند بخوانند و بنویسند و کسانی که فقط می‌توانند بخوانند، بی‌سواد تلقی می‌شوند.
- **در حال تحصیل:** به فردی گفته می‌شود که طبق برنامه‌های رسمی آموزشی کشور به تحصیل اشتغال دارد.
- **فارغ‌التحصیل:** به فردی گفته می‌شود که آموزش‌های علمی و عملی مربوط به یک دوره‌ی آموزش رسمی را فرا گرفته و حائز شرایط دریافت گواهینامه آن دوره شناخته شده است.
- **سرپرست خانوار:** عضوی از خانوار است که اعضای خانوار وی را به این عنوان معرفی می‌کنند. در مواردی که اعضای خانوار کسی را به‌عنوان سرپرست معرفی نکنند، مسن‌ترین فرد به‌عنوان سرپرست خانوار شناخته می‌شود. در خانوارهای یک‌نفره، همان فرد سرپرست خانوار به شمار می‌آید.

۲-۲-مفهوم داده‌کاوی

واژه‌ی کاوش یا کاویدن[1] به معنای فرایند یافتن مجموعه‌ی کوچک و با ارزش از حجم زیاد ماده‌ی خام است. ادغام این کلمه با کلمه‌ی داده[2]، بر جستجویی عمیق از داده‌های قابل دسترس با حجم زیاد، برای یافتن اطلاعات مفید که قبلاً نهفته بودند، تأکید دارد [۷]. به‌عبارتی داده‌کاوی، مجموعه‌ای از روش‌ها در فرایند کشف دانش است که برای تشخیص الگوها و کشف ساختارهای جالب توجه و با ارزش از داخل مجموعه‌ی وسیعی از داده‌ها مورد استفاده قرار می‌گیرد [۹]، [۱۰].

۱-۲-۲- فن‌های داده‌کاوی

فن‌های مختلف داده‌کاوی در قالب الگوریتم‌های هر بخش و متناسب با نوع عملیاتی که انجام می‌دهند، دارای دو رویکرد پیش‌گویی و توصیف است. در رویکرد پیش‌گوگری[۳]، هدف ایجاد مدلی است که بتوان با آن پیش‌گویی و تشخیص موارد نامعلوم را انجام داد. از طرفی در رویکرد توصیف‌گری[۴] هدف تشریح الگوهای رفتاری متغیرها بوده و یک شناخت از داده‌ها توسط الگوها و روابط بین آن‌ها ارایه می‌شود [۱۰]. فن مورد استفاده در این تحقیق، از نوع توصیف‌گری است.

فن قاعده‌های پیوند: از میان فن‌های داده‌کاوی، قاعده‌های پیوند[۵] یکی از روش‌های مورد توجه است. این روش در سال ۱۹۹۳ توسط اگراوال مطرح شد [۵]. مثال متداول در رابطه با کشف قاعده‌های پیوند «تحلیل سبد خرید» است. در این فرایند با توجه به اقلام مختلفی که مشتریان در سبد خریدشان قرار می‌دهند، عادات و رفتار خرید مشتریان مورد تحلیل قرار می‌گیرد و الگوهای موجود در اقلام خریداری‌شده کشف می‌شود [۱۰]. این فن، رابطه‌ها و وابستگی‌های متقابل بین مجموعه‌ی بزرگی از اقلام داده‌ای را نشان می‌دهند. قالب کلی قاعده‌های پیوند به‌صورت $x \to y$ بوده که نشان‌دهنده‌ی نوعی پیشامد همزمان بین اقلام x و y است (x و y دو مجموعه قلم دلخواه از داده‌انبار هستند) x و y به ترتیب مقدم[۶] و تالی[۷] قانون نامیده می‌شوند.

معیارهای مختلفی برای ارزیابی میزان صحت و ارزشمندی قاعده‌های ارایه شده‌اند. معروف‌ترین و پرکاربردترین این معیارها دو پارامتر کمینه‌ی ضریب پشتیبانی[۸] و کمینه‌ی ضریب اطمینان[۹] است که به‌صورت رابطه‌های زیر تعریف می‌شود:

ضریب پشتیبان[۱۰]: ضریب پشتیبانی قاعده‌ی پیوند $\boldsymbol{x \to y}$ در پایگاه داده‌ی تراکنشی یک نسبت است که در قالب رابطه‌ی (۱) به‌صورت زیر است.

$$support(x \to y) = \frac{\text{تعداد مجموعه اقلام شامل } x,y}{\text{تعداد همه‌ی مجموعه اقلام}} \qquad (۱)$$

در واقع ضریب پشتیبان، درصد مجموعه‌ی تراکنش‌هایی است که شامل هر دو زیر مجموعه‌ی x و y باشد. ضریب پشتیبانی بیش‌تر، نشان‌دهنده‌ی پرتکرار بودن قانون است.

یعنی احتمال رخداد مجموعه‌ی اقلام بیش‌تر خواهد بود. کمینه‌ی پشتیبانی توسط مشتری و بر طبق نیاز او به‌عنوان کمینه‌ی پشتیبان تعیین می‌شود [۵]، [۱۰].

ضریب اطمینان[۱۱]: نسبت میان تعداد دفعات تکرار همزمان $\boldsymbol{x}$ و $\boldsymbol{y}$ در تراکنش‌ها به تعداد تراکنش‌های شامل $\boldsymbol{x}$ است و در قالب رابطه‌ی (۲) نمایش داده می‌شود. ضریب اطمینان، درصد تراکنش‌هایی است که $\boldsymbol{x}$ و $\boldsymbol{y}$ هر دو با هم ظاهر شده‌اند و در واقع یک احتمال شرطی است.

$$conf(x \rightarrow y) = \frac{\text{تعداد تراکنش‌های شامل } x,y}{\text{تعداد تراکنش‌های شامل } x} \qquad (۲)$$

در نهایت، قواعدی مورد قبول واقع می‌شوند که مقادیر قابل قبولی برای هر دو معیار فوق داشته باشند [۵]، [۱۰].

۲-۲-۲-الگوریتم Apriori

الگوریتم Apriori، یکی از الگوریتم‌های پایه‌ای و پرکاربرد در حوزه‌ی استخراج قاعده‌های پیوند است [۴]. این الگوریتم دو گام دارد:

گام اول، تولید مجموعه‌ی قلم‌های متناوب: در مرحله‌ی اول، با جستجوی کامل پایگاه داده‌ها، اقلام تکرارشونده‌ی تک‌عضوی شناسایی می‌شوند. سپس با استفاده از اقلام تکرارشونده‌ی تک‌عضوی، شروع به ساختن قلم‌های نامزد دو عضوی می‌کند و از بین آن‌ها مجموعه‌ی قلم‌های متناوب دو عضوی شناسایی می‌شوند و این روال تا زمانی که با مجموعه‌ی قلم‌ها، نامزد بیش‌تری نتوان تولید کرد، پایان می‌یابد.

گام دوم، تولید قاعده‌های پیوند: بعد از کشف تمام قلم‌های پرتکرار از مجموعه‌ی داده‌ها گام بعدی تولید قاعده‌ها است. برای تولید قاعده‌های پیوند از مجموعه‌ی قلم‌های متناوب استفاده می‌شود. در اکثر موارد، تنها قاعده‌هایی جالب و مفید هستند که شامل قلم‌هایی با دفعات تکرار زیاد باشند و درصد رخداد تراکنش‌های آن از کمینه‌ی ضریب اطمینان[۱۲] تعیین‌شده بیش‌تر باشد [۵]، [۴].

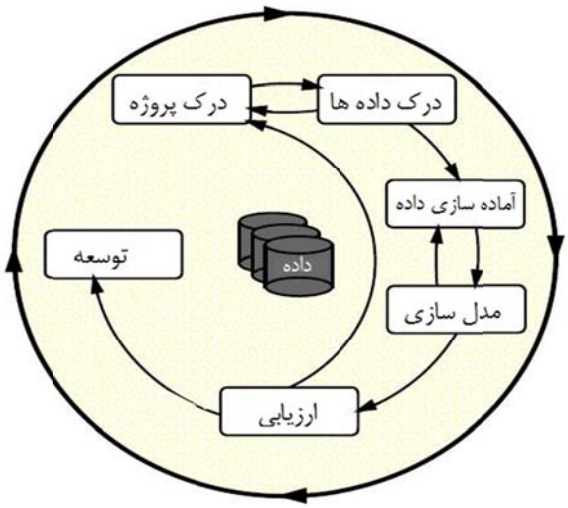
۳- روش پژوهش

۱-۳- انتخاب مدل فرایند

امروزه پروژه‌های داده‌کاوی سهم مهمی در تحلیل داده‌های موجود و کمک به تصمیم‌گیری در این زمینه می‌کند [۱۱]. پروژه‌های داده‌کاوی نیز مانند سایر پروژه‌های نرم‌افزاری نیاز به یک مدل فرایند استاندارد برای انجام فعالیت‌های مربوط دارند. یکی از این مدل‌های فرایند، استاندارد کریسپ است که با استفاده از آن می‌توان یک پروژه‌ی داده‌کاوی را در قالب چند مرحله و فعالیت‌های مشخص، سامان‌دهی کرد.

۱-۱-۳- استاندارد کریسپ[۱۳]

این مدل خلاصه‌ای از چرخه‌ی حیات پروژه‌های داده‌کاوی را ارایه می‌کند که شامل تمام مرحله‌های پروژه، کارهای مربوط و خروجی‌ها است. چرخه‌ی حیات داده‌کاوی به شش مرحله تقسیم می‌شود که در شکل ۱ نشان داده شده است [۱۳].



شکل ۱- چرخه‌ی حیات داده‌کاوی

- مرحله‌ی درک پروژه: مرحله‌ی اولیه بر درک هدف‌ها و نیازمندی‌های پروژه تمرکز دارد. در این مرحله، تعریف مسئله‌ی داده‌کاوی و مهیا کردن ساز و کارهای اولیه برای نایل شدن به هدف، تعریف می‌شود.
- مرحله‌ی درک داده: این مرحله به‌منظور شناخت کیفیت داده، کشف بینشی نسبت به داده و یا کشف زیرمجموعه‌هایی برای ساختن فرضیاتی برای داده‌های مخفی است. ارتباط نزدیکی بین مرحله‌ی درک پروژه و درک داده وجود دارد. اجرای صحیح داده‌کاوی نیازمند حداقل درک اولیه‌ای از داده‌ها است.
- مرحله‌ی آماده‌سازی داده: مرحله‌ی آماده‌سازی داده‌ها تمام فعالیت‌هایی است که به‌منظور ساخت پایگاه داده نهایی، انجام می‌شود. آماده‌سازی شامل ساخت جدول و رکوردها، انتخاب ویژگی‌ها، پاک‌سازی داده، ساخت ویژگی‌های جدید و انتقال داده به ابزارهای مدل‌سازی است.
- مرحله‌ی مدل‌سازی: در این مرحله، فنون مختلف مدل‌سازی انتخاب و اجرا می‌شوند. ارتباط نزدیکی بین آماده‌سازی داده و مدل‌سازی وجود دارد. به‌طور معمول یکی مشکلات داده را کشف می‌کند و دیگری ایده‌هایی برای ساخت داده‌ی جدید می‌دهد.
- مرحله‌ی ارزیابی: در این مرحله یک یا دو مدلی که کیفیت بالایی دارند ساخته می‌شود. پیش از توسعه‌ی نهایی مدل‌ها، مهم است که مدل را به‌طور کامل ارزیابی و قدم‌هایی که ما را به مدل رسانده است، مرور کنیم. در انتهای این مرحله، تصمیم‌گیری روی استفاده از نتایج داده‌کاوی باید به‌دست آمده باشد.
- مرحله‌ی توسعه: ساختن مدل، عموماً پایان پروژه نیست. به‌طور معمول دانش به‌دست آمده باید مرتب شود و به روشی که برای مشتری قابل استفاده باشد، به کار رود.

۲-۱-۳- فرایند داده‌کاوی طرح اطلاعات نیروی کار بر اساس استاندارد کریسپ

در این مرحله از کار، مطابق با استاندارد کریسپ، فرایند کشف دانش و استخراج الگوهای طرح نیروی کار مبتنی بر فن قاعده‌های پیوند، پیاده‌سازی می‌شود. در این زمینه تلاش بر این بوده است که ترتیب مراحل استاندارد کریسپ، در حد نیاز این پروژه بر روی داده‌ها اعمال شود.

- درک پروژه: همان‌طور که اشاره شد این گام شامل تعیین اهداف پروژه و بررسی چگونگی ایجاد یک پروژه‌ی داده‌کاوی است. طرح آمارگیری نیروی کار با هدف محاسبه‌ی شاخص‌های بازار کار، از جمله نرخ مشارکت اقتصادی طراحی شده است. هدف‌های کلی از اجرای این طرح، شناخت ساختار نیروی کار، تعیین وضعیت جاری ویژگی‌های جمعیت فعال، شاغل و بیکار در کل کشور، نقاط شهری و نقاط روستایی است [۳]. از پایگاه اطلاعاتی طرح آمارگیری نیروی کار در بازه‌ی زمانی سال‌های ۱۳۸۸-۱۳۹۳ به‌منظور کاوش قاعده‌های پیوند استفاده شده است. هدف اصلی یافتن الگوهای پنهان در بانک اطلاعاتی طرح آمارگیری نیروی کار است که بیان‌کننده‌ی ویژگی‌های افراد غیر فعال به لحاظ اقتصادی است.
- درک داده: در این بخش توضیحات کلی در خصوص فیلدهای پرسش‌نامه ذکر می‌شود. تشریح مفاهیم و واژه‌های مورد استفاده در حوزه‌ی نیروی کار در پیوست ذکر شده است.
- گرداوری داده‌های اولیه: داده‌های اولیه مربوط به اطلاعات بازار کار از مجموعه داده‌های طرح نیروی کار دریافت شده است. داده‌های این طرح به شرح جدول ۱ است [۳]. لازم به ذکر است پایگاه داده‌های این طرح در برگیرنده‌ی ریزداده‌های زیادی در زمینه‌ی بازار کار است، ولی در این جدول فقط فهرست مجموعه داده‌ها تشریح شده است.

جدول ۱- مجموعه داه‌های بانک اطلاعات طرح نیروی کار

نام بخش	شرح بخش
بخش الف	وضعیت اشتغال در طول هفته‌ی گذشته
بخش ب	ویژگی‌های اشتغال در شغل اصلی
بخش پ	ساعات کار انجام‌شده برای همه‌ی مشاغل
بخش ت	تمایل برای افزایش ساعت کار و جستجوی کار دیگر
بخش ث	ویژگی‌های اشتغال در شغل دوم
بخش ج	جستجوی کار
بخش چ	تجربه‌ی کار قبلی
بخش ح	ویژگی‌های شغل مورد نظر

- آماده‌سازی داده: در این مرحله هدف تجمیع اطلاعات مورد نظر تحقیق از اطلاعات جداول فوق در قالب یک جدول تجمیع‌شده‌ی نهایی است. از آنجا که در این تحقیق هدف بررسی، وضعیت مشارکت اقتصادی افراد و استخراج الگوهای مؤثر در داده‌ها است، بنا بر این اطلاعات در جدول نهایی در سطح فرد گردآوری می‌شود. در این مرحله گام‌های زیر انجام شده است.
- انتخاب داده‌ها: در ابتدا جداول اطلاعات کل خانوار و اطلاعات اعضاء خانوار با یکدیگر تجمیع و ویژگی‌های مورد نظر از این جداول استخراج می‌شوند. در این مرحله، جدول خروجی شامل اطلاعات اعضاء خانوار بوده که دارای کلید اصلی اطلاعات خانوار به همراه فیلدهای حاوی مشخصات عضو مورد نظر است. فهرست آن در جدول ۲ نشان داده شده است.
- پاک‌سازی داده: این عملیات به‌منظور برطرف شدن و اصلاح مشکل کیفی داده‌ها، نظیر پرکردن داده‌های گم‌شده، هموارکردن نویزها، شناخت و حذف داده‌های دورافتاده و برطرف کردن ناسازگاری‌ها است. به‌منظور تحلیل اولیه و پاک‌سازی داده‌ها از نرم‌افزار کلمنتاین استفاده شده است [۶].

در اینجا عملیات ممیزی داده‌ها[۱۴] را به‌منظور بررسی شاخص‌های کیفی داده‌ها انجام می‌دهیم. این عملیات در نرم‌افزار کلمنتاین با عملگر ممیزی داده‌ها انجام می‌شود. این عملگر، اطلاعات مربوط به میانگین، کمینه و بیشینه و مقدارهای خارج از بازه‌ی نرمال هر فیلد را استخراج می‌کند. پس از ممیزی، خروجی داده‌ها گرفته می‌شود و با توجه به اطلاعات میانگین، کمینه و بیشینه‌ی مقادیر هر فیلد، یک دید کلی و اولیه نسبت به داده‌ها بوجود می‌آید. از آن‌جا که هدف تحقیق، بررسی الگوهای مربوط به ویژگی‌های جمعیت غیر فعال اقتصادی است، لذا به‌منظور افزایش دقت، برخی داده‌ها به‌دلیل خارج بودن از محدوده‌ی بررسی تحقیق، پیش از شروع کار از مجموعه داده‌ها کنار گذاشته می‌شود، زیرا این داده‌ها هیچ تأثیری در نحوه‌ی انتخاب و تصمیم‌گیری در پژوهش نداشتند. این اطلاعات شامل موارد زیر است:

- افراد واقع در سنین ۱۵ تا ۶۵ سال به‌عنوان جمعیت دارای قدرت کار، مورد بررسی قرار گرفته‌اند؛ لذا رکوردهای خارج از این محدوده‌ی سنی کنار گذاشته شده‌اند.
- اطلاعات افراد با کد بستگی با سرپرست ۱ تا ۳ مد نظر قرار گرفته‌اند (جدول ۲) و بقیه‌ی کدهای بستگی با سرپرست کنار گذاشته شده‌اند.

جدول ۲- اطلاعات اعضای خانوار

نام فیلد	شرح فیلد
استان	استان
درجه‌ی توسعه‌یافتگی استان	رتبه‌ی استان طبقه‌ی استان بر اساس متغیر توسعه‌یافتگی
منطقه‌ی محل سکونت	۱- شهری ۲- روستایی
بعد خانوار	(فرد در خانوار چند نفره عضو است)
بستگی با سرپرست خانوار	۱- سرپرست خانوار ۲- همسر ۳- فرزند ۴- عروس یا داماد ۵- نوه یا نتیجه ۶- پدر یا مادر ۷- خواهر یا برادر ۸- پدر یا مادر همسر ۹- خواهر یا برادر همسر R–سایر خویشاوندان و B–غیر خویشاوندان
جنس	۱-مرد ۲- زن
سن	سن
وضع تحصیل (آیا در حال حاضر تحصیل می‌کند ؟)	۱-در حال تحصیل است، ۲- در حال تحصیل نیست
وضع سواد (آیا توانایی خواندن و نوشتن یک متن ساده را دارد؟)	۱- باسواد، ۲- بی‌سواد
دوره یا مدرک تحصیلی	۱- ابتدایی، ۲- راهنمایی، ۳-متوسطه و دیپلم، ۴- پیش‌دانشگاهی، ۵ – فوق دیپلم، ۶- لیسانس و فوق‌لیسانس، ۷- دکترای حرفه‌ای، ۸ – دکترای تخصصی و ۹- سوادآموزی و غیر رسمی
وضع زناشویی	۱- دارای همسر، ۲- بی‌همسر بر اثر فوت، ۳- بی‌همسر بر اثر طلاق و ۴- هرگز ازدواج نکرده
وضعیت مشارکت (فعالیت) اقتصادی	۱- فعال اقتصادی ۲- غیر فعال اقتصادی

در ادامه نیاز است که مسئله‌ی مقادیر گم‌شده را بررسی و آن را برطرف کنیم.

- مقدارهای گم‌شده: در طی انجام یک فرایند آمارگیری، ممکن است اطلاعات حاصل از یک طرح آمارگیری، در برخی از فیلدها دارای مقدار نامعلوم، گرداور نشده و یا مقدارهای اشتباه باشد. به این‌گونه موارد، مقدارهای گم‌شده گفته می‌شود [۶]، [۱۲]. به‌عنوان مثال، فرض کنید فیلد مدرک تحصیلی برای افراد زیر ۱۰ سال پر شده باشد و یا فیلدهایی نظیر جنسیت که کدهای ۱ و ۲ را به‌ترتیب برای مرد و یا زن دارد، با

مقداری به جز این دو کد تکمیل شده باشد. به‌منظور افزایش دقت در ساخت مدل، می‌بایست مسئله‌ی داده‌های گم‌شده، برطرف شود. به‌طور کلی در این زمینه می‌توان دو رویکرد را دنبال کرد. [۶]:

- ✓ فیلدها و رکوردهایی که مقادیر گمشده دارند را کنار گذاشت، و
- ✓ مقدارهای گم‌شده را با مقادیر دیگری جانهی[۱۵] کرد.

از آنجایی که بیش‌تر مدل‌ها با وجود مقدارهای خالی درست عمل نمی‌کنند، همچنین به‌منظور حفظ تمامی رکوردها، از رویکرد دوم استفاده کرده و مقدارهای گم‌شده را با مقدار دیگر جانهی می‌کنیم. لذا مقدارهای گم‌شده شامل موارد خالی یا دارای @ در اطلاعات، با مقدار عددی صفر جایگزین شده‌اند. انتخاب عدد صفر به این دلیل است که این مقدار عددی در بررسی ممیزی انجام‌شده بر روی داده‌ها، به‌عنوان مقدار عددی مورد استفاده در فیلدها نبوده است.

- قالب‌بندی داده‌ها: در این گام، فیلدهایی که برای پردازش مدل دارای فرمت مناسب نیستند می‌بایست تغییر فرمت داده شوند. به‌عنوان نمونه برای انجام محاسبات و سهولت کار و با توجه به ماهیت مقادیر فیلدهای مورد استفاده‌ی طرح در این تحقیق، در قالب فیلدهای عددی و یا دسته‌بندی قرار داده شده است.
- تجمیع داده‌ها: در این مرحله اطلاعات مربوط به مشخصات فردی اعضای خانوار و اطلاعات وضعیت شغلی و فعالیت اقتصادی آن‌ها با توجه به تعریف‌های مربوط در یک جدول نهایی گرداوری می‌شوند. علاوه بر آن با توجه به پراکندگی مقدارهای داده‌ای در اطلاعات برخی فیلدها، به‌منظور ساخت مدل و استخراج الگوهای مؤثر، نیاز است که برخی فیلدها به‌صورت گروه‌بندی و دارای کد طبقه در داده‌ی نهایی لحاظ شوند.
- مدل‌سازی و انتخاب فن: هدف کاربردی این تحقیق، شناسایی قاعده‌های پیوند مشخصات افراد خانوار در ارتباط با وضعیت مشارکت اقتصادی آنان است. از این‌رو، فن انتخابی مدل‌سازی، فن قاعده‌های پیوند است که با استفاده از آن بتوان قاعده‌های پیوند در مجموعه‌ی داده‌ها را استخراج کرد.

۴- تحلیل داده‌ها

۴-۱- تحلیل و ارزیابی فرایند داده‌کاوی

۴-۱-۱- مدل‌سازی و ارزیابی داده

همان‌طور که اشاره شده است به‌منظور انجام مطلوب و دقیق فرایند حوزه‌ی داده‌کاوی، از استاندارد کریسپ استفاده می‌شود. در فرایند مذکور و در گام آماده‌سازی داده، فعالیت‌های مربوط به انتخاب، پاک‌سازی داده، قالب‌بندی و تجمیع داده‌ها انجام می‌شود. پس از بررسی اولیه فیلدهای زیر به‌منظور مدل‌سازی انتخاب شدند:

- مشخصات فردی اعضای خانوار شامل فیلدهای سکونت در منطقه‌ی شهری/روستایی، بستگی با سرپرست خانوار، جنس، سن، وضع تحصیل، وضع سواد، سطح تحصیلات، وضع زناشویی، بعد خانوار، سطح توسعه‌یافتگی استان،
- وضعیت مشارکت اقتصادی (فعال و غیر فعال) در ادامه، مدل‌سازی و ارزیابی داده انجام می‌شود. برای انجام گام‌های مدل‌سازی داده، نمونه‌ی کوچکی از داده‌ها با استفاده از فن قواعد پیوند مدل شد. قواعد پیوند حاصل از آن با توجه به مبانی نظری و نظریه‌های بازار کار، در راستای نزدیک شدن به تعریف پروژه مورد ارزیابی قرار گرفت. با توجه به ماهیت برخی فیلدها نظیر سن، استان و همچنین مؤثر بودن استان محل سکونت در قاعده‌های ایجاد شده، به‌منظور افزایش دقت مدل و قاعده‌های خروجی، گروه‌بندی زیر در داده‌های ورودی انجام شده است،
- گروه سنی افراد در محدوده‌ی سنی ۱۵ تا ۶۵ سال، به ۵ گروه سنی تقسیم شد (گروه سنی یک: ۱۵ تا ۲۵ سال، گروه سنی دو: ۲۵ تا ۳۵ سال، گروه سنی سه: ۳۵ تا ۴۵ سال، گروه سنی چهار: ۴۵ تا ۵۵ سال و گروه سنی پنج: ۵۵ تا ۶۵سال)، و
- رتبه‌بندی توسعه‌یافتگی استان‌ها بر اساس نتیجه‌های یک طرح انجام‌شده در مرکز آمار ایران، در این مورد برای هر فرد رتبه‌ی استان محل سکونت وی در این فیلد قرار می‌گیرد. به عبارتی صفت استان به افراد تعمیم داده شده است. در ادامه، قاعده‌های استخراج‌شده که تکراری هستند و یا با توجه به تعریف مسئله‌ی مورد نظر نبوده است از مجموعه قاعده‌ها حذف شده است.

۲-۱-۴- جایگزینی کد یکتا برای مقدارها

با توجه به این که داده‌های مورد نظر در زمان پردازش در سامانه‌ی فایل توزیع‌شده هدوپ به‌صورت فایل متنی قرار می‌گیرند، دارای ساختار داده‌ای مشتمل بر نام فیلد نیستند. لذا به دلیل این که اطلاعات در برخی فیلدها دارای کدهای مشابه است، این مسئله انجام فرایند استخراج الگو و یافتن قاعده‌های پیوند را دچار مشکل می‌کند. از این‌رو، مقدارهای فیلدهای مجموعه با کدهای منحصر به فرد جایگزین شده است. به‌عنوان نمونه، جایگزینی کد برای برخی فیلدها در جدول ۳ نمایش داده می‌شود. فهرست کامل این کدها در پیوست آمده است.

جدول ۳ – نمونه‌ای از جایگزینی مقدار داده‌ای جدول نهایی با کدهای یکتا

نام فیلد	مقدار قبلی فیلد	مقدار جایگزین شده
وضعیت مشارکت اقتصادی	۱= فعال	۱۰۱
	۲= غیر فعال	۱۰۲
شهری/روستایی	۱= شهری	۳۰۱
	۲= روستایی	۳۰۲

۳-۱-۴- اجرای الگوریتم و کاوش قاعده‌های پیوند

الگوریتم مطابق با مرحله‌های ذکرشده در بخش‌های قبلی مورد پیاده‌سازی قرار گرفت. در ادامه پس از آماده‌سازی بستر اجرایی، الگوریتم Apriori را در قالب این چارچوب هدوپ اجرا شده است. پارامترهای مربوط به ارزیابی قاعده‌های پیوند در این تحقیق به‌صورت کمینه ضریب پشتیبان برابر با ۱۰% و کمینه ضریب اطمینان برابر با ۷۰% در نظر گرفته شده است.

در این بخش هدف یافتن رابطه‌ی وابستگی بین مشخصات افراد خانوار و وضعیت مشارکت اقتصادی آنان است. از آن جایی که هدف اصلی تحقیق و بررسی این وابستگی، برای افراد غیر فعال از نظر اقتصادی است، بنا بر این نتیجه‌های حاصل به‌صورت

قاعده‌های پیوند تولیدشده بر روی اطلاعات طرح نیروی کار فقط در گروه افراد غیر فعال جامعه ارایه می‌شود. در جدول ۴ به برخی از مهم‌ترین الگوهای پنهان در گروه افراد با وضعیت مشارکت اقتصادی غیر فعال این طرح اشاره شده است.

در این جدول، قاعده‌های استخراجی به شرح زیر، بیانگر ویژگی‌های جامعه‌ی افراد با وضعیت مشارکت اقتصادی غیر فعال است:

قانون ۱: زنان بی‌سواد و کم‌سواد (با تحصیلات ابتدایی یا راهنمایی) بیش‌تر در معرض غیر فعال شدن قرار دارند.

قانون ۲: زنان دارای همسر (شهری و روستایی) بیش‌تر در معرض غیر فعال شدن قرار دارند.

قانون ۳: زنان دارای همسر شهری در خانوارهای پرجمعیت (۴ نفره و بیش‌تر) بیش‌تر در معرض غیر فعال شدن قرار دارند.

قانون ۴: زنان ساکن در استان‌های کم‌تر توسعه‌یافته بیش‌تر در معرض غیر فعال شدن قرار دارند.

قانون ۶: دختران هرگز ازدواج نکرده در استان‌های کم‌تر توسعه‌یافته، بیش‌تر در معرض غیر فعال شدن قرار دارند.

قانون ۵: جوانان ۱۵ تا ۲۵ سال (پسر و دختر) در حال تحصیل و ازدواج‌نکرده بیش‌تر در معرض غیر فعال شدن قرار دارند.

قانون ۷: جوانان (پسر و دختر) در حال تحصیل در جامعه‌ی روستایی، بیش‌تر در معرض غیر فعال شدن قرار دارند.

قانون ۸: زنان باسواد در جامعه‌ی روستایی، بیش‌تر در معرض غیر فعال شدن قرار دارند.

تأمل در قاعده‌های استخراج‌شده و مروری بر مبانی نظری و ادبیات بازار کار در خصوص مشارکت اقتصادی، حاکی از هم‌خوانی در دستاوردهای دو رویکرد سنتی و رویکرد مبتنی بر داده‌کاوی است. هر چند عوامل متعدد اقتصادی، اجتماعی، فرهنگی و تعیین‌کننده‌ی وضعیت مشارکت اقتصادی افراد است، اما با در نظر گرفتن اطلاعات موجود در بانک داده‌ای مورد استفاده، متغیرهای زیر (مطابق با جدول ۴) به‌عنوان پارامترهای مورد بررسی وضعیت مشارکت اقتصادی افراد در نظر گرفته شده است:

جدول ۴- بخشی از الگوهای استخراجی - وابستگی مشخصات فردی با وضعیت مشارکت اقتصادی

ضریب اطمینان	۱۰۰	۹۵	۱۰۰	۱۰۰	۱۰۰	۱۰۰	۱۰۰	۱۰۰	۱۰۰	۱۰۰	۸۱	۸۶	۸۵
ضریب پشتیبان	۱۳	۱۲	۱۶	۱۱	۱۲	۱۲	۱۷	۱۳	۱۳	۱۶	۱۴	۱۳	۱۴
محدوده سنی		۱۵۰											
وضعیت زناشویی	۹۰۴	۹۰۴					۹۰۱	۹۰۱			۹۰۱		
مدرک تحصیلی						۲۱/۱۱							
وضع سواد		۸۰۱			۸۰۲		۸۰۱	۸۰۱	۸۰۱	۸۰۱			۸۰۱
در حال تحصیل		۷۰۱										۷۰۱	
جنسیت	۶۰۲		۶۰۲	۶۰۲	۶۰۲	۶۰۲	۶۰۲	۶۰۲	۶۰۱	۶۰۲	۶۰۲		۶۰۲
بستگی باسرپرست		۵۰۳							۵۰۳	۵۰۳		۵۰۳	
طبقه‌ی استان	۲۶۰		۲۶۰	۳۶۰									
بعد خانوار							۴	۵					
شهری /روستایی		۳۰۱					۳۰۱	۳۰۱			۳۰۲	۳۰۲	۳۰۲

- جنس افراد،
- سن افراد،
- وضع سواد و تحصیلات افراد،
- محل سکونت (شهر/ روستا) افراد،
- ساختار جمعیتی خانواده (بعد خانواده)، و
- درجه‌ی توسعه‌یافتگی استان محل سکونت.

همان‌گونه که مشاهده می‌شود متغیرهای خرد (فردی)، متغیرهای میانی (خانوادگی) و متغیرهای کلان (استان) تعیین‌کننده‌ی وضعیت مشارکت اقتصادی افراد است. جنسیت یکی از عوامل تعیین‌کننده‌ی وضعیت مشارکت اقتصادی افراد است، به گونه‌ای که تقریباً در همه‌ی جامعه‌ها نرخ مشارکت اقتصادی زنان کم‌تر از مردان است. در این ارتباط، حسب اطلاعات جهانی بین میزان مشارکت اقتصادی به‌ویژه مشارکت اقتصادی زنان و سطح توسعه‌یافتگی کشورها رابطه‌ای ناخطی وجود دارد. به‌عبارتی، کشورهای توسعه‌یافته حداکثر مشارکت اقتصادی، و کشورهای در حال توسعه و توسعه‌نیافته، حداقل مشارکت اقتصادی را تجربه می‌نمایند. این بدان معنی است که نوع، ماهیت، شرایط، کارکرد و علل مشارکت زنان با توجه به سطح توسعه‌ی کشورها متفاوت است [۸].

۵- بحث و نتیجه‌گیری

بر اساس نتایج این پژوهش، عمده‌ترین دلایل پایین بودن نرخ مشارکت اقتصادی ایران در سال‌های اخیر، سهم اندک زنان در بازار کار، غیر فعال بودن جوانان به‌علت تحصیل در مقاطع تکمیلی، کم‌سوادی و بی‌سوادی بخشی از جامعه‌ی زنان، مسئولیت‌های خانوادگی زنان، پائین بودن سطح توسعه‌یافتگی استان محل زندگی و فقدان فرصت‌های مناسب برای فعالیت اقتصادی زنان است.

در تحقیق حاضر، نتیجه‌های به‌دست آمده با یافته‌های پژوهش‌های سنتی از نظر موضوعی همخوان است. بر اساس منابع و داده‌های بین‌المللی کشورهایی‌که از نرخ مشارکت اقتصادی پایین برخوردارند، بدون تردید مشارکت زنان در بازار کار آنان محدود است و سهم زیادی از زنان در سن کار به‌دلایل فردی یا ساختاری انگیزه‌ی ورود به بازار کار را ندارند. نتایج تحلیل اکتشافی حاضر نیز نشان می‌دهد که متغیر جنس در ترکیب با پاره‌ای از صفات

دیگر که در بخش پیشین به آن‌ها اشاره شد، مهم‌ترین دلایل استخراج‌شده در توضیح غیر فعال بودن افراد جامعه است. در این راستا بیش‌تر زنان به‌دلایل مسئولیت‌های شخصی و خانوادگی حاضر به ورود، به بازار کار نبوده‌اند. به‌عبارتی، سهم زیادی از زنان در سن کار صرفاً به ایفای نقش‌های سنتی می‌پردازند. شایان ذکر است که تعداد زیادی از این افراد دارای مدارک دانشگاهی نیز هستند. با این حال، از آن جایی که پیش بردن چرخ توسعه در گرو مشارکت همه‌ی اقشار جامعه در حوزه‌های مختلف است، بهره‌گیری از ظرفیت حضور بیش‌تر زنان در عرصه‌ی کسب و کار یکی از پیش‌شرط‌های توسعه است. به نظر می‌رسد در کنار جایگزینی نسلی و بازنگری در جایگاه سنتی زنان و توسعه‌ی فرصت‌های شغلی متناسب با سطح مهارتی آنان، می‌توان به افزایش نرخ مشارکت اقتصادی زنان امیدوار بود. انتظاری که به نظر نمی‌رسد در کوتاه‌مدت تحقق یابد.

همچنین، سن افراد در سطح خرد و ساختار سنی جمعیت در سطح کلان، بر مشارکت اقتصادی شهروندان مؤثر است. در این ارتباط جمعیت‌شناسان برای بیان اهمیت و نقش سن در مشارکت اقتصادی افراد جامعه، از مفهوم پنجره‌ی جمعیتی[۱۶] استفاده می‌کنند. پنجره‌ی جمعیتی یکی از مهم‌ترین مراحل در ساخت سنی یک جمعیت است و در این مرحله ساخت سنی جمعیت، عرضه و تقاضای نیروی کار را تحت تأثیر قرار می‌دهد. در دوره‌ی پنجره‌ی جمعیتی عرضه‌ی نیروی کار به حداکثر می‌رسد و بیش‌تر افراد جامعه در سن کار قرار می‌گیرند [۸]. در حال حاضر، بیش از یک دهه است که کشور ایران در وضعیت پنجره‌ی جمعیتی قرار گرفته، ساختارهای آموزشی توسعه‌یافته و بسیاری از نظامات اجتماعی نیز متحول شده است. در چنین شرایطی انتظار می‌رود که کفه‌ی افراد فعال اقتصادی بر افراد غیر فعال اقتصادی سنگینی نماید، اما واقعیت به این گونه نیست.

با نگاه جامعه‌شناختی و به تأسی از بوردیو[۱۷] می‌توان گفت که به‌صورت تاریخی در ذهن عمومی جامعه در خصوص توانایی، جایگاه و کارکرد پیامدهای ورود زنان به عرصه‌ی فعالیت، باید و نبایدهایی نهادینه شده است. وی معتقد است که باورهای ناصحیح جنسیتی که برآمده از جایگاه سنتی و تاریخی زنان است، به نحوی ملکه‌ی ذهن کنش‌گران انسانی شده است که نه تنها تحمیل‌کنندگان آن، بلکه زنان جامعه که هدف این کلیشه‌ها هستند نیز آن‌ها را پذیرفته‌اند. بوردیو بر نقش و اهمیت جامعه‌پذیری ایجاد و توسعه‌ی این کلیشه‌ها تأکید دارد [۱]. به‌عبارتی، هر چند گشایش‌های نسبی برای حضور زنان در بازار کار به وجود آمد، ولیکن این کافی نیست و افزایش نرخ مشارکت اقتصادی کشور بدون شک در

گرو افزایش مشارکت اقتصادی زنان و آن مستلزم عزم جدی در پذیرش اهمیت نقش زنان در توسعه و تعالی جامعه (توسط خود آنان و جامعه) و در نتیجه بسترسازی برای تحقق این مهم است. نکته‌ای که نباید از نظر دور داشت آن است که فرصت جمعیتی به‌صورت خودکار عمل نمی‌کند بلکه یک پتانسیل و نیروی بالقوه است که در صورت هدایت و برنامه‌ریزی می‌تواند به فعل درآید که این امر خود نیازمند محیط و بستر اجتماعی، اقتصادی و نهادی مناسب است.

هرچند بیش از یک دهه است که کشور ایران در شرایط پنجره‌ی جمعیتی قرار گرفته و از مرحله‌ی جوانی جمعیت عبور کرده است، اما رشد چشم‌گیر و توسعه‌ی تحصیلات تکمیلی نقش به‌سزایی در به تأخیر انداختن ورود جوانان به بازار کار داشته است. اما، می‌توان آن را نوعی سرمایه‌گذاری تلقی کرد. سرمایه‌گذاری که در بسیاری از متون توسعه از آن به سرمایه‌ی انسانی تعبییر شده است. از سرمایه‌ی انسانی در کنار سرمایه‌ی اجتماعی، سرمایه‌ی اقتصادی و سرمایه‌ی فرهنگی به‌عنوان ارکان ضروری توسعه نام برده می‌شود. در دنیای امروز و با توجه به تخصص‌گرایی در حرف، سرمایه‌ی انسانی یا همان دانش و مهارت، کلید ورود و موفقیت در کسب و کار است. هرچند در یافته‌های پژوهش، کسب تحصیلات یکی از علل غیر فعال شدن به‌طور موقت است، اما بعد از آن، افراد پتانسیل ورود به بازار کار را پیدا می‌کنند. در این ارتباط، از نظر کلمن[۱۸]، سرمایه‌ی اجتماعی و انسانی به‌عنوان یک پدیده‌ی اجتماعی باعث بروز خلاقیت، ایده‌پروری، تسهیل رفتارهای نوآورانه و ریسک‌پذیری می‌شود [۲]. به تعبیری، سرمایه‌ی اجتماعی ضمن توانمندسازی و قدرتمندسازی افراد، کنش اقتصادی آنان را تسریع و در کمیت و کیفیت مشارکت اقتصادی مؤثر است.

توضیحات

1. Mine
2. data
3. predictive
4. descriptive
5. association rules
6. antecedent
7. consequent

8. minimum support
9. minimum confidence
10. Support
11. Confidence
12. Minimum confidence
13. Cross Industry Standard Process for Data Mining (CRISP-DM))
14. Data audit
15. Imputation
16. demographic window
17. Pierre Bourdieu (۱۹۳۰-۲۰۰۲) جامعه‌شناس و مردم‌شناس سرشناس فرانسوی
18. James Samuel Coleman (۱۹۲۶-۱۹۹۵) جامعه‌شناس

مرجع‌ها

[۱] احمدی، حبیب؛ گروسی، سعیده (۱۳۸۳). بررسی تأثیر برخی عوامل اجتماعی و فرهنگی بر نابرابری جنسیتی در خانواده‌های شهر کرمان و روستاهای پیرامون آن، فصلنامه‌ی مطالعات زنان، سال ۲، شماره‌ی ۶.

[۲] فاین، بن (۱۳۸۵). سرمایه‌ی اجتماعی و نظریه‌ی اجتماعی، ترجمه محمد کمال سروریان، تهران، پژوهشکده‌ی مطالعات راهبردی.

[۳] مرکز آمار ایران (۱۳۹۳). مجموعه مدارک طرح آمارگیری نیروی کار، تهران.

[4] Agrawal, R. and Srikant, R. (1994). Fast Algorithms for Mining Association Rules. In Proceedings of the 20th International Conference on Very Large Databases (VLDB'94), Santiago, Chile.

[5] Agrawal, R., Imielinski, T. and Swami, A. (1993).Mining Association Rules between Sets of Items in Large Databases. Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data - SIGMOD '93. p. 207.

[6] Book from SPSS, Clementine® 12.0 User's Guide, Printed in the United States of America, ISBN-13: 978-1-56827-395-2.

[7] Fayyad, M., Piatetsky-Shapiro, G., Smyth, P. and Uthurusamy, R. (1996). Advances in Knowledge Discovery and Data Mining. Menlo Park, AAAI Press, California

[8] Guy, S. (1981). Labour Force Participation and Development, International Lablabour Office, Geneva.

[9] Hand, D.J., Mannila, H. and Smyth, P. (2001). Principles of Data Mining, MIT press.

[10] Han, J. and Kamber, M. (2006). Data Mining: Concepts and Techniques, Second Edition, University of Illinois at Urbana-Champaign, by Elsevier Inc.

[11] Oruganti, S., Ding, Q. and Tabrizi, N. (2013) Exploring HADOOP as a Platform for Distributed Association Rule Mining. FUTURE COMPUTING 2013: The Fifth International Conference on Future Computational Technologies and Applications.

[12] SPSS Learning Module Missing data. UCLA: Statistical Consulting Group. From https://stats.idre.ucla.edu/spss/modules/missing-data

[13] Wirth, R. and Hipp, J. üdiger "CRISP-DM: Towards a Standard Process Model for Data Mining", DaimlerChrysler Research & Technology FT3/KL PO BOX 2360 89013 Ulm, Germany.

پیوست

مقدار داده‌های جایگزین‌شده با کدهای عددی یکتا

نام فیلد	مقدار قبلی فیلد	مقدار جایگزین شده
وضع فعالیت	فعال	۱۰۱
	غیر فعال	۱۰۲
وضعیت شغل	شاغل	۲۰۱
	بیکار	۲۰۲
	غیر فعال	۲۰۳
شهری/روستایی	شهری	۳۰۱
	روستایی	۳۰۲
تعداد افراد خانوار	بعد خانوار (فرد در خانوار چند نفره عضو است)	کد ۴۰۰ + بعد خانوار
بستگی با سرپرست	۱- سرپرست خانوار	۵۰۱
	۲- همسر	۵۰۲
	۳- فرزند	۵۰۳
جنسیت	۱-مرد یا پسر	۶۰۱
	۲- زن یا دختر	۶۰۲
وضع تحصیل	در حال حاضر تحصیل می‌کند	۷۰۱
	در حال حاضر تحصیل نمی‌کند	۷۰۲
وضع سواد	توانایی خواندن و نوشتن یک متن ساده را دارد	۸۰۱
	توانایی خواندن و نوشتن یک متن ساده را ندارد	۸۰۲
وضع زناشویی	دارای همسر	۹۰۱
	بی‌همسر در اثر فوت همسر	۹۰۲
	بی‌همسر بر اثر طلاق	۹۰۳
	هرگز اردواج نکرده	۹۰۴

ادامه‌ی جدول مقدار داده‌های جایگزین‌شده با کدهای عددی یکتا

نام فیلد	مقدار قبلی فیلد	مقدار جایگزین شده
گروه سنی	۱۵ تا ۲۵ سال	۱۵۰
	۲۵ تا ۳۵ سال	۲۵۰
	۳۵ تا ۴۵ سال	۳۵۰
	۴۵ تا ۵۵سال	۴۵۰
	۵۵ تا ۶۵ سال	۵۵۰
رتبه استان	طبقه خیلی پایین	۱۶۰
	طبقه پایین	۲۶۰
	طبقه متوسط	۳۶۰

سمیه آهنگر
فوق لیسانس مهندسی کامپیوتر، نرم‌افزار
تهران، خیابان دکتر فاطمی، خیابان رهی‌معیری، مرکز آمار ایران.
رایانشانی: ahangar.somaye@gmail.com

مجید خلیلیان
دکترای کامپیوتر
استان البرز، دانشگاه آزاد اسلامی واحد کرج
رایانشانی: khalilian@gmail.com

نریمان یوسفی
دانشجوی دکترای جامعه‌شناسی
تهران، خیابان دکتر فاطمی، خیابان رهی‌معیری، مرکز آمار ایران.
رایانشانی: nyosofi@yahoo.com