





دانشگاه پیام نور

دانشکده فنی و مهندسی

گروه مهندسی فناوری اطلاعات 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

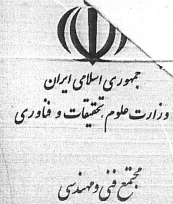
برای دریافت درجه کارشناسی ارشد

در رشته مدیریت فناوری اطلاعات

شماره:

تاریخ:

پیوست:



مجمع علمی و تخصصی

تصویب نامه

پایان نامه کارشناسی ارشد رشته رشته مدیریت فناوری اطلاعات
تحت عنوان:

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

ساعت: ۱۵-۱۳:۳۰

تاریخ دفاع: ۸۹/۴/۲۰

درجه ارزشیابی: عالی

نمره:

۱۶/۵
نمره ارزشیابی

هیات داوران:

امضاء	مرتبه علمی	نام و نام خانوادگی	داوران
	دانشیار	دکتر حسن ابوالحسنی	استاد راهنما
	استاد	دکتر احمد فراهی	استاد مشاور
	استاد	دکتر رضا عسگری مقدم	استاد داور
	استاد	دکتر لیلا شریف	استاد داور مدعو
	استاد	دکتر قربان نیا	نماینده گروه

تهران، خیابان استاد نجات
اللهی، تقاطع خیابان سپند
کوچه هشتم، پلاک ۱۱
تلفن: ۸۸۹۰۴۱۲۷
دورنگار: ۸۸۹۰۲۳۱۰
www.fani.tpnu.ac.ir
fani@tpnu.ac.ir

تقديم به

همه عزيزانم

مَنْ لَمْ يَشْكُرِ الْمَخْلُوقَ لَمْ يَشْكُرِ الْخَالِقَ

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 نوراللهی 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. در این 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 تک‌والده، 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

صفحه

عنوان

فصل اول: مقدمه

- ۱-۱ مقدمه ۲
- ۲-۱ تعریف مسأله و بیان موضوع اصلی تحقیق ۵
- ۳-۱ سؤالات تحقیق ۵
- ۴-۱ ضرورت و اهمیت انجام تحقیق ۶
- ۵-۱ فرضیات تحقیق ۶
- ۶-۱ اهداف و نوآوری تحقیق ۷
- ۷-۱ روش انجام تحقیق ۷
- ۱-۷-۱ روش گردآوری داده‌ها ۸
- ۲-۷-۱ جامعه آماری ۸
- ۳-۷-۱ فنون تجزیه و تحلیل اطلاعات ۸
- ۸-۱ ساختار پایان نامه ۸

فصل دوم: ادبیات تحقیق

- ۱-۲ مقدمه ۱۰
- ۲-۲ تاریخچه و سابقه داده‌کاوی ۱۰

- ۳-۲ مفهوم داده‌کاوی ۱۲
- ۴-۲ چه چیزی سبب پیدایش داده‌کاوی شده است؟ ۱۳
- ۵-۲ داده‌کاوی چه کارهایی نمی‌تواند انجام دهد؟ ۱۴
- ۶-۲ کاربرد داده‌کاوی در آمار رسمی ۱۴
- ۷-۲ کارهای مشابه در حوزه داده‌های سرشماری ۱۹
- ۸-۲ روش‌های داده‌کاوی ۲۲
- ۱-۸-۲ خوشه‌بندی ۲۲
- ۱-۱-۸-۲ روش‌های اعتبارسنجی خوشه‌بندی ۲۴
- ۲-۸-۲ قواعد وابستگی ۲۶
- ۳-۸-۲ طبقه‌بندی ۲۷
- ۱-۳-۸-۲ روش‌های ارزیابی طبقه‌بندی ۲۸
- ۲-۳-۸-۲ طبقه‌بندی مبتنی بر قواعد ۲۹
- ۹-۲ داده‌کاوی و انبار داده ۳۱
- ۱۰-۲ داده‌کاوی و OLAP ۳۴
- ۱۱-۲ داده‌کاوی و آمار ۳۵
- ۱۲-۲ پیشنهادی جهت انتخاب یک سیستم داده‌کاوی ۳۶
- ۱۳-۲ سرشماری عمومی نفوس و مسکن ۴۰

۴۱ ۱۴-۲ پرسشنامه سرشماری

۴۸ ۱۵-۲ جمع بندی

فصل سوم: روش تحقیق

۵۰ ۱-۳ مقدمه

۵۰ ۲-۳ متدلوژی CRISP-DM

۵۱ ۱-۲-۳ گام شناخت سیستم

۵۲ ۲-۲-۳ گام شناخت داده‌ها

۵۳ ۳-۲-۳ گام آماده سازی داده‌ها

۵۶ ۴-۲-۳ گام مدل سازی

۵۷ ۵-۲-۳ گام ارزیابی

۵۷ ۶-۲-۳ گام توسعه

۵۸ ۳-۳ متدلوژی بکار گرفته شده در این تحقیق

۵۹ ۱-۳-۳ فاز شناخت

۵۹ ۲-۳-۳ فاز آماده سازی داده‌ها

۶۲ ۳-۳-۳ فاز مدل سازی

۶۲ ۱-۳-۳-۳ خوشه بندی

۶۴ ۲-۳-۳-۳ قواعد وابستگی

۶۷ طبقه‌بندی ۳-۳-۳-۳
۶۸ فاز تجزیه و تحلیل ۴-۳-۳
۶۹ جمع‌بندی ۴-۳

فصل چهارم: تحلیل داده‌ها (مطالعات موردی)

۷۱ مقدمه ۱-۴
۷۱ هدف از تحقیق ۲-۴
۷۲ آماده‌سازی داده ۳-۴
۷۶ مدلسازی ۴-۴
۷۶ خوشه‌بندی ۱-۴-۴
۷۶ سناریو اول ۱-۱-۴-۴
۸۷ سناریو دوم ۲-۱-۴-۴
۹۱ قواعد وابستگی ۲-۴-۴
۹۱ سناریو سوم ۱-۲-۴-۴
۹۷ سناریو چهارم ۲-۲-۴-۴
۱۰۴ طبقه‌بندی ۳-۴-۴
۱۰۴ سناریو پنجم ۱-۳-۴-۴
۱۱۱ سناریو ششم ۲-۳-۴-۴

۴-۵ تجزیه و تحلیل ۱۱۹

۴-۶ جمع‌بندی ۱۲۱

فصل پنجم: نتیجه‌گیری و پیشنهادات

۵-۱ بررسی یافته‌های تحقیق ۱۲۵

۵-۲ محدودیت‌های تحقیق ۱۲۷

۵-۳ پیشنهاداتی برای تحقیقات آتی ۱۲۸

فهرست منابع

منابع فارسی ۱۳۱

منابع لاتین ۱۳۱

ضمائم

واژه‌نامه انگلیسی به فارسی ۱۳۴

واژه‌نامه فارسی به انگلیسی ۱۳۸

پیوست الف: معرفی مرکز آمار ایران ۱۴۳

پیوست ب: فرم سرشماری عمومی نفوس و مسکن سال ۱۳۸۵ ۱۴۹

پیوست پ: تعاریف و مفاهیم به کار رفته در فرم سرشماری سال ۱۳۸۵ ۱۶۶

فهرست اشکال

صفحه

شکل

-
- ۱-۱ روند تکاملی سیستم‌های پایگاه داده و جایگاه داده‌کاوی ۴
- ۱-۲ مراحل فرایند کشف دانش از پایگاه داده‌ها ۱۳
- ۲-۲ نمونه‌ای از خوشه‌بندی داده‌ها ۲۴
- ۳-۲ نمونه‌ای از انبار داده ۳۴
- ۴-۲ عوامل مؤثر در انتخاب سیستم نرم‌افزاری داده‌کاوی ۳۸
- ۱-۳ متدلوژی CRISP-DM ۵۱
- ۲-۳ متدلوژی تحقیق ۵۸
- ۱-۴ فرایند آماده‌سازی داده‌های زنان و مردان تک‌والده ۷۸
- ۲-۴ نمایش سه بعدی داده‌های زنان تک‌والده ۷۸
- ۳-۴ فرایند مدل‌سازی داده‌های زنان تک‌والده با الگوریتم *k-means* ۷۹
- ۴-۴ مشخصات خوشه‌های زنان تک‌والده ۸۰
- ۵-۴ محاسبه مقدار *k* بهینه با شاخص دیویس-بولدین برای خوشه‌های زنان تک‌والده ۸۱
- ۶-۴ چهار خوشه زنان تک‌والده ۸۲
- ۷-۴ فرایند مدل‌سازی داده‌های مردان تک‌والده با الگوریتم *k-means* ۸۳

- ۸-۴ مشخصات خوشه‌های مردان تک والده..... ۸۴
- ۹-۴ محاسبه مقدار k بهینه با شاخص دیویس- بولدین برای خوشه‌های مردان تک والده..... ۸۵
- ۱۰-۴ شش خوشه مردان تک والده..... ۸۶
- ۱۱-۴ فراوانی ۹ علت معلولیت در مناطق جغرافیایی کشور (۳۰ استان)..... ۸۸
- ۱۲-۴ فراوانی جنسیت افراد معلول..... ۸۸
- ۱۳-۴ فراوانی سطح تحصیلات افراد معلول..... ۸۸
- ۱۴-۴ فرایند خوشه‌بندی افراد دارای معلولیت با الگوریتم Twostep..... ۸۹
- ۱۵-۴ خوشه افراد دارای معلولیت..... ۸۹
- ۱۶-۴ نمودار توزیع علت مهاجرت..... ۹۲
- ۱۷-۴ رابطه علت مهاجرت و سطح تحصیلات..... ۹۳
- ۱۸-۴ رابطه علت مهاجرت و شغل..... ۹۳
- ۱۹-۴ فرایند قواعد وابستگی مهاجرین با الگوریتم GRI..... ۹۵
- ۲۰-۴ فرایند قواعد وابستگی دختران و پسران با الگوریتم GRI..... ۹۹
- ۲۱-۴ رابطه دختران مجرد با شغل، رشته‌ی تحصیلی و وضع فعالیت..... ۱۰۰
- ۲۲-۴ رابطه پسران مجرد با وضع شغلی، شغل و سطح تحصیلات..... ۱۰۲
- ۲۳-۴ مراحل آماده سازی داده و مدلسازی کارآفرینان و مزدبگیران..... ۱۰۶
- ۲۴-۴ صحت درستی مدل سناریوی پنجم با الگوریتم C5.0..... ۱۰۶

- ۲۵-۴ صحت درستی مدل سناریوی پنجم با الگوریتم CHAID ۱۰۷
- ۲۶-۴ درصد اهمیت فیلدهای ورودی سناریوی پنجم ۱۰۷
- ۲۷-۴ قوانین استخراج شده برای کارآفرینان و مزدبگیران ۱۰۸
- ۲۸-۴ نمودار درختی سناریوی پنجم ۱۱۰
- ۲۹-۴ مراحل آماده سازی داده و مدلسازی سناریوی ششم ۱۱۳
- ۳۰-۴ صحت درستی مدل سناریوی ششم با الگوریتم C5.0 ۱۱۳
- ۳۱-۴ صحت درستی مدل سناریوی ششم با الگوریتم QUEST ۱۱۴
- ۳۲-۴ صحت درستی مدل سناریوی ششم با الگوریتم CHAID ۱۱۴
- ۳۳-۴ صحت درستی مدل سناریوی ششم با الگوریتم C&R ۱۱۵
- ۳۴-۴ درصد اهمیت فیلدهای ورودی سناریوی ششم ۱۱۵
- ۳۵-۴ قوانین استخراج شده برای بناهای نوساز و کهنه ۱۱۶
- ۳۶-۴ نمودار درختی سناریوی ششم ۱۱۸

فهرست جداول

صفحه

جدول

۲۷	۱-۲ داده‌های خرید مشتریان از یک فروشگاه
۳۶	۲-۲ مقایسه روش‌های آماری و داده‌کاوی
۶۳	۱-۳ مشخصات سناریوی اول
۶۴	۲-۳ مشخصات سناریوی دوم
۶۵	۳-۳ مشخصات سناریوی سوم
۶۶	۴-۳ مشخصات سناریوی چهارم
۶۷	۵-۳ مشخصات سناریوی پنجم
۶۸	۶-۳ مشخصات سناریوی ششم
۷۲	۱-۴ شرح فیلدهای به کار رفته در سناریوها
۸۲	۲-۴ خلاصه‌ی خوشه‌های زنان تک‌والده
۸۶	۳-۴ خلاصه‌ی خوشه‌های مردان تک‌والده
۹۰	۴-۴ خلاصه‌ی خوشه‌های معلولین
۹۵	۵-۴ قوانین علت مهاجرت و تحصیلات و اشتغال
۱۰۰	۶-۴ قوانین دختران بالاتر از متوسط سن ازدواج
۱۰۲	۷-۴ قوانین پسران بالاتر از متوسط سن ازدواج

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^۴ داده‌ها دارند. در این روش استخراج، هیچ تمرکز و توجهی بر روی سابقه^۵ 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

^۴ Quantitative

^۵ 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^۶ و دانش گذشته^۷ دارد. به منظور رفع چنین مشکلاتی دانشمندان به جستجو و تفحص ایده‌ها و روش‌ها در علوم هم‌چون یادگیری ماشین^۸، الگوشناسی^۹، آنالیز داده‌ها به روش آماری^{۱۰}، تجسم داده^{۱۱}، شبکه‌های عصبی^{۱۲} و ... پرداختند. نتیجه این مطالعات، گسترش شاخه جدید و ترکیبی^{۱۳} از علوم مختلف به نام داده‌کاوی^{۱۴} و کشف دانش^{۱۵} 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 دایره‌المعارف، تمرکز بر روی داده‌های درست است. هم اکنون کشمکش زیادی بر روی ساختاردهی داده‌ها به روشی قانونمند وجود دارد. علاوه بر تمامی این موارد، می‌توانیم از این فعالیت‌های ذخیره شده از هر سازمان، به منظور برنامه‌ریزی استراتژیک^{۱۶} استفاده نماییم و هم‌چون جستجوی تکه‌های طلا در معادن به کاوش در داده‌ها پردازیم. 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^{۱۷}، طبقه‌بندی^{۱۸} و نمایش قواعد حاکم بر روی داده‌ها را فراهم می‌نماید، است. می‌توان داده‌کاوی را به عنوان نتیجه سیر تکاملی فناوری اطلاعات بیان کرد. همانند آنچه که در شکل ۱-۱ نمایش داده شده است، سیستم‌های پایگاه داده دارای روند تکاملی هستند که شامل چهار بخش کلی

^۶ Facts

^۷ Background Knowledge

^۸ Machine Learning

^۹ Pattern Recognition

^{۱۰} Statistical Data Analysis

^{۱۱} Data Visualization

^{۱۲} Neural Networks

^{۱۳} Interdisciplinary

^{۱۴} Data Mining

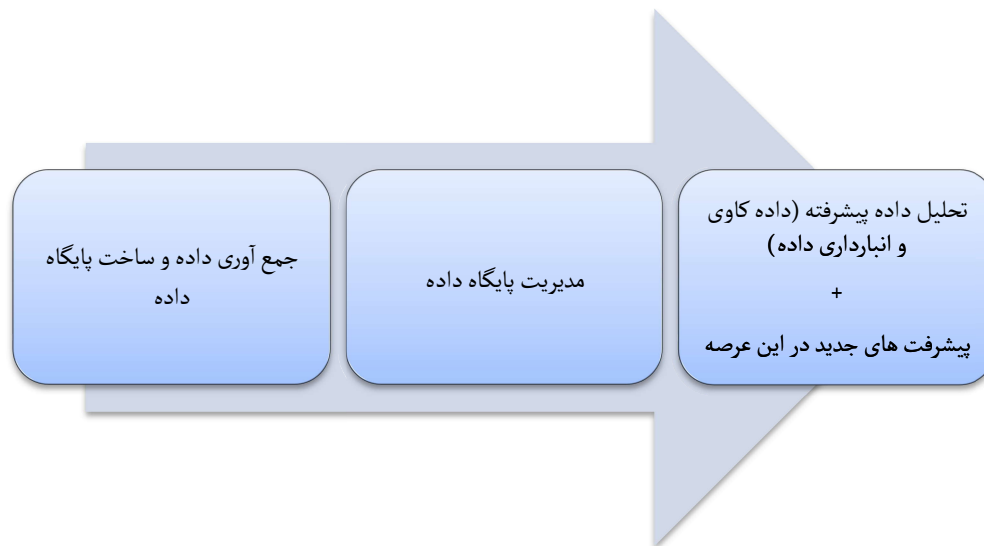
^{۱۵} Knowledge Discovery

^{۱۶} Strategic Planning

^{۱۷} Predict

^{۱۸} 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 در سالهای ۱۳۴۵، ۱۳۵۵، ۱۳۶۵ و ۱۳۷۵ به وسیله مرکز آمار ایران انجام شد. سرشماری سال ۱۳۸۵، ششمین سرشماری عمومی است که در تاریخ ششم آبان ماه ۱۳۸۵ 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 طرح‌های آمارگیری نمونه‌ای در حوزه‌های مربوط به جمعیت و خانوار، یکی از ارکان اصلی و زیربنایی نظام آمار 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؟