داده‌کاوی در آمار رسمی

مطالعه موردی: بررسی الگوی مصرف خانوارهای شهری براساس طرح هزینه و درامد خانوار

مجری

عباس مرادی

همکاران

کاوه کیانی

اشکان شباک

ایوب فرامرزی

حامد لروند

پژوهشکده‌ی آمار

گروه پژوهشی پردازش داده‌ها و اطلاع رسانی

بهار ۱۳۹۸

به نام خداوند جان و خرد

پیش‌گفتار

عصر حاضر، عصر تغییرات سریع، پی‌درپی و متنوع است. همزمان با این تغییرات، به ویژه برخط شدن و به‌روز شدن سامانه‌ها و سیستم‌های ثبت و گرداوری داده‌ها، سازمان‌های آماری خصوصا سازمان‌های متولی آمار رسمی را با انباشت زیاد داده‌ها و اطلاعات آماری ساختار نیافته در قالب‌ها و فرمت‌های متنوع مواجه کرده است. در واقع امروزه «انفجار داده‌ها» به وضعیتی رسیده است که با آمار کلاسیک و روش‌های سنتی تجزیه و تحلیل آن امکان‌پذیر نیست. برطرف شدن این چالش بزرگ، محققان ریاضی، آمار و علوم کامپیوتر را بر آن داشت تا راهی برای تجزیه و تحلیل این داده‌ها کشف نمایند. مهمترین یافته و راه حل این چالش برای محققان علوم داده، روش‌ها و الگوریتم‌هایی است که امروزه «داده‌کاوی» نام گرفته است. سازمان‌های متولی آمار رسمی و در راس این سازمان‌ها، مرکز آمار ایران به عنوان یگانه مرجع آمارهای رسمی کشور نیز نیازمند دانشی است که به کمک آن داده‌ها را تجزیه و تحلیل نماید و در اختیار سیاست‌گذاران و مدیران ارشد کشور به منظور تدقیق نمودن تصمیم‌گیری‌ها قرار دهد. پر واضح است که دانش و خرد برگرفته از داده‌های خام و متنوع، تصمیم‌گیری و برنامه‌ریزی بهینه برای برنامه‌های کلان و آتی کشور را تسهیل می‌نماید. امروزه مهم‌ترین و در واقع تنهاترین ابزار برای رفع این چالش، استفاده از تکنیک‌های داده‌کاوی است. با استفاده از دانش داده‌کاوی، روابط کشف و الگوهای پنهان در داده‌ها آشکار می‌شوند. در واقع داده‌کاوی به دانشی بین‌رشته‌ی (ریاضی، آمار و علوم کامپیوتر) که هدف آن ایجاد ابزارهایی در جهت پردازش مه‌داده‌ها و استخراج الگوهای مخفی و قوانین حاکم بر آنهاست، اطلاق می‌شود. داده‌کاوی، در تمامی زمینه‌هایی که داده‌ها گرداوری شوند و نیازمند تجزیه و تحلیل باشند، نقش بسزایی خواهد داشت.

پژوهشکده‌ی آمار با توجه به رسالت خود در جهت تسهیل این مهم برای تجزیه و تحلیل داده‌های آمار رسمی اجرای طرح پژوهشی «داده‌کاوی در آمار رسمی» را در دستورکار خود قرار داد. هدف این طرح پژوهشی معرفی برخی از روش‌های داده‌کاوی و کاربردهای آن در آمار رسمی است. این پژوهش در گروه پژوهشی پردازش داده‌ها و اطلاع‌رسانی پژوهشکده‌ی آمار توسط آقای دکتر عباس مرادی به عنوان مجری و آقایان دکتر ایوب فرامرزی، حامد لروند، دکتر کاوه کیانی و دکتر اشکان شباک به‌عنوان همکار به انجام رسیده است، که از آنان صمیمانه تشکر و قدردانی می‌شود. همچنین از دکتر محمد شیری، به خاطر نظرات ارزشمندشان تشکر و قدردانی می‌شود.

گروه پژوهشی پردازش داده‌ها و اطلاع رسانی

پژوهشکده‌ی آمار

فهرست مطالب

فهرست جداول

فهرست شکل‌ها

۱

کلیات تحقیق

۱-۱- مقدمه

داده‌کاوی[1] بین محققان علوم آمار، کامپیوتر و ریاضی در جهت تحلیل داده‌ها و پایگاه‌های داده‌ای که شامل مه‌داده‌ها[2] و یا داده‌های با ابعاد بالا[3] هستند، به عنوان یکی از ابزارهای پرتوان شناخته شده است. بر همین اساس مرکز آمار ایران به عنوان یگانه مرجع آمار رسمی کشور روز به روز به اهمیت داده‌کاوی پی برده و این مهم را در اولویت‌های برنامه‌ای خود قرار داده است. پرواضح است که ایجاد یک نظام کارآمد و مؤثر در تولید و عرضه آمار از الزامات اولیه و ضروری در برنامه‌ریزی است و زیربنای برنامه‌ریزی مناسب، اطلاعات جامع، منسجم و به‌روز است. داده‌کاوی که به معنای کشف دانش و استخراج از مقادیر زیادی از داده‌های خام است. از این رو داده‌کاوی مفیدترین ابزار برای استفاده از تحلیل وقایع گذشته در پایگاه‌های داده‌ای مراکز آمار ایران خواهد بود. بدین معنی برای کشف و استخراج اطلاعات از آمار رسمی کشور که آمار مد نظر، سیاست‌گذاران، برنامه‌ریزان و مدیران کشور است نیازمند استفاده از داده‌کاوی خواهیم بود. هدف این طرح پژوهشی، معرفی برخی از روش‌های سودمند داده‌کاوی در آمار رسمی خواهد بود.

۱-۲- بیان مسئله

[1] Data Mining
[2] Big Data
[3] High Dimensional Data

آمار رسمی[4] به اطلاعات عددی گفته می‌شود که توسط دولت یا مراجع صلاحیت‌دار که در قوانین و مقررات مشخص هستند، تولید و منتشر می‌شود و اطلاعاتی را در مورد وضعیت عمومی کشور برای امور مدیریتی (برنامه‌ریزی، سیاست‌گذاری، و تصمیم‌گیری) به دست می‌دهد. در واقع آمار رسمی به اطلاعات عمومی مربوط می‌شود که به نفع جامعه و با بودجه دولتی تولید می شود. آمار رسمی برای همگان قابل دسترس خواهد بود. از این‌رو ذی‌نفعان را قادر می‌سازد تا با اتکا بر این آمارها، تصمیم‌های مهمی در زندگی شخصی و تجاری خویش انجام دهند. آمار رسمی مطابق با رده‌بندی‌های بین‌المللی و با رویکرد اصل بی‌طرفی، قابلیت اطمینان، اثربخشی، محرمانه بودن و شفافیت تولید و منتشر می‌شود. اما از یک سو حجم این اطلاعات بسیار زیاد و همواره در حال تغییر است. بدین معنی که این داده‌ها با حجم، تنوع و سرعت بالا در حال تولید هستند. در واقع اکثر سازمان‌های آماری دارای پایگاه‌های داده‌ای هستند که داده‌های آن از نظر حجم و تنوع بسیار گوناگون است. از این‌رو می‌توان گفت این سازمان‌ها با داده‌ها و اطلاعات آماری جمع آوری شده که دارای رشد حیرت‌آوری است مواجه شده‌اند به گونه‌ای که تحلیل داده‌ها با روش‌های کلاسیک برای آن‌ها امکان‌پذیر نخواهد بود. از سوی دیگر با گسترش روزافزون فناوری اطلاعات و تجهیز سازمان‌ها، شرکت‌های دولتی و خصوصی امکان جمع آوری اطلاعات دقیق و به روز فراهم آمده است، به‌طوری که مجموعه وسیع و حجیمی از داده‌ها شامل داده‌های ثبتی خلق گردیده است. این اطلاعات مد نظر مدیران و پژوهشگران بوده و از آن‌ها برای تهیه گزارش‌های مختلف و تصمیم‌گیری‌های روزمره و راهبردی استفاده می‌شود. استخراج اطلاعات از این داده‌ها نیازمند کاوش و جست‌وجوی اساسی است. تصمیم‌گیری و برنامه‌ریزی بهینه توسط سیاست‌گذاران کلان یک کشور نیازمند داشتن دانش مشتق شده از اطلاعات آماری و آمار رسمی کشور است. بدیهی است که ارتقا و توسعه علمی کشور نیازمند تصمیم‌گیری بهینه از الگوهای نهان در پایگاه‌های داده‌ای کشور است. از این‌رو سازمان‌ها و شرکت‌های دولتی نیازمند به کارگیری دانش داده‌کاوی و تکنیک‌های هوشمند آن در پایگاه‌های داده‌ای خود می‌باشند. یکی از مهمترین روش‌های داده‌کاوی الگویابی و مدلسازی است. داده‌کاوی داده‌ها را به منظور کشف و استخراج دانش مورد تحلیل و کندوکاوهای خودکار و نیمه خودکار قرار می‌دهد. با استقرار پایگاه داده در سازمان مربوطه، زیرساخت‌های امکان داده‌کاوی با ابزارهای مربوطه برای سازمان پیدا خواهد شد. امروزه استفاده از رایانه‌ها در تحلیل و ذخیره‌سازی داده‌ها نقش بسزایی دارند. پایگاه‌های داده شامل چند صد میلیارد رکورد ثبت شده وجود دارند که امکان تحلیل و استخراج اطلاعات با روش‌های معمول و کلاسیک آماری از این پایگاه داده‌ها مستلزم داشتن دانش و ابزارهای توانمندتر است. شدت رقابت در عرصه‌های علمی، اجتماعی، اقتصادی، سیاسی و نظامی نیز اهمیت سرعت یا زمان دسترسی به اطلاعات را دو چندان کرده است. نیاز به طراحی سیستمهایی که قادر به اکتشاف سریع اطلاعات بدون خطای انسانی باشد مورد علاقه مدیران عصر حاضر است. در حال حاضر، داده‌کاوی مهمترین فناوری برای بهره‌برداری موثر، صحیح و سریع از داده‌های حجیم بوده و اهمیت آن رو به فزونی است و به‌عنوان ابزاری توانمند نه تنها دسترسی به اطلاعات را تسهیل می‌سازد بلکه باعث می‌شود تا الگوها و مدل‌های پنهانی در دل این پایگاه‌های داده نمایان شود و اطلاعات مفید و قابل اعتمادی که تا کنون نهفته بوده را به آشکار سازد.

[4] Official Statistics

۱-۳- اهداف

هدف از این طرح پژوهشی معرفی روش‌های داده‌کاوی و به کارگیری برخی از این روش‌ها در آمارهای رسمی کشور است. پس از معرفی روش‌های داده‌کاوی، داده‌های هزینه و درامد خانوار به‌طور خاص مورد مطالعه قرار می‌گیرد. خانوارهای شهری در سال ۱۳۹۰ و درسال ۱۳۹۶ به خوشه‌های دهگانه تقسیم‌بندی می‌شوند و رفتار خوشه‌ها با استفاده از قوانین داده‌کاوی کشف می‌گردد و در نهایت وابستگی این خوشه‌ها بررسی می‌گردد.

۱-۴- پیشینه تحقیق

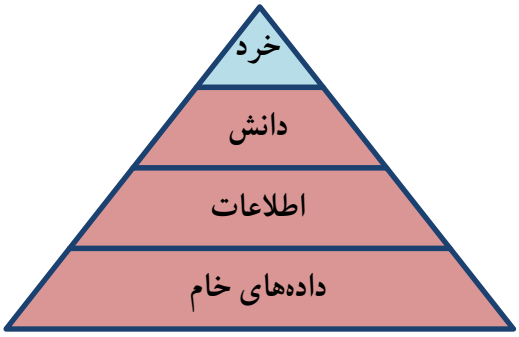
داده‌کاوی دانشی بین‌رشته‌ای است و ترکیبی از علومی مانند آمار، ریاضیات، کامپیوتر، علوم اطلاعات، نظریه پایگاه داده و یادگیری ماشین می‌باشد. در قرن هجدهم قضیه بیز توسط توماس بیز چاپ و منتشر گردید که بعدها احتمال شرطی نام گرفت. قرن بعد قرن پیدایش الگوریتم رگرسیون توسط گاوس بود که به کمک آن پیش‌بینی‌هایی از داده‌های خام برای امور آتی انجام می‌شد. در قرن بعد الگوریتم شبکه‌های عصبی توسط دانشمندان کشف و منتشر شد. این ایده برگرفته از شبکه‌های عصبی مغز بود. برای اولین بار در سال ۱۹۸۰ اصطلاح «داده‌کاوی» مطرح شد. در این دوره کارشناسان می‌توانستند به روابط معنادار پی ببرند. سپس، در سال ۱۹۸۹ اصطلاح «کشف دانش در پایگاه داده[5]» مطرح شد. در همین زمان اولین کارگاه آموزشی با نام KDD برای استخراج دانش در پایگاه‌های داده گوناگون، شروع به فعالیت نمود. در سال ۲۰۰۱ علم داده به عنوان یک رشته مستقل معرفی شد[6]. امروزه داده‌کاوی در اقتصاد، مهندسی، سیاست، پزشکی و... کاربرد فروانی دارد. داده‌کاوی تراکنش‌های مالی، سود سهام، شبکه‌های عصبی، امنیت ملی، الگوریتم‌های ژنتیک نمونه‌هایی از این کاربردهاست. در سال ۲۰۱۴ آقای حسنی و همکاران به بررسی اهمیت و ارزش کاربردهای داده‌کاوی در آمار رسمی پرداختند [۳]. در این طرح پژوهشی رفتار مصرفی خانوارها در سال‌های ۱۳۹۰ و ۱۳۹۶ بررسی و تغییرات آن با استفاده از تکنیک‌های داده‌کاوی نمایان خواهد شد.

۱-۵- روش تحقیق

آنچه امروزه اهمیت بسیاری پیدا کرده کمبود یا نبود داده‌های مورد نیاز تحقیق نیست بلکه کمبود یا نبود روش‌هایی مناسب و استاندارد به منظور نگهداری، به روز کردن، در دسترس قرار دادن و در حالت آرمانی‌تر، کشف دانش جدید در داده‌های موجود است. یکی از راهکارهای پیشنهادی برای حصول به این هدف، استفاده از روش‌های داده‌کاوی است. روش‌های داده‌کاوی این امکان را به کاربر می‌دهند که بتواند انبوه داده‌های جمع‌آوری شده را تحلیل و تفسیر نماید و الگوها و دانش نهفته در آن را استخراج کند. از آن‌جا که مراکز و مؤسسه‌های آماری با انبارداده‌های با حجم بالا سروکار دارند، خواسته یا ناخواسته نیازمند استفاده از این دانش سودمند می‌باشند. در حال حاضر، داده‌کاوی مهمترین فناوری برای بهره‌برداری

[5] Knowledge Discovery in Databases
[6] William S. Cleveland

موثر، صحیح و سریع از داده‌های حجیم بوده و اهمیت آن رو به فزونی است و به عنوان ابزاری توانمند نه تنها دسترسی به اطلاعات را تسهیل می‌سازد بلکه باعث می‌شود تا الگوها و مدل‌های پنهانی در دل این پایگاه‌های داده نمایان شود و اطلاعات مفید و قابل اعتمادی که تا کنون نهفته بوده را به آشکار سازد. همان‌طور که پیش‌تر اشاره شد، امروزه داده‌کاوی در دو بخش دولتی و بخش خصوصی نقش حیاتی دارد. در بخش خصوصی (بانکداری، بیمه، بهداشت، بازاریابی و ...) برای کاهش هزینه‌ها، ارتقاء کیفی پژوهش‌ها و بالاتر بردن میزان فروش از داده‌کاوی استفاده می‌نمایند. در بخش دولتی نیز به منظور برنامه‌ریزی و تعیین سیاست‌های کلان برای اداره جامعه و آن سازمان دولتی از داده‌کاوی استفاده می‌شود. ابزارهای داده‌کاوی ممکن است شامل مدل‌های آماری الگوریتم‌های ریاضی و روش‌های یادگیرنده باشد. داده‌کاوی منحصر به گردآوری و مدیریت داده‌ها نبوده و تجزیه و تحلیل اطلاعات و پیش‌بینی را نیز شامل می‌شود. داده‌کاوی با سه قسمت اولیه‌ی هرم دانش ارتباط تنگاتنگی دارد. این بخش‌ها در شکل ۱ به رنگ نارنجی نشان داده شده است.



شکل ۱-۱- هرم دانش

به کمک داده‌کاوی هم می‌توان روش‌های توصیفی مانند خوشه‌بندی، کشف قواعد پیوند و کشف الگوهای ترتیبی را انجام داد و هم می‌توان روش‌های پیش‌بینی کننده مانند رگرسیون، رده‌بندی را انجام داد. برخی از روش‌های مهم داده‌کاوی عبارتند از:

- رده‌بندی (Classification): رده‌بندی یعنی هر متغیری در یک کلاس خاص قرار گیرد. هر کلاس شامل مجموعه‌ای از متغیرهاست. یعنی نیازمند مدلی هستیم که کلاس متغیرها را به عنوان خروجی یک تابع توصیف نماید. این الگوریتم تحت نظارت کاربر صورت می‌پذیرد. مهم‌ترین الگوریتم‌های رده‌بندی، درخت تصمیم، شبکه عصبی و روش بیز می‌باشند.
- خوشه‌بندی (Clustering): گروه‌بندی مجموعه‌ای از اعضاء، رکوردها یا اشیاء به نحوی که اعضای موجود در یک خوشه بیشترین شباهت را به یکدیگر و کمترین شباهت را به اعضای خوشه‌های دیگر داشته باشند. این الگوریتم تا حدی بدون نظارت کاربر صورت می‌پذیرد.
- قواعد پیوند (Association Rules): قواعد انجمنی که به تحلیل سبد بازار نیز معروف است. الگوریتمی است که بر اساس آن وابستگی رویدادی به رویداد دیگر نمایش داده خواهد شد. به‌عنوان مثال خرید کره و مربا به هم وابسته‌اند.

- رگرسیون (Regression): پیش‌بینی یک متغیر براساس سایر متغیرها بر مبنای یک مدل وابستگی خطی یا غیرخطی، رگرسیون نام دارد.
- تحلیل‌های دنباله‌ی (Sequence Analysis): تجزیه و تحلیل رویدادهایی است که به صورت سلسه مراتبی رخ می‌دهد.

در واقع الگوریتم‌های داده‌کاوی مبتنی بر دو اصل است:

- الگوریتم‌های توصیفی: یافتن الگوها و روابط بین متغیرها، روش‌های مرتبط با آن.
- الگوریتم‌های پیش‌بینی: یافتن مدل‌های مناسب و نتایجی مفید برای تصمیم‌گیری‌های آتی.

در واقع به کمک داده‌کاوی هم می توان روش‌های توصیفی مانند خوشه‌بندی، کشف قواعد پیوند و کشف الگوهای ترتیبی را انجام داد و هم می توان روش‌های پیش‌بینی کننده مانند رگرسیون، دسته‌بندی را انجام داد. با توجه به ضرورت به‌کارگیری داده‌کاوی در آمار رسمی، در این طرح به طور خاص داده‌های هزینه و درامد خانوار مورد مطالعه قرار می‌گیرد. خوشه‌های ده‌گانه با استفاده از قوانین داده‌کاوی کشف می‌گردد و وابستگی این خوشه‌ها نیز بررسی می‌شود.