
بسم الله الرحمن الرحيم

دانشکده فنی و مهندسی
گروه مهندسی فناوری‌اطلاعات و ارتباطات

عنوان پایان نامه :

تحلیل داده‌های سرشماری عمومی سال ۱۳۸۵ کشور بر اساس روشهای داده‌کاوی

دانشجو:
نسترن فروغی

استاد راهنما :
دکتر حسن ابوالحسنی

استاد مشاور :
دکتر احمد فراهی

نگارش:
خرداد ۱۳۸۹

پایان نامه

برای دریافت درجه کارشناسی ارشد

در رشته مدیریت فناوری‌اطلاعات

شماره: .....
تاریخ: .....
پیوست: .....

جمهوری اسلامی ایران
وزارت علوم، تحقیقات و فناوری

مجتمع فنی و مهندسی

دانشگاه پیام نور
دانشگاه پیام نور استان تهران

تصویب نامه

پایان نامه کارشناسی ارشد رشته رشته مدیریت فناوری اطلاعات

تحت عنوان:

"تحلیل داده‌های سرشماری عمومی سال ۱۳۸۵ کشوربراساس روشهای داده کاوی"

تاریخ دفاع : ۸۹/٤/۲۰ **ساعت:** ۱۵-۱۳:۳۰

نمره: ۱۹/۵ نوزده و نیم

درجه ارزشیابی: عالی

هیات داوران:

داوران	نام و نام خانوادگی	مرتبه علمی	امضاء
استاد راهنما	دکتر حسن ابوالحسنی	دانشیار	
استادمشاور	دکتر احمد فراهی	استادیار	
استادداور	دکتر رضا عسگری مقدم	استادیار	
استادداورمدعو	دکتر لیلا شریف	استادیار	
نماینده گروه	دکتر قربان نیا	استادیار	

تهران،خیابان استاد نجات الهی،تقاطع خیابان سپند کوچه هشتم، پلاک ۱۱
تلفن: ۸۸۹۰۴۱۲۷
دورنگار: ۸۸۹۰۳۳۱۰
www.fani.tpnu.ac.ir
fani@tpnu.ac.ir

تقدیم به

همه عزیزانم

مَن لم یَشکُر المَخلوق لم یَشکُر الخالِق

بر خود فرض می دانم از زحمات خالصانه‌ی تمامی افرادی که بنده را در تدوین این اثر یاری نمودند صمیمانه تشکر و قدردانی نمایم.

- در وهله‌ی اول استاد راهنمای ارجمندم جناب آقای دکتر حسن ابوالحسنی و استاد مشاور محترم جناب آقای دکتر احمد فراهی که از ابتدا تا انتهای کار با راهنمایی‌های عالمانه خود موجب پیشرفت این پایان نامه شدند.

- مدیران محترم مرکز آمار ایران آقای نورالهی و آقای تهرانی که در تهیه داده و اطلاعات کمک شایانی به انجام این تحقیق داشتند.

- همچنین از استاد گرانقدر خانم دکتر سمیه علیزاده که با راهنمایی‌های خویش مشوق و راهبر من در این راه بودند.

چکیده:

در طول دهه گذشته حجم زیادی از داده‌ها در پایگاه داده‌ها انباشته و ذخیره شده‌اند و نتیجه این انباشتگی این است که سازمان‌ها در داده غنی، ولی در کسب دانش بسیار ضعیف می‌باشند. دانش داده‌کاوی سازمانها را قادر می‌سازد تا از سرمایه داده‌هایشان بهره برداری نمایند. داده‌کاوی با پردازش جامع داده و انجام فرایند تصمیم سازی از طریق استخراج دانش با ارزش از داده، تصمیم‌گیری را برای مدیران سازمان تسهیل می‌نماید. از طرفی سازمان‌ها و مؤسسه‌های آماری دارای داده انباره‌های حجیمی از اطلاعات هستند که از منابع مختلف و موضوع‌های متفاوت نشأت گرفته و جمع‌آوری شده‌اند. دراین خصوص داده‌کاوی به عنوان ابزاری توانمند نه تنها دسترسی به اطلاعات را تسهیل می‌سازد بلکه باعث می‌شود تا از دل این داده انباره‌ها اطلاعات مفید و قابل اعتمادی که تا کنون نهفته بوده را به دست آورد.

هدف از این تحقیق بررسی روشهای داده‌کاوی در استخراج الگوهای مناسب از داده‌های سرشماری عمومی نفوس و مسکن سال ۱۳۸۵ ایران می‌باشد. برای این منظور با توجه به نوع داده‌های موجود در سرشماری، تعدادی سناریو برای استخراج الگوها پیشنهاد گردید. برای هر سناریو، استفاده از روشهای خوشه‌بندی، قواعد وابستگی و طبقه‌بندی بررسی شد و روش مناسب انتخاب و بکار گرفته شد تا الگوهایی استخراج گردد. نتایج حاصل از این تحقیق نشان می‌دهد که در صورت تعریف مناسب سناریوها می‌توان از هر یک از تکنیک‌های خوشه‌بندی، قواعد وابستگی و طبقه‌بندی استفاده کرد. برای بکارگیری هر یک از این تکنیک‌ها نوع پیش‌پردازش داده‌ها و الگوریتم مورد استفاده در نتایج حاصله تأثیر بسزایی خواهد داشت که در این تحقیق مورد بررسی و تجزیه و تحلیل قرار گرفته است.

مدل‌های تولید شده در این تحقیق، الگوهایی در زمینه‌های خانواده‌های تک والده، معلولیت، مهاجرت، دختران و پسران مجرد بالاتر از متوسط سن ازدواج، کارآفرینان و مزدبگیران، و قدمت بناها و مصالح ساختمانی بکار رفته شده در آنها را فراهم می‌سازد.

کلمات کلیدی

داده‌کاوی، خوشه‌بندی، قواعد وابستگی، طبقه‌بندی، تحلیل داده‌های سرشماری عمومی

فهرست مطالب

عنوان **صفحه**

فصل سوم: روش تحقیق

فصل چهارم: تحلیل داده‌ها (مطالعات موردی)

فصل پنجم: نتیجه‌گیری و پیشنهادات

فهرست منابع

ضمائم

فهرست اشکال

شکل صفحه

فهرست جداول

جدول | صفحه

فصل اول

مقدمه

فصل اول

مقدمه

۱-۱ مقدمه

یکی از مشخصه‌های بارز عصر حاضر افزایش فرایند تولید و ذخیره‌سازی داده‌ها در تمامی عرصه‌های زندگی بشر است. این داده‌ها عموماً در قالب پایگاه داده ذخیره و بازیابی می‌شوند. میزان داده‌های خام ذخیره شده در پایگاه‌های داده روز به روز در حال افزایش است. با این وجود میزان اطلاعات و دانش قابل کسب از این داده‌ها بسیار ناچیز است. به طوری که بسیاری از دانشمندان از این مشکل به عنوان "وفور داده، ضعف در اطلاعات[۱]" یاد می‌کنند. رشد بیش از حد پایگاه داده در تمام عرصه‌های زندگی بشر منجر به افزایش تقاضا بر ابزارهایی شده است که داده‌ها را به دانشی[۲] سودمند و تکلیف‌گرا[۳] تبدیل نماید. در گذشته و تاحدی در حال از روشهای آنالیز داده، مانند آنالیز رگرسیون، طبقه‌بندی عددی، آنالیزهای چندبعدی، .... بدین منظور استفاده می‌شد. ولی تمامی این روشها عموماً گرایش به استخراج خصوصیات آماری و مقداری[٤] داده‌ها دارند. در این روش استخراج، هیچ تمرکز و توجهی بر روی سابقه[٥] داده‌ها وجود ندارد. لذا به منظور تحلیل می‌بایست مقدار زیادی از سوابق گذشته دخیل می‌شد تا قابلیت استنتاج حاصل و در نهایت دانش تولید شود. با این وجود این کار توسط تحلیلگر انجام می‌شد و نه ابزار، و دانش تولید شده مستقیماً حاصل آنالیز تحلیل‌گران از دانش گذشته بود.

---

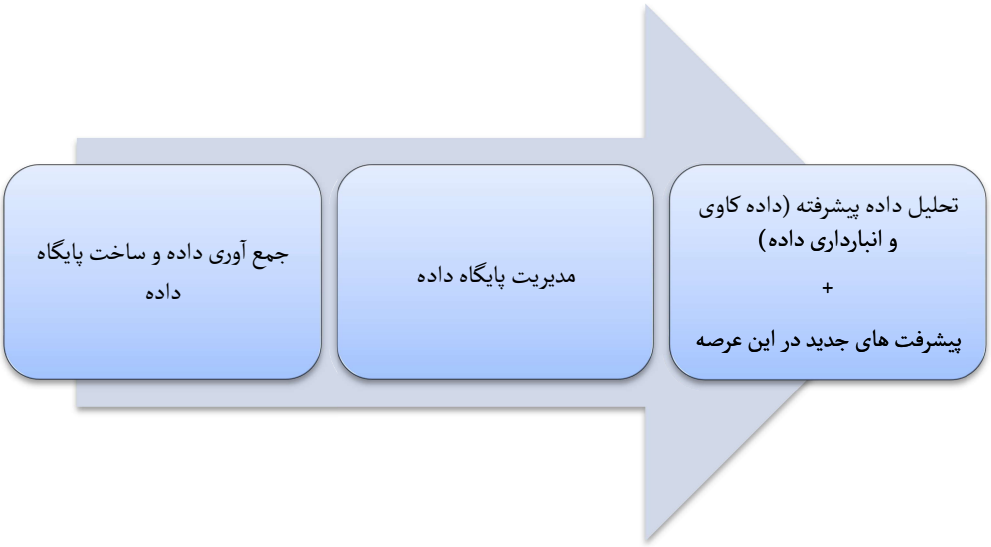
[۱] Data Rich, Information Poor
[۲] Knowledge
[۳] Task Oriented
[٤] Quantitative
[٥] Background

پس از تحقیقات بسیار دانشمندان دریافتند که کلید اصلی حل این مشکلات در یادگیری ماشین است. زیرا عصاره تمامی تحقیقات در این حوزه، ریشه در توسعه مدلهای محاسباتی به منظور کسب دانش با استفاده از حقایق[٦] و دانش گذشته[٧] دارد. به منظور رفع چنین مشکلاتی دانشمندان به جستجو و تفحص ایده‌ها و روش‌ها در علومی همچون یادگیری ماشین[٨]، الگوشناسی[٩]، آنالیز داده‌ها به روش آماری[١٠]، تجسم داده[١١]، شبکه های عصبی[١٢]و ... پرداختند. نتیجه این مطالعات، گسترش شاخه جدید و ترکیبی[١٣] از علوم مختلف به نام داده‌کاوی[١٤] و کشف دانش[١٥] شده است.

همانطور که بیان شد، در عصر اخیر با افزایش حجم عظیمی از داده‌ها روبرو هستیم. در این عصر به جای ذخیره داده‌ها مانند یک دایرةالمعارف، تمرکز بر روی داده‌های درست است. هم اکنون کشمکش زیادی بر روی ساختاردهی داده‌ها به روشی قانونمند وجود دارد. علاوه بر تمامی این موارد، می توانیم از این فعالیت‌های ذخیره شده از هر سازمان، به منظور برنامه‌ریزی استراتژیک[١٦] استفاده نماییم و همچون جستجوی تکه‌های طلا در معادن به کاوش در داده‌ها بپردازیم. این همان عرصه تحقیق داده‌کاوی و کشف دانش است که رشدی شگرف در چند سال اخیر داشته است. به طور کلی اهداف روشها و الگوریتم‌های آن در استخراج دانش مفید از حجم عظیم داده‌ها به صورت مستقیم، به شکل دانشی که روابط موجود مابین خصوصیات مورد انتظار را معین می‌نماید، یا غیر مستقیم، در غالب کارکردهایی که امکان پیشگویی[١٧]، طبقه‌بندی[١٨] و نمایش قواعد حاکم بر روی داده‌ها را فراهم می نماید، است. می‌توان داده‌کاوی را به عنوان نتیجه سیر تکاملی فناوری اطلاعات بیان کرد. همانند آنچه که در شکل ١-١ نمایش داده شده است، سیستم‌های پایگاه داده دارای روند تکاملی هستند که شامل چهار بخش کلی

[٦] Facts
[٧] Background Knowledge
[٨] Machine Learning
[٩] Pattern Recognition
[١٠] Statistical Data Analysis
[١١] Data Visualization
[١٢] Neural Networks
[١٣] Interdisciplinary
[١٤] Data Mining
[١٥] Knowledge Discovery
[١٦] Strategic Planning
[١٧] Predict
[١٨] Classify

جمع‌آوری داده و ساخت پایگاه داده[۱۹]، مدیریت پایگاه داده[۲۰]، تحلیل داده پیشرفته[۲۱] و پایگاه داده‌ها و سیستم‌های مدیریت پایگاه داده جدید (مانند پایگاه داده مبتنی بر وب، پایگاه داده بر مبنای مدلهای داده‌ای جدید مانند کاربردگرا[۲۲]، فضایی[۲۳]، چندرسانه‌ای، جریانی[۲۴] و ...) می‌باشند. همانطور که در تصویر نیز مختصر بیان شده است تحلیل پایگاه داده پیشرفته شامل داده‌کاوی و انبارداری داده[۲۵] است. در حقیقت با سیر تکاملی در تکنولوژی‌های پایگاه داده و پیشرفت تکنولوژی‌های سخت‌افزار و رکود قیمت رسانه‌های ذخیره‌سازی اطلاعات شکافی مابین حجم داده‌های تولید شده و اطلاعات و دانش به وجود آمد که این شکاف در دهه اخیر با ابزارهای داده‌کاوی در حال از میان برداشتن است.



شکل ۱-۱: روند تکاملی سیستم‌های پایگاه داده و جایگاه داده‌کاوی

[۱۹] Data Collection & Database Creation
[۲۰] Database management
[۲۱] Advanced Data Analysis
[۲۲] Application Oriented
[۲۳] Spatial
[۲۴] Stream
[۲۵] Data warehousing

۱-۲- تعریف مسأله و بیان موضوع اصلی تحقیق

سرشماری عمومی در ایران به موجب قانون هر ده سال یک بار به اجرا در می آید. نخستین سرشماری عمومی کشور در سال ۱۳۳۵ به وسیله اداره آمار عمومی و سرشماری‌های بعدی درسالهای ۱۳٤٥، ۱۳٥٥، ۱۳٦٥ و ۱۳۷٥ به وسیله مرکز آمار ایران انجام شد. سرشماری سال ۱۳۸٥، ششمین سرشماری عمومی است که در تاریخ ششم آبان ماه ۱۳۸٥ در سراسر کشور به اجرا در آمد. سرشماری با ارائه یک تصویر کلی از اندازه، ساختار و ویژگی‌های جمعیت یکی از منابع اصلی برنامه‌ریزی در زمینه‌های اقتصادی، اجتماعی و فرهنگی به شمار آمده و با ترسیم وضعیت موجود، آگاهی‌های لازم برای تهیه برنامه‌های توسعه را در اختیار سیاستگذاران، برنامه‌ریزان و مسئولان کشور قرار می‌دهد. همچنین سرشماری عمومی، با ارائه چارچوب‌های نمونه‌گیری لازم برای اجرای طرح های آمارگیری نمونه‌ای درحوزه‌های مربوط به جمعیت و خانوار، یکی از ارکان اصلی و زیربنایی نظام آماری کشور محسوب می‌شود.

با توجه به اینکه داده‌های سرشماری یکی از مهم‌ترین منابع کشوری محسوب می‌شود، جا دارد که با تحلیل داده‌های سرشماری ، تعدادی الگوی جالب و مفید بر اساس تکنیک‌های داده‌کاوی یافته تا مورد تجزیه و تحلیل قرار گرفته و در اختیار مسئولان و برنامه‌ریزان کشور قرار گیرد.

لذا در این تحقیق هدف اصلی بررسی تکنیک‌های مؤثر داده‌کاوی بر روی این داده‌ها، ساخت مدل‌های داده‌کاوی از روی آنها، تحلیل مدل‌های ایجاد شده و تفسیر آنها می‌باشد.

۱-۳- سؤالات تحقیق

۱- الگوریتم‌های مناسب طبقه‌بندی افراد چیست؟ چه پیش‌پردازش‌هایی مورد نیاز است؟
۲- برای ایجاد قواعد وابستگی بسته به ذات داده‌ها چه الگوریتم‌هایی را بکار ببریم؟
۳- چگونه می‌توان قواعد استخراج شده را تفسیر کرده و به آنها معنا بخشید؟
۴- از بین روشهای خوشه‌بندی، قواعد وابستگی و طبقه‌بندی کدام یک مناسب‌تر است؟