کاربرد داده‌کاوی در استفادهٔ بهینه از بانک‌های اطلاعاتی پلیس؛ راهبردهای نوین

کامران فیضی [۱]
رسول لطفی [۲]
ناصر پیکری [۳]

تاریخ دریافت: ۹۶/۱۱/۲۸ تاریخ پذیرش نهایی: ۹۷/۲/۱۵
فصلنامه‌ی مطالعات راهبردی ناجا/ سال سوم / شماره هفتم - بهار ۱۳۹۷* ۱۴۱-۱۵۸

چکیده

در هر سازمانی، رشد فزایندهٔ پایگاه داده‌ها به‌کارگیری ابزارهای جدید و قدرتمند تحلیلی در امر پردازش و بهره‌برداری از داده‌ها را ضروری کرده است. علم داده‌کاوی یکی از ابزارهای مورد نیاز در تحلیل داده‌های حجیم می‌باشد که این حوزه را به‌عنوان یکی از مهم‌ترین حوزه‌های پژوهشی مطرح نموده است. یکی از عرصه‌های مهم کاربرد داده‌کاوی حوزهٔ جرم‌شناسی است.

نیروی انتظامی به‌عنوان یکی از سازمان‌های فعال در عرصهٔ نظم و امنیت، باتوجه‌به ماهیت مأموریتی، با حجم، تنوع و گسترهٔ فراوانی از فرایند ذخیره‌سازی و به‌کارگیری بانک‌های داده‌ای انبوه در حوزه‌های گوناگون اجتماعی، امنیتی، اقتصادی و فرهنگی در ارتباط است. ازاین‌رو، ضروری است برای اجرا و تحقق مأموریت‌های محوله، استخراج الگوهای مفید و کاربردی از پایگاه داده‌ای، با کمک علم داده‌کاوی، از اولویت‌های اصلی این سازمان می‌باشد. در این مقاله پس از معرفی روش‌های داده‌کاوی و الگوریتم‌های آن به بیان مدلی کاربردی در آماده‌سازی و پیش‌پردازش داده‌های جمع‌آوری شدهٔ مواد مخدر، شامل ویژگی‌های شخصیتی و دیگر اطلاعات جمع‌آوری‌شده از این مجرمان، به خوشه‌بندی در پنج دسته پرداخته شده است. هدف نهایی در این نگاشته آشنایی مدیران عالی برای درک اهمیت راهبردی علم داده‌کاوی در اقدامات آتی پلیس است. گفتنی است، این مقاله بخشی از کار عملی بوده که نگارندگان آن در حوزهٔ پردازش اطلاعات انجام داده‌اند و مثال‌های آن واقعی و اطلاعات سیستمی آن موجود است و به‌همین‌دلیل، نگاه کتابخانه‌ای و ارجاع به منابع در آن کمتر دیده می‌شود.

واژگان کلیدی

داده‌کاوی، پلیس، مواد مخدر، مجرمان، قاچاق

۱.استاد فناوری اطلاعات دانشگاه علامه طباطبائی (ره) .
۲.دکتری فناوری اطلاعات دانشگاه علامه طباطبائی (ره)(نویسنده مسئول) r.lotfi@yahoo.com
۳.کارشناس ارشد جامعه‌شناسی.

بیان مسئله

در دهۀ اخیر داده‌کاوی مفهومی نوظهور و به‌عنوان زیرمجموعه‌ای از فرایند استخراج دانش است که به صورت روزافزون بر کاربردهای آن در حوزه‌های گوناگون پژوهشی، صنعتی و تجاری، پزشکی و سلامت، تبلیغات و ... افزوده می‌شود. درواقع، داده‌کاوی استخراج الگوها و روابط نهفته در میان حجم عظیم داده‌های موجود در پایگاه‌های داده است که امکان دستیابی به آن‌ها با روش‌های متداول تحلیل داده‌ها وجود ندارد. داده‌کاوی به صورت هم‌زمان مفاهیمی چون هوش مصنوعی، یادگیری ماشینی، الگوریتم‌های پیشرفته و روش‌های آماری را در کنار تکنیک‌های بهینه‌سازی به‌کار می‌گیرد تا بتواند رویکردی جدید برای استخراج دانش و الگوهای نهفته در میان داده‌های حجیم ارائه دهد. داده‌کاوی فرایند مرتب‌سازی و طبقه‌بندی داده‌های حجیم و آشکارسازی اطلاعات مرتبط با هم است و امروزه، به‌عنوان یکی از ابزارهای بسیار مهم مدیران برای شناخت وضعیت دقیق‌تر سازمان و همچنین، کمک در اتخاذ تصمیم‌های مناسب کاربرد دارد. با استفاده از این تکنیک، داده‌های موجود در سازمان با به‌کارگیری ابزارهای نرم‌افزاری بررسی و تحلیل می‌شوند تا الگوهای پنهان و پیچیده‌ای که در آن‌ها وجود دارد، کشف و استخراج گردد. برای کاربردهای داده‌کاوی تاکنون زمینه‌های مختلف و متعددی شناخته شده و در حال گسترش است، ولی یکی از کاربردی‌ترین زمینه‌های داده‌کاوی به مبحث انتظامی و حوزۀ جرایم مربوط می‌شود. حجم زیاد داده‌هایی که در ارتباط با جرایم، مجرمان و شیوه‌های ارتکاب و متغیر زمان و مکان وجود دارد، پتانسیل مناسبی برای انجام دادن فرایند داده‌کاوی و استخراج دانش پنهان فراهم کرده است (Roos,2009) آنچه در نگاشتۀ پیش‌رو به آن پرداخته شده، مروری بر مفاهیم، رویکردها، تکنیک‌ها و کاربردهای داده‌کاوی در مدیریت انتظامی، با تأکید بر بانک‌های اطلاعاتی در مدیریت پیشگیری از جرم، است.

داده‌کاوی

داده‌کاوی این امکان را ایجاد می‌کند تا اطلاعات باارزش را در حجمی بسیار بزرگ از داده‌ها جست‌وجو و استخراج کنیم(Weiss & Indurkhya: 1998). همچنین[1]، شاخه‌ای از هوش مصنوعی [2]است که از سال ۱۹۶۰ در نوآوری‌های رایانه‌ای پدید آمد و به‌عنوان مقدمه‌ای بر

1 . DTM Data Mining Techniques
2 . AI Artificial intellignece

فناوری‌های جدید معرفی شد(Ha, Bae, & Park, 2000). داده‌کاوی فرایند کشف الگوهای روشنگرانه، جالب و همچنین، به‌عنوان مدل توصیفی فهم‌پذیر، و پیش‌بینی از داده‌ها در مقیاس بزرگ است (J. ZAKI,2014). در برخی از این تعاریف داده‌کاوی در حدّ ابزاری که کاربران را به ارتباط مستقیم با حجم عظیم داده‌ها قادر می‌سازد، معرفی شده است و در برخی دیگر، تعاریف دقیق‌تری که در آن‌ها به جست‌وجو در داده‌ها توجه می‌شود، موجود است. برخی از این تعاریف عبارت‌اند از: داده‌کاوی عبارت است از فرایند استخراج اطلاعات معتبر، از پیش‌ناشناخته، فهم‌پذیر و قابل‌اعتماد از پایگاه‌های دادهٔ بزرگ و استفاده از آن در تصمیم‌گیری در فعالیت‌های تجاری مهم (مهریزی،حائری،۱۳۸۲: ۴۷). در تعریف دیگری، داده‌کاوی یعنی استخراج دانش کلان، استنادپذیر و جدید از پایگاه داده‌های بزرگ (شاه‌سمندی،۱۳۸۴،۶۱). در تعریف سوم، داده‌کاوی به فرایند نیم‌خودکار تجزیه‌وتحلیل پایگاه داده‌های بزرگ با هدف یافتن الگوهای مفید اطلاق می‌شود (گودرزی،۱۳۸۸،۲۹). تعریف چهارم از داده‌کاوی یعنی تجزیه‌وتحلیل مجموعه‌داده‌های مشاهده شدنی برای یافتن روابط مطمئن بین داده‌ها (کانتاردزیک،۱۳۸۹ :۱۱۱). داده‌کاوی استخراج مفاهیم باارزش از میان انبوه داده‌هاست (Cho,2007). همان‌گونه‌که در تعاریف گوناگون داده‌کاوی مشاهده می‌شود، تقریباً در تمام تعاریف به مفاهیمی چون استخراج دانش، تحلیل و یافتن الگو بین داده‌ها اشاره شده است.

عملیات داده‌کاوی

تجزیه‌وتحلیل سطح بالا از فناوری‌های داده‌کاوی باید متمرکز روی حفظ اطلاعات باشد (S.-H. Liao et 2012)، زیرا این داده‌ها هستند که در آیندهٔ تحلیلی کمک می‌کنند تا الگوهای ایده آل استخراج شوند.

به‌طورکلی، عملیات یا وظایف مختلف داده‌کاوی به دو دسته تقسیم می‌شوند:

الف) الگوریتم‌های یادگیری با نظارت: در الگوریتم‌های یادگیری با نظارت هدف داده‌کاوی مشخص است و می‌دانیم که به دنبال چه نوع دانشی می‌گردیم. درواقع، به دنبال پیش‌بینی پارامترهای مشخص و ازپیش‌تعیین‌شده هستیم؛

ب) الگوریتم‌های یادگیری بدون نظارت: در روش‌های یادگیری بدون نظارت، هدف کاملاً تعریف‌شده نیست و بیشتر به دنبال کسب درک توصیفی از روابط شباهت‌های داده‌ها هستیم، مانند خوشه‌بندی، خلاصه‌سازی، قواعد وابستگی و الگوهای مکرر.

تکنیک‌ها، روش‌ها و الگوریتم‌های داده‌کاوی راه‌های اجرای عملیات‌های داده‌کاوی‌اند. اگرچه هر تکنیکی نقاط ضعف و قوت دارد، ابزارهای گوناگون داده‌کاوی تکنیک‌ها را براساس معیارهای ویژه‌ای انتخاب می‌کنند. این معیارها عبارت‌اند از:

۱- تناسب با نوع داده‌های ورودی

۲- شفافیت خروجی داده‌کاوی

۳- مقاومت در مقابل اشتباه در مقادیر داده‌ها

۴- میزان صحت خروجی

۵- توانایی کار کردن با حجم بالای داده‌ها (صنیعی‌آباده و دیگران، ۱۳۹۱: ۸۷)

شاید مهم‌ترین نکته این باشد که هیچ مدل یا الگوریتمی نمی‌تواند و نباید به‌تنهایی استفاده شود. برای هر مسئلهٔ داده شده طبیعت داده استفاده شده روی انتخاب مدل‌ها و الگوریتم‌هایی که برگزیده می‌شوند تأثیر خواهد گذاشت. نمی‌توان هیچ مدل یا الگوریتمی را دراین‌زمینه بهترین نامید. درنتیجه، به یک‌سری ابزار و تکنولوژی برای یافتن بهترین مدل ممکن نیاز خواهد بود. برای فهم بهتر اینکه داده‌کاوی چه کاری انجام می‌دهد، عملیات و تکنیک‌های گوناگون آن در جدول زیر بیان شده است:

عملیات و تکنیک‌های داده‌کاوی

عملیات	دسته‌بندی، پیش‌بینی، خوشه‌بندی، تعیین‌روابط بین متغیرها، قواعد هم‌بستگی
روش‌ها و تکنیک‌ها	شبکه‌های عصبی، درخت‌های تصمیم‌گیری، الگوریتم‌های نزدیک‌ترین همسایگی، تحلیل خوشه‌بندی، استنتاج قانون، الگوریتم ژنتیک، شبکه‌های بیزین و ...

۱- دسته‌بندی

دسته‌بندی یکی از عملیاتی است که در داده‌کاوی بسیار رایج بوده و استفاده می‌شود. دسته‌بندی سازمان‌ها را قادر می‌سازد که در حلّ مسائل خاص در مجموعه‌های بزرگ و پیچیده به کشف الگوهایی دست یابند. دسته‌بندی فرایندی است که مجموعه‌داده‌ها را براساس یک ویژگی خاص به قسمت‌های مشخص تقسیم می‌کند. مسائل دسته‌بندی به شناسایی خصوصیاتی منجر می‌شوند که مشخص می‌نمایند هر مورد به کدام دسته تعلق دارد. این الگو هم می‌تواند برای فهم دادهٔ موجود و هم برای پیش‌بینی اینکه هر نمونهٔ جدید چگونه کار می‌کند، استفاده شود. نتایجی که در این حالت دریافت می‌شود، بسیار ساده‌تر است و راحت‌تر تفسیر می‌شود.

مسئله‌های دسته‌بندی دنیای واقعی معمولاً ابعاد بسیار بالا و درنتیجه، پیچیدگی زیاد دارند. چگونگی تحلیل ابزارهای داده‌کاوی در این‌گونه داده‌ها و همچنین، اطلاعاتی که این ابزارها فراهم می‌نمایند، به روش استفاده شده، بستگی دارد. معمول‌ترین روش‌های استفاده شده در دسته‌بندی عبارت‌اند از: درخت تصمیم‌گیری، دسته‌بندی بیزین، شبکه‌های عصبی، SVM، نزدیک‌ترین همسایگی KNN، و دسته‌بندی جمعی.

۲- پیش‌بینی

در دسته‌بندی گروه‌هایی مشخص می‌شوند که اقلام به آن‌ها تعلق دارند.پیش‌بینی‌هایی که براساس مدل‌های دسته‌بندی مطرح می‌شوند، یک خروجی گسسته دارد که مشخص می‌کند که مثلاً یک مشتری جزء گروه با پاسخ مثبت است یا منفی و یک دانشجو جزءِ گروه با ریسک بالاست یا پایین؛ ولی پیش‌بینی، برخلاف دسته‌بندی، یک مقدار پیوسته را پیش‌بینی می‌کند، مثلاً معدل نیمسال آینده یا میزان فروش در ماه آتی و ... ابزارهای داده‌کاوی، نظیر شبکه‌های عصبی، نیز به وفور برای پیش‌بینی استفاده می‌شوند. پیش‌بینی معمولاً به‌وسیلهٔ تکنیک‌های گوناگونی از قبیل شبکهٔ عصبی، SVM، رگرسیون خطی و غیرخطی و سری‌های زمانی صورت می‌گیرد. از مسائل سادهٔ پیش‌بینی می‌توان پیش‌بینی مقادیر پیوسته براساس یک‌سری داده‌های موجود را ذکر نمود که برای مثال می‌توان به پیش‌بینی درآمد یک فرد براساس مشخصات آن اشاره کرد. ابزارهایی نظیر درخت تصمیم‌گیری و شبکه‌های عصبی چنین کاری را انجام می‌دهند.

۳- خوشه‌بندی

درواقع، خوشه‌بندی عملیاتی غیرنظارتی است. این عملیات هنگامی به‌کار برده می‌شود که ما به دنبال یافتن گروه‌هایی از داده‌های مشابه هستیم، بدون اینکه از قبل یک دربارهٔ شباهت‌های موجود پیش‌بینی داشته باشیم. فرایند گروه‌بندی مجموعه‌ای از اشیای فیزیکی یا انتزاعی به گروه هایی از اشیای شبیه به هم خوشه‌بندی نامیده می‌شود. یک خوشه شامل مجموعه‌ای از اشیاست که به هم شبیه‌اند، ولی با اشیای خارج از آن خوشه متفاوت می‌باشند. به‌وسیلهٔ خوشه‌بندی می‌توانیم نحوهٔ توزیع الگوها را مشخص نماییم. برای اینکه بتوانیم داده‌ها را خوشه‌بندی کنیم باید بتوانیم میزان شباهت آن‌ها را به‌دست آوریم. رایج‌ترین روش

اندازه‌گیری میزان شباهت داده‌ها، اندازه‌گیری فاصله‌ای است. این کار به‌طورمعمول به‌صورت غیرمستقیم انجام می‌شود، یعنی با استفاده از معیارهای اندازه‌گیری فاصله، میزان تفاوت بین دو شیء به‌دست می‌آید؛ به‌طوری‌که اشیای شبیه به هم فاصلهٔ کمتری خواهند داشت. چندین معیار اندازه‌گیری فاصله برای خوشه‌بندی وجود دارد که معروف‌ترین آن فاصلهٔ اقلیدسی است.

۴- تعیین روابط و وابستگی‌ها

تحلیل وابستگی‌های یک عملیات غیرنظارتی داده‌کاوی است که به جست‌وجو برای یافتن ارتباط در مجموعه‌داده‌ها می‌پردازد. هدف اصلی داده‌کاوی در پیدا کردن وابستگی، یافتن قانون‌های محکم و توجه‌برانگیز است. جست‌وجوی قواعد انجمنی یکی از مشهورترین و شناخته‌شده‌ترین روش‌ها در داده‌کاوی است. قوانین انجمنی یا جمعی نخستین‌بار برای کشف الگوهای خرید مشتریان با استفاده از آنالیز داده‌های فروش فروشگاه استفاده شد. چنین قوانینی کاربردهای زیادی دارند که از جملهٔ این موارد کمک در تصمیم‌گیری است (همان،۱۰۱).

فرایند داده‌کاوی

ازآنجاکه داده‌کاوی یک فرایند است، اجرای آن نیاز به الگوی فرایندی مناسب و جامع دارد. مشهورترین این فرایند براساس الگوریتم CRISP-DM است؛ هرچند بدون در نظر گرفتن این الگوریتم نیز بیشتر مراحل اصلی به‌صورت ناخودآگاه در ذهن فرد یا تیم داده‌کاو تداعی می‌شود، ولی بیان مراحل کار در یک قالب مدون و استاندارد به‌طورمسلم به سهولت ارتباط با مخاطبان، کارشناسان و مدیران منجر می‌شود. گفتنی است، متدلوژی این الگوریتم مستقل از نوع مسئلهٔ داده‌کاوی و جنس داده‌هاست. الگوریتم CRISP-DM در شش فاز متوالی طراحی شده است.

۱- درک فضای کسب‌وکار

سیستم‌های دانش‌پایه مبتنی‌بر داده‌کاوی‌اند و داده‌کاوی می‌تواند تصمیم‌سازی را،که توجیه‌پذیر است، برای سازمان‌ها انجام دهد (Wiig, 1994). درک فضای کسب‌وکار و تصمیم‌گیری مهم‌ترین فاز پروژهٔ داده‌کاوی است. برای انجام دادن این فاز باید مجموعه‌فعالیت‌های زیر مدّنظر قرار گیرد:

- آشنایی با فرایندهای تجاری مرتبط
- تعیین اهداف و نیازمندی‌های تجاری
- ارائهٔ اهداف و محدودیت‌ها در قالب مسئلهٔ داده‌کاوی
- تعیین اهداف پروژهٔ داده‌کاوی
- تهیهٔ برنامه و راهبرد اولیه برای انجام دادن پروژه.

۲. فهم داده‌ها

بانک‌های اطلاعاتی و مجموعه‌داده‌ها موادّ خام یک پروژهٔ داده‌کاوی‌اند. این فاز به شناسایی مراحل مورد نیاز برای جمع‌آوری داده‌ها و تحلیل مشخصات آن‌ها می‌پردازد. این مرحله شامل فعالیت‌های زیر است:

- جمع‌آوری داده‌های اولیه
- تشریح و توصیف داده‌ها
- کاوش ابتدایی در داده‌ها
- اعتبارسنجی کیفیت داده‌ها
- انتخاب زیرمجموعه‌های مطلوب و جالب توجه

۳. آماده‌سازی داده‌ها

پس از جمع‌آوری داده‌ها برای اجرای عملیات داده‌کاوی باید به آماده‌سازی و پیش‌پردازش مجموعه‌داده‌ها پرداخت. این مرحله شامل فعالیت‌های زیر است:

- انتخاب داده‌ها
- پاک‌سازی داده‌ها
- یکپارچه‌سازی داده‌ها
- ساختاردهی به داده‌ها
- نرمال‌سازی داده‌ها
- تنظیم فرمت داده‌ها

۴. مدل‌سازی

این فاز مرحله و هدف اصلی یک پروژهٔ داده‌کاوی است. در این مرحله از تکنیک‌ها و روش‌های تحلیلی پیچیده‌ای برای استخراج دانش و اطلاعات از مجموعه داده‌ها به‌کار گرفته می‌شوند. این مرحله شامل فعالیت‌های زیر است:

- طراحی آزمایش
- انتخاب تکنیک‌های ساخت مدل
- تنظیمات تکنیک‌های منتخب برای دستیابی به نتایج بهینه
- اجرای مدل‌ها و تکنیک‌های مختلف برای حل مسئله
- ارزیابی عملکرد مدل

۵. ارزیابی

پس از انتخاب مدل در این مرحله ضروری است به بررسی و ارزیابی این موضوع پرداخت که آیا نتایج داده‌کاوی ما را در رسیدن به اهداف پروژه یاری می‌کند. این مرحله شامل فعالیت‌های زیر است:

- ارزیابی کیفیت و کارایی نتایج
- مرور فرایند داده‌کاوی
- تعیین گام‌های بعدی

۶. اجرا و گسترش

این مرحله آخرین‌فاز از پروژهٔ داده‌کاوی است. تمرکز این فاز روی یکپارچه‌سازی دانش کسب شده و در فرایندهای داده‌کاوی برای حلّ مسائل اساسی حوزه مورد مطالعه و بررسی است. این مرحله شامل فعالیت‌های زیر می‌باشد:

- اجرای برنامه
- نظارت و نگهداری
- آماده‌سازی گزارش نهایی
- مرور و شناسایی نقاط بهبودپذیر پروژه (صنیعی‌آباده و دیگران، ۱۳۹۲: ۱۱۲)

کاربرد داده‌کاوی در پلیس

باتوجه‌به گسترهٔ مأموریتی پلیس و تنوع، حجم و سرعت انباشت داده‌ها در بانک‌های اطلاعاتی پلیس و لزوم استفاده از دانش استخراجی از داده‌های انباشتی در پیش‌بینی، خوشه‌بندی و دسته‌بندی جرایم، به‌کارگیری تکنیک‌های داده‌کاوی و اجرای آن اهمیت ویژه‌ای دارد. مجموعه‌داده‌های مورد استفاده در این مقاله حاوی ۹۹۹ رکورد مربوط به مجرمان دستگیر شده به دلیل حمل، مصرف یا فروش مواد مخدر است. این مجموعه‌داده حاوی متغیرهای جمعیت‌شناختی مجرمان مانند سن، وضعیت تأهل، شغل و غیره و همچنین، برخی از متغیرهای مربوط به نحوهٔ دستگیری آن‌ها می‌باشد. در جدول زیر متغیرهای موجود در مجموعه‌داده به همراه توضیحاتی دربارهٔ هر متغیر موجود است:

جدول ۱ متغیرهای وارد شده در تحقیق

نام انگلیسی	نام فارسی	آماره‌ها	دامنه	مقدار از دست رفته
Jensiat	جنسیت	مد = مرد = least , زن	مرد (۹۷۹)، زن (۲۰)	۰
VaziateTaahol	وضعیت تأهل	مد = متأهل = least ,مجرد	متأهل (۶۹۲)، مجرد (۳۰۲)	۵
Shoghl	شغل	مد = مشاغل آزاد = least , پزشک	مشاغل آزاد (۴۳۲)، مشاغل دیگر (۳۸۰)، دانشجو (۴)، پزشک (۱)، آزاد (۲)، کارگر ساده (۱۷) **و غیره**	۱۱۵
Vaziate Maskan	وضعیت مسکن	مد= استیجاری = least , بی‌خانمان	مالک (۱۵۸)، بی خانمان (۴۶)، پدری (مالک) (۱۵۶)، استیجاری (۵۳۴)، پدری (استیجاری) (۵۶)	۴۹
Ostane Mahale Sokunat	استان محل سکونت	مد = تهران بزرگ=least , رشت	تهران بزرگ (۵۵۶)، استان تهران (۱۵۱)، مازندران (۱۱)، زنجان (۱۲)، قم (۳)، گلستان (۱۰)، اردبیل (۲۱) **و غیره**	۱۳
Sen	سن	میانگین10.63 -/+ 33.82 =	[17.00 ; 75.00]	۰
Mizane Tahsilat	میزان تحصیلات	مد = دیپلم = least , فوق لیسانس	لیسانس (۲۸)، بی‌سواد (۸۳)، ابتدایی (۱۴۸)، دیپلم (۲۳۹)، سیکل (۲۲۸) **و غیره**	۵۴

به منظور تشریح هرچه بیشتر مجموعه‌داده در جدول یادشده برخی از شاخص‌های آماری تک‌متغیرهٔ مربوط به هر متغیر بیان شده است، تا بدین‌ترتیب مشخصه‌های مجموعه‌داده بیشتر توصیف شوند، به‌عنوان مثال، مشخص است که میانگین سنی افراد متهم موجود در این مجموعه‌داده در حدود ۳۴ سال است و جنسیت اکثر آن‌ها مرد می‌باشد.

۱-۱. پاک‌سازی و آماده‌سازی داده‌ها

به دلیل اینکه از ۹۹۹ رکورد موجود در این مجموعه‌داده ۹۷۹ رکورد مربوط به مردهاست و فقط بیست رکورد مربوط به زن‌ها بود، رکوردهایی که جنسیت آن‌ها زن بود، از مجموعه‌داده حذف شدند. به منظور ورود متغیرهای جمعیت‌شناختی به مرحلهٔ مدل‌سازی مقادیر برخی از متغیرها دوباره تعریف شدند. این کار به دلیل تشریح راحت‌تر نتایج خوشه‌بندی انجام شد. بدین‌ترتیب، مقادیر متغیرهای موجود در مجموعه‌داده به شکل زیر دوباره تعریف شدند.

جدول ۲ مقادیر متغیرهای موجود در مجموعه‌داده

نام متغیر	مقدار قبلی	مقدار فعلی
وضعیت تأهل	مجرد	۲
	متأهل	۱
وضعیت مسکن	مالک	۱
	پدری (مالک)	۲
	استیجاری	۳
	استیجاری (پدری)	۴
	بی‌خانمان	۵
شغل	پزشک،صنعتگر،کارگر ماهر	۱
	کارمند بخش خصوصی، کارمند بخش دولتی، معلم	۲
	بازنشسته، کارگر ساده، راننده	۳
	آزاد، مشاغل آزاد، مشاغل دیگر	۴
	دانشجو، محصل، دانش‌آموز	۵
	سرباز، خانه‌دار، ولگرد	۶
میزان تحصیلات	فوق لیسانس، لیسانس، فوق دیپلم	۱
	دیپلم، متوسطه	۲
	سیکل، راهنمایی، ابتدایی	۳
	بی‌سواد	۴

در این امتیازدهی تلاش شده است که باتوجه‌به وخامت، هر متغیر از نظر وضعیت اجتماعی امتیازها به ترتیب صعودی بالا رود؛ به‌عنوان نمونه، وضعیت مسکن فردی که بی‌خانمان می‌باشد، امتیاز ۵ و شخصی که خود مالک خانه است، دارای امتیاز ۱ است. همچنین، فردی که تحصیلات دانشگاهی دارد، امتیاز یک و فردی که بی‌سواد است، امتیاز پنج دریافت

کرده است. بدین‌ترتیب می‌توان اظهار نمود که هر فردی که امتیاز بیشتری را در هر متغیر کسب کرده باشد، وخامت او از نظر مشخصه‌های اجتماعی بیشتر از دیگر افراد موجود در مجموعه‌داده است.

۱-۲. جایگذاری مقادیر از دست رفته

همان‌طورکه در جدول شمارهٔ ۱ نیز مشخص است، برخی از متغیرها، از جمله متغیر شغل، میزان تحصیلات و وضعیت مسکن به ترتیب دارای ۱۱۵، ۵۴ و ۴۹ مقدار از دست رفته می‌باشند. در این مجموعه‌داده فقط دو متغیر سن و جنسیت مقدار از دست رفته ندارند.

به منظور جایگذاری مقادیر از دست رفته در این تحقیق از روش تخمین زدن استفاده شده است. بدین‌ترتیب که با استفاده از روش رده‌بندی k نزدیک‌ترین همسایگی تمام مقادیر از دست رفته تخمین زده شده و با مقدار مناسبی جایگذاری شده‌اند. در مدل k نزدیک‌ترین همسایگی، تعداد همسایگی به منظور پیش‌بینی مقادیر از دست رفته برابر یک قرار داده شده است. به این معنا که هر نمونهٔ از دست رفته با یک نزدیک‌ترین همسایگی خود جایگذاری شده است.

۱-۳. خوشه‌بندی داده‌ها

در این تحقیق به منظور خوشه‌بندی از یکی از پرکاربردترین روش‌های خوشه‌بندی، یعنی خوشه‌بندی k، نزدیک‌ترین همسایگی استفاده شده است. تنها پارامتر تنظیم شده برای این روش تعداد خوشه است. الگوریتم k میانگین به شرح زیر عمل می‌کند:

الف - ابتدا K نقطه به صورت تصادفی به‌عنوان مراکز خوشه‌ها انتخاب می‌شوند؛

ب - هر رکورد در مجموعه‌داده به خوشه‌ای که مرکز آن خوشه کم‌ترین فاصله تا آن رکورد را دارد، نسبت داده می‌شود. معیار محاسبهٔ فاصله در این مرحله هر معیاری می‌تواند باشد. این معیار با ماهیت مجموعه‌داده ارتباط تنگاتنگی دارد. مشهورترین معیارهای محاسبهٔ فاصلهٔ رکوردها در روش خوشه‌بندی، معیارهای فاصلهٔ اقلیدسی و فاصلهٔ همینگ هستند که به ترتیب در ۱ و ۲ به آن‌ها اشاره شده است:

$$(۱)\mathrm{d_E}(\mathrm{x,y}) = \sqrt{\sum_{k=1}^{n}(x_k - y_k)^2}\mathrm{d_E}(\mathrm{x,y}) = \sqrt{\sum_{k=1}^{n}(x_k - y_k)^2}$$

$$(۲)\mathrm{d_H}(\mathrm{x,y}) = \sum_{K=1}^{n}|\mathrm{x_K} - \mathrm{y_K}\mathrm{d_H}(\mathrm{x,y}) = \sum_{K=1}^{n}|\mathrm{x_K} - \mathrm{y_K}$$

در روابط یاد شده n بیانگر تعداد ابعاد مسئلهٔ خوشه‌بندی یا همان تعداد ویژگی‌هاست.

همچنین x_k و y_k هم به ترتیب، مبین K امین ویژگی‌های دو رکورد x و y است؛

ج- پس از تخصیص تمام رکوردها به یکی از خوشه‌های تشکیل شده، برای هر خوشه یک نقطهٔ جدید به‌عنوان مرکز محاسبه می‌شود؛

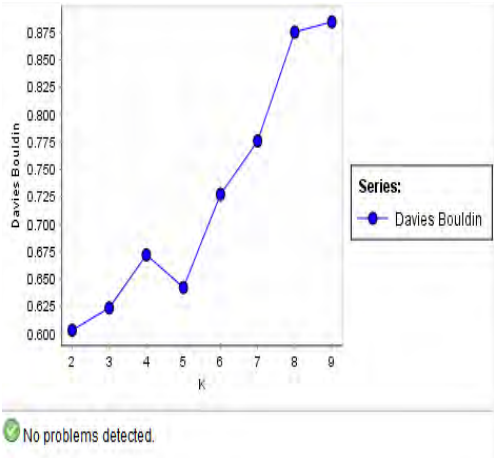
د- مراحل دو و سه تا زمانی‌که دیگر هیچ تغییری در مراکز خوشه‌ها حاصل نشود، تکرار خواهند شد.

در این تحقیق به منظور مشخص شدن تعداد خوشه‌های موجود در مجموعه‌داده از شاخص دیویس بولدین استفاده شده است. بدین‌ترتیب‌که تعداد خوشه‌ها از دو خوشه تا نُه خوشه اجرا شد و در هر مرحله شاخص دیویس بولدین برای ارزیابی صحت خوشه‌بندی محاسبه شد.

در جدول ۳ و نمودار ۱ مقدار محاسبه شدهٔ شاخص دیویس بولدین برای هر مرحله از خوشه‌بندی موجود می‌باشد. باتوجه‌به نتایج به دست آمده، مشخص است که در تعداد پنج خوشه شاخص دیویس بولدین در بهترین وضعیت به میزان ۶۴درصد قرار دارد. در تعداد پنج خوشه شاخص دیویس بولدین هم از تعداد چهار خوشه بهتر است و هم از تعداد شش خوشه و این موضوع به آن دلیل است که در پنج خوشه داده‌ها در یک الگوی محلی قرار گرفته‌اند.

جدول ۳ مقادیر متغیرهای موجود در مجموعه‌داده

تعداد خوشه	شاخص دیویس بولدین
۲	۰.۶۰
۳	۰.۶۲
۴	۰.۶۷
۵	۰.۶۴
۶	۰.۷۳
۷	۰.۷۸
۸	۰.۸۸
۹	۰.۸۸



شکل ۱ نمودار شاخص دیویس بولدین در خوشه‌بندی (K تعداد خوشه می‌باشد)

۱-۴تحلیل خوشه‌ها

باتوجه‌به اینکه در مرحلهٔ پیش مشخص شد که تعداد پنج خوشه مقدار مناسبی برای بخش‌بندی رکوردهای مربوط به متهمان است، در این مرحله داده‌ها به پنج خوشه خوشه‌بندی شدند. در جدول شمارهٔ ۳ نتایج مربوط مراکز هر خوشه مشخص شده است. در ادامه می‌توان با تحلیل اعضای موجود در هر خوشه ویژگی‌های افرادی که در آن خوشه قرار گرفته اند را بررسی نمود.

جدول ۴ مراکز خوشه‌ها (میانگین متغیرهای اعضا در هر خوشه)

متغیرها	خوشه_۰	خوشه_۱	خوشه_۲	خوشه_۳	خوشه_۴
استان	۱.۳۱	۱.۲۵	۱.۵۰	۱.۲۷	۱.۳۴
تحصیلات	۲.۹۷	۲.۷۰	۳.۵۳	۲.۷۴	۲.۹۳
شغل	۳.۹۰	۳.۹۶	۳.۷۴	۳.۸۹	۳.۸۷
مسکن	۲.۶۵	۲.۸۰	۲.۵۰	۲.۵۸	۲.۶۲
تأهل	۱.۱۴	۱.۶۳	۱.۰۶	۱.۲۸	۱.۰۷
سن	۳۸.۶۲	۲۲.۸۴	۶۴.۴۷	۳۰.۲۴	۴۸.۸۲
وخامت	**۱۱.۹۷**	**۱۲.۳۴**	**۱۲.۳۲**	**۱۱.۷۷**	**۱۱.۸۳**
تعداد اعضای هرخوشه	**۲۱۸**	**۲۵۶**	**۳۴**	**۳۳۷**	**۱۳۴**

در این جدول با جمع مقادیر میانگین متغیرهای استان، تحصیلات، شغل، مسکن و تأهل، وخامت هر خوشه بررسی‌پذیر است. همان‌طورکه مشخص می‌باشد، اعضای موجود در خوشهٔ

۱ و ۲ وخامت بیشتری را نسبت‌به خوشه‌های دیگر از نظر ۵ متغیر اشاره شده دارد. در ادامه اعضای موجود در هر خوشه، به‌صورت جداگانه، تفسیر و برچسب‌دهی شده‌اند.

- خوشهٔ شمارهٔ ۲

به‌طورمیانگین، اعضای موجود در این خوشه دارای مقدار ۱/۵ برای متغیر استان می‌باشند. باتوجه‌به اینکه رکوردهایی که در تهران قرار گرفته‌اند، مقدار استانی برابر ۱ و رکوردهایی که مربوط به افراد خارج از تهران هستند، دارای مقدار ۲ برای این متغیر می‌باشند، مشخص است که بیشتر افراد موجود در این خوشه خارج از تهران زندگی می‌کنند. همچنین، باتوجه‌به اینکه میانگین سطح تحصیلات در این خوشه برابر با ۳/۵۳ است، که نسبت‌به تمام خوشه‌ها مقدار بیشتری را به خود اختصاص داده است، می‌توان برداشت کرد که سطح تحصیلات افراد در این خوشه به‌صورت میانگین در وضعیت نامناسبی قرار دارد و بیشتر این افراد تحصیلاتی در سطح ابتدایی و راهنمایی دارند.

به‌طورمیانگین، میزان متغیر شغل، مسکن و تأهل در این افراد در وضعیت مناسب‌تری نسبت‌به افراد موجود در دیگر خوشه‌هاست و این موضوع به این معناست که افراد این خوشه از لحاظ شغلی و تأمین مسکن در وضعیت مناسبی قرار دارند. باید توجه داشت که اعضای موجود در این خوشه باتوجه‌به میانگین ۶۴/۴۷ سال مسن‌ترین گروه از متهمان می‌باشند.

- خوشهٔ شمارهٔ ۱

اعضای موجود در این خوشه، برخلاف اعضای خوشهٔ ۲، به‌طورمیانگین دارای سطح تحصیلاتی نسبت به تمام اعضا می‌باشند. همچنین، بیشتر این افراد مقیم تهران هستند. باتوجه‌به این موضوع که میانگین سن افراد موجود در این خوشه ۲۲/۸۴ سال است و جوان‌ترین خوشهٔ موجود در مجموعه‌داده می‌باشند، این موضوع حائز اهمیت است که این افراد از لحاظ شغل، مسکن و وضعیت تأهل بدترین وضعیت را در مقایسه با دیگر خوشه‌ها دارد. ممکن است که این دلیل مناسبی به منظور اعتیاد این دسته از افراد در جامعه باشد.

- خوشهٔ شمارهٔ ۴ و ۰

اعضای موجود در خوشهٔ ۰ و خوشهٔ ۴ از نظر وخامت وضعیت اجتماعی در یک سطح قرار گرفته‌اند. با این تفاوت که افراد موجود در خوشهٔ ۰ به‌طور میانگین ۳۸/۶۲ سال و نسبت به اعضای خوشهٔ ۴، که به‌طور میانگین دارای سنی برابر ۴۸/۸۲ می‌باشند، ۱۰ سال جوان‌ترند.

-خوشهٔ شمارهٔ ۳

باتوجه‌به میانگین وخامت وضعیت در اعضای این خوشه که به‌طور میانگین برابر ۱۱/۷۷ است و کمترین میزان در بین دیگر خوشه‌ها را به خود اختصاص داده است، می‌توان اظهار نظر کرد که افراد موجود در این خوشه از نظر شاخص‌های اجتماعی نسبت‌به افراد موجود در دیگر خوشه‌ها وضعیت مناسب‌تری دارند.

تحلیل بیشتر بر روی خوشه‌ها

در این مرحله به منظور تحلیل بیشتر افراد موجود در هر خوشه از دو متغیر نوع مادهٔ مخدر مصرفی و وضعیت اعتیاد استفاده شد. در جدول زیر امتیاز اختصاص داده شده به مقادیر این دو متغیر موجود است. باید توجه داشت که مواد مخدر صنعتی مانند شیشه و کراک امتیاز بیشتری را دریافت کرده‌اند و در این مرحله نیز امتیازدهی براساس وخامت صورت گرفته است.

جدول ۵- امتیازدهی به مقادیر متغیرها

نام متغیر	مقدار قبلی	مقدار فعلی
ماده مصرفی	شیشه	۴
	کراک، هروئین، قرص روان‌گردان	۳
	حشیش، گراس	۲
	سوخته، شیره	۱
وضعیت اعتیاد	معتاد است	۱
	معتاد نیست	۰

بدین‌ترتیب، میانگین نوع مادهٔ مصرفی و وضعیت اعتیاد در هر خوشه براساس جدول زیر استخراج شد:

جدول ۶ - مراکز خوشه‌ها (میانگین متغیرهای اعضا در هر خوشه)

متغیرها	خوشهٔ_۰	خوشهٔ_۱	خوشهٔ_۲	خوشهٔ_۳	خوشهٔ_۴
ماده	۳.۹۴۵	۳.۹۵۳	۳.۹۷۱	۳.۹۰۸	۳.۹۶۳
اعتیاد	۰.۸۹۹	۰.۷۹۷	۰.۸۵۳	۰.۸۱۹	۰.۹۱۸

همان‌طورکه مشخص است، اعضای موجود در خوشهٔ شمارهٔ ۴ امتیاز بیشتری از لحاظ مادهٔ مصرفی به خود اختصاص داده‌اند. به‌طوری‌که، افراد موجود در این خوشه بیشتر به مصرف مواد مخدر صنعتی تمایل دارند. همچنین، باتوجه‌به اینکه، به‌طورمیانگین، بیشتر افراد موجود در این خوشه معتادند، امتیاز اعتیاد در آن‌ها نیز بیشتر است. بدین‌ترتیب می‌توان به این نتیجه رسید که از نظر نوع مواد و وضعیت اعتیاد اعضای موجود در خوشهٔ ۴ در بدترین وضعیت قرار دارند. ازطرف‌دیگر، باتوجه‌به اینکه اعضای موجود در خوشهٔ ۳ از لحاظ وضعیت اعتیاد و نوع مادهٔ مصرفی تقریباً وضعیت مناسب‌تری را نسبت‌به اعضای موجود در خوشه‌های دیگر دارند، می‌توان به اعضای موجود در این خوشه را معتادین کم‌خطر نامید. بیان این نکته ضروری است که، افراد موجود در خوشهٔ شمارهٔ ۱ از لحاظ وضعیت بیشترین وخامت را داشتند، باتوجه‌به اینکه کمترین میانگین سنی را به خود اختصاص داده‌اند، درصد افرادی که در این خوشه قرار گرفته‌اند میانگین وضعیت اعتیاد کمتری در مقایسه با تمام خوشه‌های دیگر دارند. همچنین، نوع مادهٔ مصرفی در آن‌ها در حالت میانگین قرار دارد. بدین‌ترتیب می‌توان اظهار کرد که، اعضای جوان موجود در این خوشه اگر زودتر تحت درمان قرار بگیرند، راحت‌تر می‌توانند اعتیاد خود را ترک کنند. گفتنی است که، افرادی که در خوشهٔ شمارهٔ ۲ قرار گرفتند و میانگین سنی بسیار بیشتری نسبت‌به اعضای موجود در دیگر خوشه‌ها دارند، اعتیاد آن‌ها به مواد مخدر صنعتی نیز بیشتر از اعضای موجود در دیگر خوشه‌هاست.

نتیجه‌گیری

داده‌کاوی می‌تواند با استفاده از داده‌های موجود راهبردهای آیندهٔ سازمانی را تغییر دهد. بسیاری از سیاست‌هایی که در سازمان پلیس اتخاذ می‌شود براساس تجربه و اقتضای زمانی است و می‌تواند تحت تأثیر فشار رسانه‌ای یا متأثر از تصمیمات دیگر سازمان‌ها باشد، ولی اگر بخواهیم به روش کاملاً عملی و تحلیلی و اجراپذیر برسیم، باید مفاهیم پنهانی را که تنها از راه داده‌کاوی و سیستم‌های پردازش اطلاعات در دسترس را نیز مدّنظر قرار دهیم. واقعیت این است که در یک مقاله نمی‌توان تمامی آنچه را که ما در حوزهٔ پردازش اطلاعات تجربه کرده‌ایم را به نوشتار تبدیل کرد، تنها با مثال‌های ممکن می‌توانستیم مخاطب را در این فضا قرار دهیم. در حوزهٔ مواد مخدر فرایند داده‌کاوی می‌تواند بسیاری از هزینه‌های سازمان و کشور را پایین بیاورد و این تنها از راه جمع‌آوری صحیح اطلاعات و پردازش علمی آن‌ها حصول‌پذیر است.

منابع فارسی

- زعفریان، رضا و قاسم زعفریان (۱۳۸۰). مروری بر داده‌کاوی، فصلنامهٔ **صنایع**، ش ۲۹.
- شاه‌سمندی، پرستو (۱۳۸۴). داده‌کاوی در مدیریت ارتباط با مشتری، ، مجلهٔ **تدبیر**، ش ۱۵۶.
- صنیعی‌آباده، محمد، سینا محمودی و محدثه طاهرپرور(۱۳۹۱). داده‌کاوی کاربردی، نیاز دانش.
- گودرزی، حمیدرضا (۱۳۸۸). داده‌کاوی چیست، نشریهٔ گزیدهٔ مطالب آماری، مرکز آمار ایران، ش۵۲.
- مهریزی، حائری، علی اصغر (۱۳۸۲). **داده‌کاوی: مفاهیم، روش‌ها و کاربردها**، پایان‌نامهٔ کارشناسی ارشد، آمار اقتصادی و اجتماعی، دانشکدهٔ اقتصاد، دانشگاه علامه طباطبائی.
- مهمد، کانتاردزیک (۱۳۸۵). داده‌کاوی، ترجمهٔ امیر علیخانزاده، علوم رایانه۱۳۸۹.

منابع لاتین

- Eun-Jeong Cho And Others, Organizational Data Mining In Korea, Issues In Infor
- H.T. Roos, Data Mining For Intelligence Led Policing, 15th Acm Sigkdd International
- Conference on Knowledge Discovery and Data Mining (Kdd), Paris 2009
- Ha, S., Bae, S., & Park, S. (2000). Web Mining for Distance Education. In IEEE
- International Conference On Management Of Innovation And Technology (Pp. 715–719)
- http://www.dbmsmag.com, 1998
- mation Systems Volume Viii, No. 2, 2007
- Mohammed J. Zaki And Wagner Meira Jr, Data Mining And Analysis, Cambridge
- University Press,2014
- Shu-Hsien Lia, Pei-Hui Chu, Pei-Yuan Hsiao, Data Mining Techniques And Applications
- A Decade Review From 2000 To 2011, Expert Systems With Applications, Volume 39, Issue 12, 15 September 2012, Pages 11303–11311

- Weiss, S. H., & Indurkhya, N. (1998) Predictive Data Mining: A Practical Guide. San Francisco,

- CA: Morgan Kaufmann Publishers.

- Wiig, K. M. (1994). Knowledge Management, the central management .Expert Systems with

- Applications, Volume 10(Pp 32–45)

- Www.Accessegov.Org