مجله تحقیق در عملیات در کاربردهای آن، سال دوازدهم، شماره اول (پیاپی ۴۴)، بهار ۹۴، ص ص ۱۱۱-۹۵
شاپا ۷۲۸۶-۲۲۵۱

طراحی مدل ترکیبی داده‌کاوی و تصمیم‌گیری چند معیاره (مورد مطالعه: بانک اطلاعات یارانه‌های مرکز آمار ایران)

عادل آذر*[1]، علیرضا مهدوی راد[2]، مرتضی موسی خانی[3]
۱- استاد، دانشگاه تربیت مدرس، گروه مدیریت صنعتی، تهران، ایران
۲- کارشناس ارشد، دانشگاه آزاد اسلامی، گروه مدیریت صنعتی، قزوین، ایران
۳ - دانشیار، دانشگاه آزاد اسلامی،گروه مدیریت دولتی، قزوین، ایران

رسید مقاله: ۱ خرداد ۱۳۹۳
پذیرش مقاله: ۱۰ مهر ۱۳۹۳

چکیده
حجم بالای داده‌ها در سازمان‌ها و جامعه‌ی امروزی حاصل پیشرفت‌های چشم‌گیر بشر در علوم کامپیوتری است. دغدغه‌ی مدیران امروز دیگر کمبود داده‌ها نیست؛ بلکه استفاده‌ی درست از داده‌های موجود و تبدیل آن‌ها به اطلاعات و سپس به دانش برای اتخاذ درست‌ترین تصمیمات با درصد خطای پایین‌تر می‌باشد. تکنیک‌های داده‌کاوی و روش‌های تصمیم‌گیری چند معیاره در دهه‌های اخیر هرکدام به شکلی کمک‌رسان مدیران در عرصه‌ی تصمیم‌گیری بوده است؛ اما این روش‌ها همواره نقاط ضعفی داشته‌اند؛ که موجب جستجوی پیوسته محققان در طراحی و کاربرد روش‌های جدید تصمیم‌گیری شده است. در پژوهش حاضر سعی بر آن داریم که با ارایه‌ی مدلی ترکیبی از داده‌کاوی و روش‌های تصمیم‌گیری چند معیاره، نتایج هر دو روش را بهبود دهیم. در این راستا مدل‌های تصمیم‌گیری چند معیاره می‌تواند نقش تصمیم‌گیرندگان را در نتایج مدل‌های داده‌کاوی، پررنگ‌تر کرده، از طرفی استفاده از تکنیک‌های داده‌کاوی امکان اجرای روش‌های تصمیم‌گیری چند شاخصه را بر روی حجم بالایی از داده‌ها امکان‌پذیر سازد. پس از ارایه‌ی چارچوب پیشنهادی برای مدل ترکیبی تحقیق یک مطالعه موردی بر روی پایگاه داده‌ای مناسب صورت گرفته است تا میزان کاربردی بودن چارچوب پیشنهادی مشخص گردد. این مطالعه‌ی موردی مربوط به خوشه‌بندی خانوارهای یکی از استان‌های کشور و رتبه‌بندی خانوارها با استفاده از مدل ترکیبی پیشنهادی می‌باشد.

کلمات کلیدی: داده‌کاوی، تصمیم‌گیری چند شاخصه، روش ویکور، خوشه‌بندی.

۱ مقدمه
در دنیای امروز مساله‌ی تصمیم‌گیری از مسایل بسیار مهمی است که ذهن اکثر مدیران را در سطوح مختلف به خود مشغول کرده است. در حل مسایل تصمیم‌گیری روش‌های بسیاری وجود دارد که

* عهده‌دار مکاتبات
آدرس الکترونیکی: AZARA@modares.ac.ir

روش‌های تصمیم‌گیری چند معیاره و تکنیک‌های داده‌کاوی از روش‌های پرکاربرد در این عرصه می‌باشد. مسایل بهینه‌سازی و تصمیم‌گیری زندگی واقعی نیازمند ابزارهای بهینه‌سازی چند معیاره است. با توجه به این نیاز، روش‌های تصمیم‌گیری چند معیاره از محبوبیت گسترده‌ای در علوم مختلف برخوردار می‌باشد. با روند رو به رشد کاربرد مدل‌های تصمیم‌گیری چند معیاره، افزایش تعداد اجزای تشکیل‌دهنده، متغیرها، پارامترها، اعداد ثابت و اهداف موجود در این فرآیند، این روش‌ها را بسیار پیچیده می‌کند. نسل جدید ابزارهای تصمیم‌گیری این فرآیند را هر چه بیش‌تر خودکار کرده است؛ اما شروع فرآیند و تنظیم ارزش‌های اولیه‌ی ابزارهای شبیه‌سازی و همچنین شناسایی اهداف و متغیرهای ورودی مؤثر برای رسیدن به فضای طراحی کوچک‌تر همچنان پیچیده می‌باشد [۱]. برای نمونه می‌توان به تحقیق آقای کریمی شیرازی و همکاران در صنایع غذایی ایران اشاره کرد. در تحقیق مذکور از روش ترکیبی DEMATEL،ANPو VIKOR استفاده شده است که هر کدام از این روش‌ها در جای خود پیچیدگی‌های خاص خود را دارا می‌باشد [۲].

روش دیگری که در تصمیم‌گیری‌های امروزی بسیار پرکاربرد می‌باشد داده‌کاوی است. تکنیک‌های داده‌کاوی، فرآیند تشخیص الگوها و مدل‌های موجود در داده‌هاست، الگوهایی که معتبر، بدیع، بالقوه، مفید و کاملاً قابل‌فهم می‌باشد. به‌عبارت‌دیگر، هدف تکنیک‌های داده‌کاوی یافتن الگوهای جالب موجود در پایگاه داده‌هاست که در میان حجم عظیمی از داده‌ها مخفی است[۳]. هرچند کاربرد تکنیک‌های داده‌کاوی در دنیای نرم‌افزار به‌عنوان الگوهای هوشمند حل مسایل دسته‌بندی، خوشه‌بندی، و مدل‌های پیش‌بینی، همچنان روبه‌پیشرفت و افزایش دارد؛ اما باید گفت که استفاده از این تکنیک‌ها در حل مسایل تصمیم‌گیری مدیران آن‌چنان‌که انتظار می‌رفته کاربردی نشده است. شاید بتوان گفت این دور ماندن تکنیک‌های داده‌کاوی و مدیران از یکدیگر به دلیل دیدگاه خودکار سازی هرچه بیش‌تر الگوریتم‌های داده‌کاوی در شاخه‌ی هوش مصنوعی باشد. حال‌آنکه در تصمیم‌گیری‌های مدیریتی نیاز به اعمال‌نظر مدیران در فرآیند تصمیم‌گیری است. در این تحقیق به دنبال آن هستیم تا با کاربرد تکنیک‌های داده‌کاوی و ترکیب آن‌ها با روش‌های تصمیم‌گیری چند معیاره، مدلی ترکیبی ارایه کنیم که خروجی هریک از این روش‌ها در مسایل تصمیم‌گیری را بهبود داده و تا حد امکان مدلی کاربردی باشد. نوآوری این تحقیق در استفاده از داده‌کاوری برای پیش‌پردازش داده‌های ورودی مدل‌های تصمیم‌گیری چندمعیاره و امکان کار با تعداد بالای داده‌ها در این مدل‌ها و از طرفی استفاده از مدل‌های تصمیم‌گیری چندمعیاره در جهت اعمال نظر مدیران در مدل‌های داده‌کاوی می‌باشد. مدل ارایه شده در مرکز آمار ایران و برای خوشه‌بندی و رتبه‌بندی خانوارهای یکی از استان‌های کشور مورد آزمایش قرارگرفته است.

۲ مفاهیم

۲-۱ کشف دانش و داده‌کاوی

همان‌طور که الکترون‌ها و امواج موضوع اصلی مهندسی برق شد، داده‌ها، اطلاعات و دانش نیز موضوع اصلی حوزه‌ی جدیدی از تحقیق و کاربرد به نام کشف دانش و داده‌کاوی است. به‌طورکلی،

داده‌کاوی را می‌توان به‌صورت مجموعه‌ای از مکانیسم‌ها و تکنیک‌ها که از طریق نرم‌افزار، اطلاعات مخفی‌شده در داده‌ها را استخراج می‌کند تعریف کرد [۴]. رایج‌ترین و پرکاربردترین تعریف از کشف دانش و داده‌کاوی تعریفی است که به فایاد و همکاران نسبت داده‌شده است " فرآیند شناخت و شناسایی الگوهای معتبر، بدیع، مفید و درنهایت قابل‌فهم درداده‌ها" [۵].

پرکاربردترین استاندارد رسمی برای داده‌کاوی استاندارد کریسپ (CRISP) است. به‌طورکلی استاندارد کریسپ دارای ۶ مرحله است.

شکل ۱. فرآیند کلی داده‌کاوی بر اساس استاندارد کریسپ

۲-۱-۱ خوشه‌بندی و الگوریتم k میانگین

خوشه‌بندی به معنای "تقسیم داده‌ها" به گروه‌های مشابه است. داده‌ها بر اساس اصل حداکثر کردن شباهت داخل گروهی و حداقل کردن شباهت بین گروهی، خوشه‌بندی می‌شود[۶].

الگوریتم k میانگین پارامتر k را به‌عنوان ورودی گرفته، مجموعه n شی را به k خوشه افراز می‌کند. به‌طوری‌که سطح شباهت داخلی خوشه‌ها بالا بوده و سطح شباهت اشیای بیرون خوشه‌ها پایین باشد. شباهت هر خوشه نسبت به متوسط اشیای آن خوشه، سنجیده شده که این متوسط، مرکز خوشه نیز نامیده می‌شود. ورودی و خروجی این الگوریتم به‌صورت زیر است:

ورودی: k به‌عنوان تعداد خوشه‌ها و یک پایگاه داده شامل n شیء

خروجی: یک مجموعه از k خوشه که معیار مربع خطا را حداقل می‌کند.

در عمل، این الگوریتم یک روش هیوریستیکی برای کاهش معیار مربع خطاست که در رابطه‌ی زیر آمده است:

$$E = \sum \sum |p - m_i|^2 \qquad (۱)$$

در این رابطه E مجموعه‌ی مربع خطا برای تمام اشیای پایگاه داده می‌باشد. p نقطه‌ای در فضاست که نمایانگر یک شیء می‌باشد و m_i میانگین خوشه‌ی C_i می‌باشد که نقطه‌ی p به آن متعلق است. (هم p و هم m_i چندبعدی هستند) [۶].

یکی از مهم‌ترین نقاط ضعف این روش این است که در برابر اغتشاشات و نقاط پرت حساس است؛ زیرا این داده‌ها به‌راحتی مراکز را تغییر می‌دهند و ممکن است نتایج مطلوبی حاصل نشود. به همین دلیل در این تحقیق سعی شده قبل از استفاده از این روش تا حد امکان نقاط پرت، حذف یا اصلاح شود. برای این منظور از الگوریتم آنومالی که ضمن خوشه‌بندی، نقاط پرت را هم مشخص می‌کند و روش‌های آماری استفاده‌شده است که در پیاده‌سازی مدل توضیح داده خواهد شد.

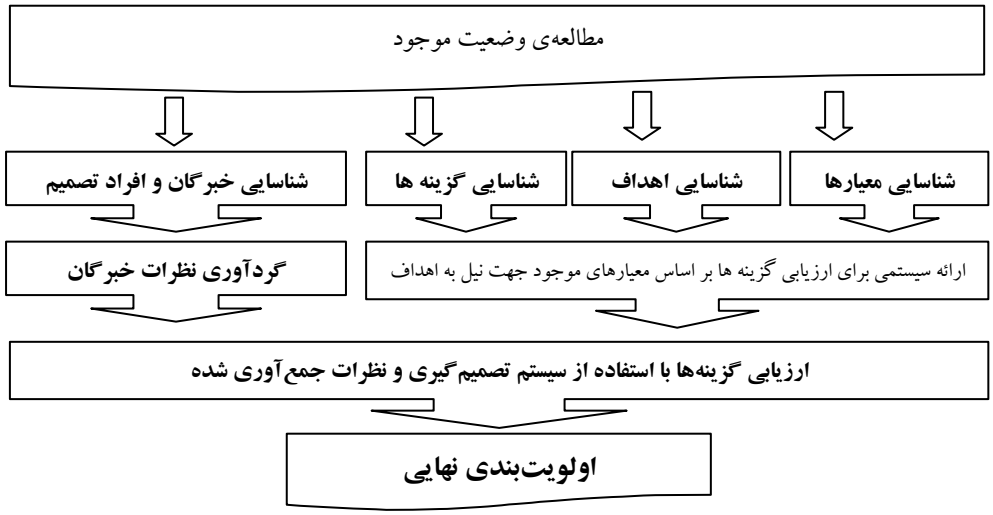
۲-۱-۲ دسته‌بندی

دسته‌بندی، فرآیند یافتن مدلی است که با تشخیص دسته‌ها یا مفاهیم داده می‌تواند دسته‌ی ناشناخته‌ی اشیای دیگر را پیش‌بینی کند. دسته‌بندی یک تابع یادگیری است که یک‌قلم داده را به یکی از دسته‌های از قبل تعریف‌شده نگاشت می‌کند. داده‌های موجود به دو قسمت آموزش و آزمون تقسیم می‌شوند. داده‌های آموزش برای یادگیری قواعد توسط سیستم استفاده می‌شود و داده‌های آزمون برای بررسی دقت دسته‌بندی و جلوگیری از بیش بردازش به کار می‌روند. یکی از روش‌های متداول دسته‌بندی تکنیک درخت تصمیم و الگوریتم کارت می‌باشد [۴].

در واقع الگوریتم درخت تصمیم‌گیری با انتخاب آزمونی شروع می‌شود که بهترین جداسازی را برای دسته‌ها انجام دهد. در مراحل بعدی درخت هم، همین کار برای گره‌های بعدی با داده‌های کم‌تر صورت می‌گیرد تا بهترین قانون‌ها حاصل شود. درخت آن‌قدر بزرگ می‌شود تا دیگر نتوان جداسازی بهتری برای داده‌های گره انجام داد [۷].

۲-۲ تصمیم‌گیری چند معیاره و روش ویکور

به‌طورکلی فرآیند تصمیم‌گیری‌چندمعیاره عبارت است از طریقه‌ی پیدا کردن وزن‌ها، رتبه‌ها و یا اهمیت یک مجموعه از فعالیت‌ها با توجه به تأثیر آن‌ها بر وضعیت موجود و هدف تصمیم. فرآیند کلی تصمیم‌گیری چندمعیاره در شکل(۲) نشان داده‌شده است [۸].



شکل ۲. فرآیند کلی تصمیم‌گیری چندمعیاره

۲-۲-۱ روش ویکور

ویکور یک روش تصمیم‌گیری چند معیاره‌ی توافقی است که توسط آپریکوویچ و زنگو بر مبنای روش ال پی متریک توسعه‌یافته است [۹].

این روش می‌تواند یک مقدار بیشینه‌ی مطلوبیت گروهی برای اکثریت و یک کمینه‌ی تأثر انفرادی برای مخالفت را فراهم نماید.

مراحل کاری روش ویکور

مراحل این روش شامل گام‌های ذیل است [۹]:

۱. محاسبه‌ی مقادیر نرمال شده

۲. تعیین بهترین و بدترین مقدار

بهترین و بدترین هریک از مقادیر در هر معیار را شناسایی می‌کنیم و به ترتیب f_j^* و f_j^- می‌نامیم.

$$f_j^* = \quad Max\ f_{ij} \quad , i = ۱,۲,...,m \qquad (۲)$$

$$f_j^- = \quad Min\ f_{ij}, j = ۱,۲,...,n \qquad (۳)$$

۳. تعیین وزن معیارها

اوزان معیارها باید برای بیان اهمیت روابط آن‌ها محاسبه‌شده باشد. برای این منظور، در این تحقیق از روش تحلیل سلسله مراتبی استفاده شده است.

۴. محاسبه‌ی فاصله‌ی گزینه‌ها از راه‌حل ایده آل

این مرحله محاسبه‌ی فاصله‌ی هر گزینه از راه‌حل ایده‌آل و سپس حاصل جمع آن‌ها برای ارزش نهایی بر اساس روابط ذیل است:

$$S_i = \sum_{j=۱}^{n} w_i (f_j^* - f_{ij}) / (f^* - f_j^-) \qquad (۴)$$

$$R_i = Max\,[w_i (f_j^* - f_{ij}) / (f^* - f_j^-)] \qquad (۵)$$

جایی که S_i بیانگر نسبت فاصله‌ی گزینه iام از راه‌حل ایده آل مثبت (بهترین ترکیب) و R_i بیانگر نسبت فاصله‌ی گزینه iام از راه‌حل ایده آل منفی (بدترین ترکیب) می‌باشد.

۵. محاسبه‌ی مقدار ویکور و رتبه‌بندی گزینه‌ها بر اساس این مقدار Q_i

این مقدار برای هریک از iها به‌صورت زیر تعریف می‌شود:

$$Q_i = v\left[\frac{S_i - S^*}{S^- - S^*}\right] + (۱ - v)\left[\frac{R_i - R^*}{R^- - R^*}\right] \qquad (۶)$$

درجایی که $S^- = \underset{i}{Max}\, S_i$ ، $S^* = \underset{i}{Min}\, S_i$ ، $R^- = \underset{i}{Max}\, R_i$ و $R* = \underset{i}{Min}\, R_i$ و v وزن استراتژی اکثریت موافق معیار یا حداکثر مطلوبیت گروهی است.

۳ پیشینه‌ی تحقیق

موسوی [۱] در تحقیق خود با عنوان" پیش‌پردازشی بر روش‌های تصمیم‌گیری چند معیاره با استفاده از ابزارهای داده‌کاوی" به لزوم کاربرد ابزارهای بهینه‌سازی چندهدفه در مسایل تصمیم‌گیری زندگی امروزی می‌پردازد. وی در ادامه به کاربرد روزافزون روش‌های تصمیم‌گیری چند معیاره در این راستا اشاره‌کرده، پیچیدگی این روش‌ها به دلیل تعداد رو به افزایش اجزای تشکیل‌دهنده، متغیرها، پارامترها، اعداد ثابت و اهداف موجود در این روش‌ها را مشکل اصلی آن‌ها می‌داند.

طلوع و همکاران [۳] در مقاله‌ای با عنوان"روشی جدید برای رتبه‌بندی الگوهای شناسایی‌شده در داده‌کاوی با استفاده از تحلیل پوششی داده‌ها" کاربرد روزافزون داده‌کاوی در تجارت را بیان می‌کنند. در استفاده از تکنیک‌های داده‌کاوی الگوهای مختلفی ممکن است به دست بیاید که تنها تعداد کمی از الگوها در اجرا استفاده می‌شود. ارزیابی و رتبه‌بندی جذابیت یا کارایی الگوهای وابستگی در داده‌کاوی نقش مهمی را ایفا می‌کند.

برخی از تحقیقات انجام‌شده در این شاخه به همراه رویکرد هر یک در جدول ۱ آورده شده است.

جدول ۱. برخی از تحقیقات انجام‌شده با روش‌های داده‌کاوی و تصمیم‌گیری چند معیاره [۶]، [۱۰-۱۸]

محقق	زمینه تحقیق
موسوی(۲۰۱۰)، هاله و همکاران(۲۰۱۲)	کاربر داده‌کاوی در پیش‌پردازش داده‌های ورودی روش‌های تصمیم-گیری چندمعیاره
مستروجیانیس و همکاران(۲۰۰۹)، چویی و همکاران (۲۰۰۵)، موآتا و بریسن(۲۰۰۴)	تصمیم گیری چند معیاره نا رتبه ای در بهبود دقت الگوریتم های طبقه بندی داده کاوی
بابایی و همکاران(۱۳۹۰)	کاربرد روش های تصمیم گیری چند معیاره در پیش پردازش داده ها در الگوریتم های داده کاوی
فروغی(۲۰۱۱)، طلوع و نلچیگر(۲۰۱۱)، طلوع و همکاران(۲۰۰۹)	ارزیابی و رتبه بندی جذابیت یا کارایی الگوهای ارتباطی درداده کاوی با استفاده از روش های تصمیم گیری چند معیاره و تحلیل پوششی داده ها
راد و همکاران(۲۰۱۱)	مقایسه روش های تصمیم گیری چند معیاره با تکنیک های داده کاوی در رتبه بندی تعداد زیادی از گزینه ها
لیو و شیه (۲۰۰۵)	روش های تصمیم گیری چند معیاره با تکنیک های داده کاوی در سیستم های پیشنهاد محصول

٤ متدولوژی تحقیق

متدولوژی استفاده‌شده در این تحقیق شامل دو رویکرد داده‌کاوی و تصمیم‌گیری چندمعیاره برای خوشه‌بندی و رتبه‌بندی حجم بالایی از رکوردها می‌باشد. چارچوب مدل ارایه‌شده در این تحقیق در شکل (۳) آورده شده است. مراحل اجرای مدل بر اساس استاندارد کریسپ در داده‌کاوی که پیش‌تر بیان شد، صورت گرفته است که در سه مرحله زیر پیاده‌سازی می‌گردد:

مرحله‌ی ۱- شناخت و انتخاب داده‌ها
مرحله‌ی ۲- پیش‌پردازش داده‌ها
مرحله‌ی ۳- مدل‌سازی و تجزیه‌وتحلیل داده‌ها و نتیجه‌گیری

٤-١ شناخت و انتخاب داده‌ها

در مرحله نخست متغییرهای موجود در پایگاه داده شناسایی‌شده(این متغییرها در سال های گذشته توسط کارشناسان و خبرگان مرکز آمار ایران شناسایی شده است) و با توجه به نظر خبرگان و با استفاده از فرآیندهای تصمیم‌گیری چندمعیاره متغیرهای تأثیرگذار مشخص و تعریف می‌گردد.

پس از آماده شدن متغیرهای موردنظر و پایگاه داده، با استفاده از روش‌های تصمیم‌گیری چند معیاره مانند تحلیل سلسله مراتبی و با استفاده ازنظر خبرگان و کارشناسان مربوط ۶ متغیر که بیش‌ترین تأثیر را بر نتیجه‌ی تحقیق حاضر دارد مشخص و اطلاعات خانوارها در مورد این ۶ متغیر به‌عنوان پایگاه داده جدید وارد فرآیند داده‌کاوی شده است.

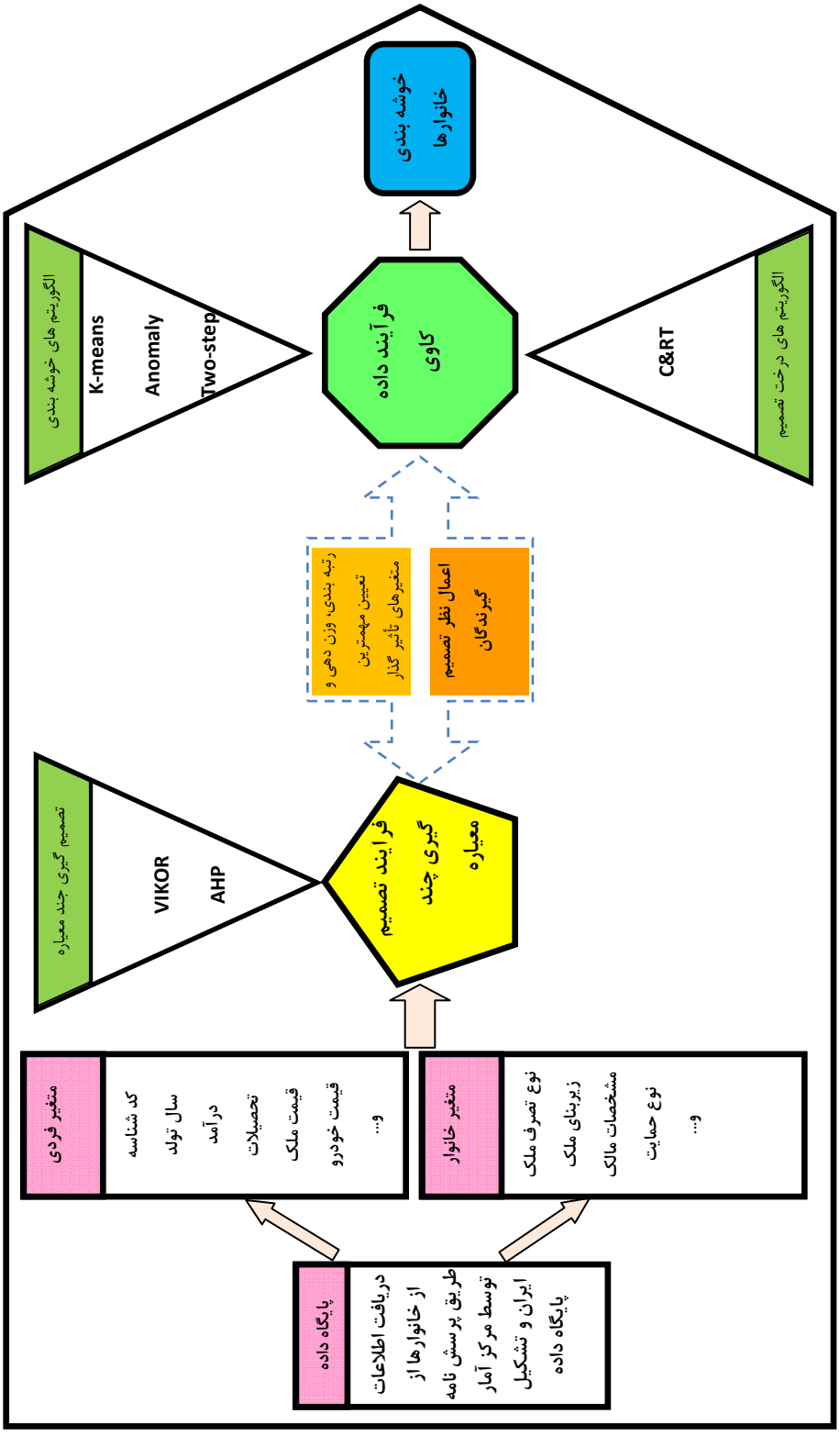
٤-٢ پیش‌پردازش داده‌ها

پیش‌پردازش شامل آماده‌سازی و نرمال‌سازی داده می‌شود. در بخش آماده‌سازی داده‌ها و برخورد با داده‌های گمشده، رکوردهای با بیش از ۵۰ درصد اطلاعات خالی حذف و در مورد سایر فیلدهای خالی، با توجه به متغیر مربوط تخصیص مقدار ثابت و یا استفاده از الگوریتم‌های پیش‌بینی کننده در داده‌کاوی در نظر گرفته‌شده است. در بخش دیگر آماده‌سازی داده‌ها که مربوط به برخورد با داده‌های غیر نرمال می‌باشد از الگوریتم آنومالی برای شناسایی داده‌های غیر نرمال استفاده‌شده و پس از شناسایی این داده‌ها رفتار مناسب در مورد آن‌ها صورت گرفته است. قبل از شروع اجرای مدل نیاز به نرمال‌سازی داده‌ها داریم. در تحقیق حاضر نرمال‌سازی داده‌ها با استفاده از روش Z استاندارد صورت گرفته است. این روش تمامی داده‌ها را به‌صورت اعدادی بین ۴ و ۴- تبدیل می‌کند.

٤-٣ مدل‌سازی

در این مرحله وزن متغیرها با توجه به نتایج به‌دست‌آمده از پرسش‌نامه‌ها و نظر خبرگان و با استفاده از روش تحلیل سلسله مراتبی محاسبه‌شده و برروی داده‌های نرمال شده اعمال می‌گردد. سپس با توجه به شاخص انحراف استاندارد و شاخص سیلوئت و تحلیل الگوریتم خوشه‌بندی دومرحله‌ای تعداد خوشه‌ی مناسب برای خوشه‌بندی به روش کای میانگین مشخص‌شده و خوشه‌بندی با توجه به مقادیر وزن‌دار شده صورت می‌پذیرد. در آخر، خوشه‌های به‌دست‌آمده با استفاده از روش ویکور، رتبه‌بندی شده، تجزیه‌وتحلیل نهایی با توجه به خوشه و رتبه‌ی هر رکورد صورت می‌گیرد.

در ادامه در قالب یک مطالعه‌ی موردی به توضیح و تشریح مراحل انجام تحقیق پرداخته خواهد شد.



شکل ۳. چارچوب مدل تحقیق

۵ مطالعه‌ی موردی

مطالعه‌ی موردی این تحقیق مربوط به بانک اطلاعات یارانه‌های مرکز آمار ایران می‌باشد. پایگاه داده یکی از استان‌های کشور به‌عنوان جامعه‌ی آماری استفاده شده است.

۵-۱ مرحله‌ی شناخت و انتخاب داده‌ها

در مرحله‌ی نخست اطلاعات و فرم‌های بانک اطلاعات یارانه‌های خانوارهای یکی از استان‌های کشور در مرکز آمار ایران مورد بررسی قرارگرفته است. با توجه به اطلاعات موجود و طبق نظر خبرگان، ۲۴ متغیر جهت خوشه‌بندی خانوارها شناسایی شدند، شش متغیر تأثیرگذار و موردنیاز برای خوشه‌بندی با توجه به نظر خبرگان مرکز آمار ایران و با استفاده از پرسشنامه‌ی مقایسات زوجی، مشخص‌شده، اطلاعات هر خانوار در مورد این شش متغیر از بانک‌های اطلاعاتی موجود استخراج و به‌صورت فیلدهای جدیدی در جداول بانک اطلاعاتی تعریف و ذخیره‌سازی گردید. این شش متغیر شامل موارد زیر می‌شود:

- تحصیلات
- قیمت خودرو
- تعداد اعضا
- قیمت مسکن
- اقساط ماهیانه
- درآمد خانوار

۵-۲ مرحله‌ی پیش‌پردازش داده‌ها

۵-۲-۱ آماده‌سازی داده‌ها

در این مرحله ابتدا برخورد مناسب با داده‌های گمشده صورت گرفته است. در حالت اولیه تنها حدود ۱۷ درصد از فیلدهای اطلاعاتی پر می‌باشد که برای شروع داده‌کاوی این درصد باید به ۱۰۰ درصد افزایش یابد. برای این منظور از میان صدوپنجاه هزار خانوار استان، کلیه خانوارهایی که بیش از ۵۰ درصد اطلاعاتشان در بانک اطلاعاتی موجود نمی‌باشد از بانک اطلاعاتی کنار گذاشته می‌شود. با این کار درصد فیلدهای اطلاعاتی پر به ۴۹ درصد افزایش می‌یابد.در این مرحله تعداد خانوارهای باقیمانده در بانک اطلاعاتی یکصدودو هزار خانوار می‌باشد. در مرحله‌ی بعد برای فیلدهای خالی مربوط به متغیرهایی مانند قیمت خودرو و قیمت مسکن مقدار ثابت صفر در نظر گرفته‌شده است(چراکه این فیلدهای خالی مربوط به خانوارهایی بوده است که دارای مسکن یا خودروی شخصی نبوده‌اند).

در مورد فیلدهای خالی مربوط به متغیری مانند درآمد خانوار از الگوریتم‌های پیش‌بینی کننده در داده‌کاوی و مشخصاً الگوریتم کارت استفاده کرده‌ایم. در این بخش و قبل از استفاده از الگوریتم کارت، به دلیل حساسیت بالای این نوع الگوریتم‌ها به داده‌های پرت، ابتدا با استفاده از روش‌های آماری داده‌های پرت شناسایی‌شده (که در مورد تحقیق حاضر تعداد ۳۶ خانوار بوده) و موقتاً کنار گذارده می‌شود و پس از اجرای الگوریتم پیش‌بینی کننده، دوباره به بانک داده‌ها اضافه می‌گردد. به‌این‌ترتیب مقادیر مربوط به درآمد خانوارهایی که درآمد خود را اعلام نکرده‌اند با توجه به سایر

اطلاعات خانوار در بانک اطلاعاتی، پیش‌بینی و در فیلد مربوطه قرار داده‌شده است.پس‌ازاین مرحله به ۱۰۰ درصد فیلدهای اطلاعاتی پرشده دست‌یافته‌ایم.

پس از تکمیل بانک اطلاعاتی نوبت به برخورد با داده‌های غیر نرمال می‌رسد. در پایگاه داده‌ی مورد مطالعه خانوارهایی دیده می‌شود که نسبت به سایرین غیرمتعارف یا غیر نرمال می‌باشند. در اینجا منظور ما داده‌های پرت و بافاصله زیاد نمی‌باشد. برای مثال فرض کنید یک خانوار اقساط ماهیانه‌ای برابر با بیست میلیون ریال پرداخت می‌کند درحالی‌که جمع درآمد خانوار چهار میلیون ریال می‌باشد. وجود چنین رکوردهایی در پایگاه داده موجب افزایش ضریب خطا در نتایج داده‌کاوی می‌گردد؛ لذا تشخیص و مدیریت این داده‌های غیرمتعارف در مرحله‌ی آماده‌سازی داده‌ها از اهمیت بسزایی برخوردار می‌باشد.

در مدل تحقیق حاضر پس از بررسی مطالعات انجام‌شده و آزمون روش‌های مختلف، استفاده از تکنیک و الگوریتم آنومالی که از تکنیک‌های خوشه‌بندی در داده‌کاوی به شمار می‌رود به‌عنوان یک روش کارا در تشخیص داده‌های غیرمتعارف استفاده شد. این تکنیک ضمن خوشه‌بندی داده‌ها، داده‌های غیرمتعارف را نیز در هر خوشه مشخص می‌کند. پس از اعمال الگوریتم آنومالی، رکوردهای با شاخص غیرمتعارفی بیش از ۳ به‌عنوان داده‌های غیرمتعارف در نظر گرفته می‌شود. اگر تعداد داده‌های غیرمتعارف زیاد باشد امکان حذف کل آن‌ها وجود نخواهد داشت که در تحقیق حاضر به دلیل تعداد کم این داده‌ها(۲۰۰۰ رکورد) و پس از مشاوره با خبرگان مرکز آمار ایران داده-های غیرمتعارف قبل از انجام خوشه‌بندی نهایی کنار گذاشته شده است. پس از مشخص شدن وضعیت داده‌های گمشده و داده‌های غیر نرمال پایگاه داده حاصل آماده ورود به مرحله‌ی بعد که همان فرآیند نرمال‌سازی است می‌باشد.

در این مرحله و برای بی مقیاس سازی داده‌ها روش‌های مختلف بی مقیاس سازی مورد آزمایش قرار گرفت و پس از مشاوره با خبرگان مرکز آمار روش استفاده از Z نرمال مناسب‌ترین روش برای این منظور تشخیص داده شد. در این روش از فرمول زیر برای بی مقیاس سازی فیلدهای مختلف استفاده می‌شود:

$$Z = \frac{X - @GLOBAL - MEAN(X)}{@GLOBAL\_SDEV(X)} \qquad (۷)$$

در این فرمول X مقدار مربوط به هر فیلد است. @GLOBAL_MEAN میانگین مربوط به یک متغیر (ستون) خاص از جدول و @GLOBAL_SDEV مقدار انحراف از معیار یک متغیر خاص را برمی‌گرداند. لازم به توضیح است توابع گلوبال در نرم‌افزار کلمنتاین باید از قبل تعریف شود تا بتوان از آن‌ها در مدل استفاده کرد که در مدل پروژه‌ی حاضر این کار صورت گرفته است.

۵-۳ مرحله‌ی مدل‌سازی و تجزیه‌وتحلیل داده‌ها و نتیجه‌گیری

در این مرحله نتایج نظر خبرگان در قالب پرسشنامه‌ی مقایسات زوجی و روش آقای ساعتی دریافت و پس از انجام محاسبات لازم نتایج به دست آماده به‌صورت وزن متغیرها بر روی داده‌ها اعمال‌شده است. به این صورت خوشه‌بندی با توجه به نظر خبرگان و اعمال اولویت‌های آن‌ها صورت می‌پذیرد. جدول مربوط به ماتریس نتایج به‌دست‌آمده از پرسش‌نامه‌ها و جدول وزن نهایی به‌دست‌آمده برای هر متغییر در ادامه آورده شده است.

جدول۲. مقایسه زوجی معیارها

	تحصیلات	قیمت مسکن	قیمت خودرو	اقساط ماهیانه	تعداد اعضا	درآمد خانوار
تحصیلات	۱	۰/۷۱۴۲۸۵۷۱۴	۰/۵۵۵۵۵۵۵۵۶	۱/۶۶۶۶۶۶۶۶۷	۲/۵	۰/۸۳۳۳۳۳۳۳۳
قیمت مسکن	۱/۴	۱	۰/۷۷۷۷۷۷۷۷۸	۲/۳۳۳۳۳۳۳۳۳	۳/۵	۱/۱۶۶۶۶۶۶۶۷
قیمت خودرو	۱/۸	۰/۲۲۲۲۲۲۲۲۲	۱	۳	۴/۵	۱/۵
اقساط ماهیانه	۰/۶	۰/۴۲۸۵۷۱۴۲۹	۰/۳۳۳۳۳۳۳۳۳	۱	۱/۵	۰/۵
تعداد اعضا	۰/۴	۰/۲۸۵۷۱۴۲۸۶	۰/۲۲۲۲۲۲۲۲۲	۰/۶۶۶۶۶۶۶۶۷	۱	۰/۳۳۳۳۳۳۳۳۳
درآمد خانوار	۱/۲	۰/۸۵۷۱۴۲۸۵۷	۰/۶۶۶۶۶۶۶۶۷	۲	۳	۱

جدول۳. وزن معیارها

ردیف	متغیر	وزن محاسبه‌شده با روش تحلیل سلسله مراتبی
۱	تحصیلات	۰/۱۵۹۹۰۵۰۳۵
۲	قیمت مسکن	۰/۲۲۳۸۶۷۰۴۸
۳	قیمت خودرو	۰/۲۶۴۴۳۶۸۴۱
۴	اقساط ماهیانه	۰/۰۹۵۹۴۳۰۲۱
۵	تعداد اعضا	۰/۰۶۳۹۶۲۰۱۴
۶	درآمد خانوار	۰/۱۹۱۸۸۶۰۴۱

۵-٤ خوشه‌بندی

۵-٤-۱ تعیین تعداد خوشه‌ی مناسب

در این مرحله داده‌های وزن‌دار شده وارد فرآیند خوشه‌بندی می‌شود. برای اطمینان از تعداد خوشه‌ی مناسب در خوشه‌بندی یک فرایند سه مرحله‌ای در مدل صورت گرفته است:

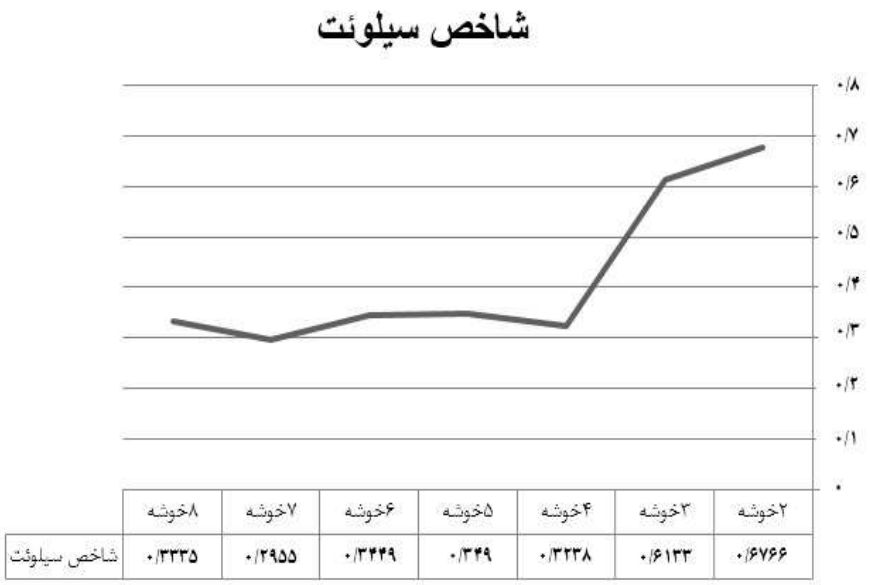
۵-٤-۱-۱ استفاده از الگوریتم دومرحله‌ای

در ابتدا با استفاده از الگوریتم خوشه‌بندی دومرحله‌ای خوشه‌بندی صورت گرفته است. این الگوریتم در طی انجام خوشه‌بندی، تعداد خوشه‌ی مناسب را تخمین می‌زند و خوشه‌بندی را انجام می‌دهد. در

این پژوهش این الگوریتم صرفاً جهت داشتن تخمینی از تعداد خوشه‌ی مناسب استفاده‌شده است که این تعداد برابر با دو خوشه‌ی به‌دست‌آمده؛ اما ازآنجاکه الگوریتم دومرحله‌ای روشی مناسب برای تشخیص تعداد مناسب خوشه با ضریب اطمینان بالا نیست؛ لذا تعداد خوشه‌های پیشنهادشده توسط این الگوریتم تنها به‌عنوان حدودی برای تعداد خوشه‌ی مناسب در نظر گرفته‌شده است.

۵-٤-۱-۲ شاخص سیلوئت

برای اطمینان از تعداد خوشه مناسب فرآیند خوشه‌بندی چندین بار و هر بار با تعداد خوشه‌ی مختلف اجرا شده و در نهایت خروجی مدل در هر بار به‌عنوان ورودی به نرم‌افزار متلب برده شده و در آنجا از شاخص سیلوئت جهت مقایسه‌ی تعداد خوشه‌ها و تشخیص تعداد خوشه‌ی مناسب استفاده‌شده است. نتایج این مقایسه در شکل ۴ آورده شده است.



	۸خوشه	۷خوشه	۶خوشه	۵خوشه	۴خوشه	۳خوشه	۲خوشه
شاخص سیلوئت	۰/۳۳۳۵	۰/۲۹۵۵	۰/۳۴۴۹	۰/۳۴۹	۰/۳۲۳۸	۰/۶۱۳۳	۰/۶۷۶۶

شکل ٤. مقادیر شاخص سیلوئت با تعداد خوشه‌های مختلف

۵-٤-۱-۳ انحراف استاندارد

خوشه‌بندی با تعداد خوشه‌های ۵،۴،۳،۲ و ۶ خوشه انجام‌شده و هر بار برای هر معیار و هر خوشه و نهایتاً هر خوشه‌بندی یک انحراف استاندارد به‌دست‌آمده است. نتایج محاسبات انحراف استاندارد خوشه‌بندی‌های مختلف در جدول ۴ آورده شده است.

با توجه به انحراف استاندارد به‌دست‌آمده در هربار خوشه‌بندی مشاهده می‌شود تعداد خوشه‌ی سه مناسب‌ترین تعداد خوشه برای این مورد مطالعه می‌باشد. درنتیجه با توجه به تخمین به‌دست‌آمده از الگوریتم دومرحله‌ای و مقایسه‌ی شاخص سیلوئت برای خوشه‌بندی با تعداد خوشه‌های مختلف و همچنین مقایسه‌ی

انحراف استاندارد در حالات مختلف خوشه‌بندی، تعداد خوشه‌ی سه مناسب‌ترین تعداد خوشه برای این مدل به‌دست‌آمده و خوشه‌بندی نهایی با در نظر گرفتن سه خوشه مبنای این تحقیق قرارگرفته است.

جدول ٤. انحراف استاندارد خوشه‌بندی‌های مختلف

تعداد خوشه در نظر گرفته‌شده	انحراف استاندارد در هر خوشه						انحراف استاندارد کل
	۱	۲	۳	۴	۵	۶	
۲	۰/۱۶۷۸۳	۰/۱۴۲۵					۰/۱۵۵۱۶۶۶۶۷
۳	۰/۱۴۲۵	۰/۱۴۵۳	۰/۱۵۴۸۳				۰/۱۴۷۵۵۵۵۵۶
۴	۰/۱۲۲	۰/۱۲۳۵	۰/۱۲۲۱۶	۰/۹۲۲۸۳			۰/۳۲۲۶۲۵
۵	۰/۱۰۵	۰/۱۱۸۸۳	۰/۱۴۳۱۶	۰/۱۵۴۸۳	۰/۸۵۲۸۳		۰/۲۷۴۹۳۳۳۳۳
۶	۰/۱۰۴	۰/۱۰۴۶	۰/۵۹۹۱۶	۰/۸۵۲۸۳	۰/۱۶۳۶	۰/۱	۰/۳۲۰۷۲۲۲۲۲

۵-٤-۲ خوشه‌بندی نهایی

در پایگاه داده یک ستون جدید اضافه می‌شود که مشخص‌کننده‌ی خوشه‌ی مربوط به هرخانوارمی‌باشد. بخش کوچکی از پایگاه داده در جدول ۵ آورده شده است:

جدول۵. پایگاه داده بعد از خوشه‌بندی

	W-C...	W-H...	W-VE...	W-E...	W-REV...	W-LOAN...	$KM-K-Means
1	-0.093	0.167	-0.073	-0.234	-0.005	-0.024	cluster-1
2	0.147	-0.114	-0.073	0.060	-0.008	0.025	cluster-2
3	-0.053	-0.114	-0.073	-0.136	-0.010	-0.023	cluster-1
4	0.027	-0.069	-0.073	-0.057	-0.013	-0.036	cluster-1
5	-0.013	-0.114	-0.073	-0.087	-0.010	-0.021	cluster-1
6	-0.013	0.027	-0.073	-0.087	-0.005	-0.023	cluster-1
7	0.027	-0.058	-0.073	0.060	-0.002	-0.030	cluster-1
8	0.027	0.167	0.416	-0.175	0.009	0.026	cluster-1
9	0.107	0.055	-0.073	0.018	-0.010	-0.015	cluster-2
10	0.027	-0.102	-0.073	-0.057	-0.014	-0.036	cluster-1

نتایج به‌دست‌آمده بر اساس مقادیر هم مقیاس شده و وزن‌دار شده می‌باشد. برای تحلیل بهتر، میانگین نتایج به‌دست‌آمده برای هر خوشه از حالت نرمال و وزن‌دار خود خارج‌شده و در جدول ۶ نمایش داده شده است.

جدول٦. میانگین معیارها در خوشه‌های به‌دست‌آمده

ردیف	متغیر	خوشه ۱	خوشه ۲	خوشه ۳
۱	تحصیلات	۱/۵۲۲۸۹۷	۱/۸۹۰۶۷۲	۲/۴۵۹۳۶۳
۲	قیمت مسکن	۱۷۵۹۴۸۳۲۱/۹	۲۳۴۷۷۶۰۹۸/۹	۲۳۸۳۴۱۴۱۸/۸
۳	قیمت خودرو	۱۲۰۲۲۳۶۴	۱۶۱۱۶۹۸۲	۲۴۵۱۰۹۴۹
۴	اقساط ماهیانه	۱۹۹۸۷۴۵	۲۲۹۱۱۵۴	۲۹۲۴۷۰۷
۵	تعداد اعضا	۳/۹۰۰۵۴۳	۶/۵۲۹۱۵۹	۲/۹۲۴۲
۶	درآمد خانوار	۲۹۱۱۲۳۱	۳۵۷۴۹۴۳	۴۶۳۶۸۸۲

۵-٤-۳ نام‌گذاری و تفسیر خوشه‌ها

با توجه به نتایج به‌دست‌آمده خوشه‌ها به‌صورت زیر نام‌گذاری می‌شود:

I. خانوارهای با سطح درآمد و تحصیلات پایین

II. خانوارهای پرجمعیت با سطح متوسط

III. خانوارهای کم‌جمعیت با سطح تحصیلات و درآمد بالا

۵-۵ رتبه‌بندی خوشه‌ها با استفاده از روش راه‌حل توافقی و بهینه‌سازی چند معیاره (ویکور)

در این مرحله خوشه‌های به‌دست‌آمده از داده‌کاوی با توجه به میانگین مقادیر به‌دست‌آمده برای هر خوشه و در نظر گرفتن وزن‌های محاسبه‌شده برای متغیرها با استفاده از روش "راه‌حل توافقی و بهینه‌سازی چند معیاره" رتبه‌بندی شده است. متغیرها به‌عنوان معیارهای تصمیم‌گیری چند معیاره می‌باشد. با توجه به نتایج به‌دست‌آمده در جدول۶ و پس از نرمال‌سازی ماتریس فوق، نتایج به‌دست‌آمده از روش راه‌حل توافقی و بهینه‌سازی چند معیاره مطابق جدول ۷ می‌باشد. همان‌طور که پیش‌تر بیان شد در این روش رتبه‌بندی نهایی بر اساس مقایسه‌ی مقادیر فاصله از راه‌حل ایده آل مثبت، فاصله از راه‌حل ایده آل منفی و مقدار شاخص ویکور به‌دست‌آمده است.

جدول۷. رتبه‌بندی با روش ویکور

معیارها		تعداد اعضا	تحصیلات	قیمت مسکن	اقساط ماهیانه	درآمد خانوار	قیمت خودرو
وزن معیارها		۰/۰۶۳۹۶۲	۰/۱۵۹۹۰۵	۰/۲۲۳۸۶۷	۰/۰۹۵۹۴۳	۰/۱۹۱۸۸۶	۰/۲۶۴۴۳۶۸
گزینه‌ها	خوشه ۱	۰/۴۷۸۶۹۳	۰/۴۴۰۶۸۳	۰/۴۶۵۴۷۱	۰/۴۷۳۷۷۱	۰/۴۴۵۲۲۳	۰/۳۷۹۲۱۸
	خوشه ۲	۰/۸۰۱۲۸۹	۰/۵۴۷۱۰۷	۰/۶۲۱۱	۰/۵۴۳۰۸۲	۰/۵۴۶۷۲۶	۰/۵۰۸۳۷۴
	خوشه ۳	۰/۳۵۸۸۷۲	۰/۷۱۱۶۶۹	۰/۶۳۰۵۳۲	۰/۶۹۳۲۵۶	۰/۷۰۹۱۳۱	۰/۷۷۳۱۴۳
f^*		۰/۳۵۸۸۷۲	۰/۷۱۱۶۶۹	۰/۶۳۰۵۳۲	۰/۴۷۳۷۷۱	۰/۷۰۹۱۳۱	۰/۷۷۳۱۴۳
f^-		۰/۸۰۱۲۸۹	۰/۴۴۰۶۸۳	۰/۴۶۵۴۷۱	۰/۶۹۳۲۵۶	۰/۴۴۵۲۲۳	۰/۳۷۹۲۱۸
گزینه‌ها	فاصله از راه‌حل ایده آل مثبت (Si)		فاصله از راه‌حل ایده آل منفی (Ri)		مقدار (Vi)	مقدار (Qi)	رتبه‌بندی
خوشه ۱	۰/۸۵۷۴۱۸		۰/۲۶۴۴۳۷		۰/۵	۱	۳
خوشه ۲	۰/۴۹۹۹۷۸		۰/۱۷۷۷۳۶		۰/۵	۰/۵۰۸۰۱۶	۲
خوشه ۳	۰/۰۹۵۹۴۳		۰/۰۹۵۹۴۳		۰/۵	۰	۱

این رتبه‌بندی بر مبنای سطح رفاه خانوارها صورت گرفته است و می‌تواند در تصمیم‌گیری‌های آتی مرکز آمار ایران و تدوین استراتژی‌های این مرکز در تخصیص منابع کاربرد داشته باشد. با این رتبه‌بندی درواقع برای هر خانوار با توجه به خوشه‌ی مربوط یک رتبه در نظر گرفته‌شده است.

۵-۶ نام‌گذاری و توصیف خوشه‌ها

با توجه به نتایج به‌دست‌آمده، خوشه‌ها به‌صورت زیر نام‌گذاری و توصیف می‌شود:

۵-۶-۱ خانوارهای با سطح درآمد و تحصیلات پایین

این خوشه با ۵۶٪ خانوارهای استان بیش‌ترین درصد خانوارهای استان را به خود اختصاص داده است. خانوارهای این خوشه پایین‌ترین سطح درآمد را در بین کل خانوارهای استان مورد مطالعه داشته و هم‌زمان با سطح درآمد پایین سطح تحصیلات نیز در این خانوارها بسیار پایین می‌باشد. به‌طوری‌که اکثر اعضای این خانوارها را افراد کم‌سواد تشکیل می‌دهد.

به نظر می‌رسد استفاده از معیار شهرنشینی و روستانشینی خانوارها در تحلیل این خوشه موردنیاز باشد چراکه خانوارهایی با سطح درآمد و تحصیلات پایین و جمعیت متوسط در روستا دارای رفاه نسبی بوده درحالی‌که خانواری با این شرایط در محیط شهر از اقشار ضعیف جامعه محسوب می‌شود.

۵-۶-۲ خانوارهای پرجمعیت با سطح متوسط

این خوشه ۲۴٪ جمعیت کل استان را شامل می‌شود. مشخصه‌ی بارز خانوارهای این خوشه پرجمعیت بودنشان است. به‌طوری‌که میانگین جمعیت خانوارها در این خوشه ۶/۵ نفر می‌باشد. خانوارهای این خوشه در سایر معیارها وضعیتی متوسط دارند؛ اما تعداد زیاد افراد خانوارهای این خوشه وضعیت رفاهی افراد این خانوارها را به‌شدت تحت تأثیر قرار می‌دهد. جمعیت زیاد این خانوارها در حالی است که دو خوشه‌ی دیگر دارای میانگین جمعیت ۳ و ۴ نفر در هر خانوار می‌باشد.

۵-۶-۳ خانوارهای کم‌جمعیت با سطح تحصیلات و درآمد بالا

خانوارهای این خوشه با بالاترین سطح تحصیلات در استان دارای کم‌ترین میانگین جمعیت خانوار(۳) و همچنین بالاترین سطح درآمد می‌باشند. با توجه به نتایج رتبه‌بندی ازنظر وضعیت رفاهی خانوارها و بر اساس معیارها و وزن‌های ذکرشده به دست آمدن بالاترین رتبه برای این خوشه‌ی طبیعی و مطابق با نتایج خوشه‌بندی می‌باشد.

همچنین این خوشه به دلیل سطح بالای تحصیلات خانوارها دارای مطمئن‌ترین داده‌ها در تحقیق می‌باشد و در مدل‌های پیش‌بینی می‌توان از این داده‌ها به‌عنوان داده‌های آزمایشی استفاده نمود.

٦ نتیجه‌گیری

٦-١ امکان کاربردی شدن هر چه بیش‌تر مدل‌های داده‌کاوی توسط مدیران

مدل ارایه‌شده در پژوهش حاضر با واردکردن نظرات مدیران در فرآیند داده‌کاوی تأثیر تصمیمات مدیران بر نتایج داده‌کاوی را پررنگ‌تر می‌سازد و این امر می‌تواند موجب رغبت هر چه بیش‌تر مدیران در استفاده از تکنیک‌های داده‌کاوی گردد. این مدل و ایده استفاده از وزن و رتبه‌های به‌دست‌آمده با روش‌های تصمیم‌گیری چند معیاره در داده‌کاوی قابل‌اجرا در سایر روش‌های داده‌کاوی ازجمله مدل‌های درخت تصمیم و تکنیک‌های پیش‌بینی می‌باشد.

٦-٢ رتبه‌بندی خوشه‌ها

رتبه‌بندی خوشه‌ها خود یک رویکرد نوین در این زمینه است که می‌تواند راهنمای مدیران در تصمیم‌گیری‌ها باشد. این رتبه‌بندی می‌تواند از دیدگاه‌های مختلف و با روش‌های مختلف صورت بگیرد. برای مثال در تخصیص منابع می‌توان از روش‌های وزن دهی استفاده کرده، تخصیص منابع را بر اساس وزن‌های به‌دست‌آمده انجام داد.

٦-٣ مزیت کاربرد این مدل در مرکز آمار ایران

استفاده از این مدل به‌عنوان یک مدل ترکیبی موجب افزایش دقت تعیین خوشه ها خواهد شد. همچنین رتبه بندی انجام شده که ورودی خود را از خروجی خوشه بندی انجام شده گرفته است باعث مشخص شدن سطح رفاه خانوارها در هر خوشه می شود و این امر از بروز خطاهای احتمالی در تعیین سطح رفاه خانوارها تا حد زیادی جلوگیری خواهد کرد؛ البته دسترسی به اطلاعات خوشه بندی‌های قبلی امکان‌پذیر نبوده، درصورت استفاده از این مدل در خوشه‌بندی‌های آتی امکان مقایسه وجود خواهد داشت.

منابع

[۲] کریمی شیرازی، ح.، مدیری، م. وفرجپورخاناپشتانی، ق.، (۱۳۹۳). یک مدل MCDM جدید ترکیبی از DEMATELوANPباVIKOR برای اولویت‌بندی کاربردهای فناوری نانو در بخش صنایع غذایی. مجله تحقیق در عملیات وکاربردهای آن، ۹(۲)، ۳۳-۱۳.

[۶] غضنفری، م.، علیزاده، س.و تیمورپور، ب.، (۱۳۸۷). داده‌کاوی و کشف دانش (نسخه چاپ اول). تهران، دانشگاه علم و صنعت ایران.

[۷] آذر، ع.، سبط، و.،(۱۳۸۹). طراحی مدل انتخاب نیروی انسانی با رویکرد داده کاوی. نشریه مدیریت فناوری اطلاعات، ۲(۴)، ۲۲-۳.

[۸] اصغر پور، م .ج .(۱۳۸۷) .تصمیم‌گیری‌های چند معیاره .موسسه انتشارات و چاپ دانشگاه تهران.

[۱۴] بابایی، م.، صفار یزدی، ز. و سرایی، م. ح.، (۱۳۹۰). معرفی و مقایسه روش‌های پیش‌پردازش داده برای کاربردهای مختلف داده‌کاوی. تکنولوژی و علوم دانشگاه اصفهان .

[1] Mosavi, A.,(2010). Multiple Criteria Decision Making Preprocessing Using Data Mining Tools. International Journal of Computer Science Issues, 7(2): 26-34.
[3] Toloo, M., Sohrabi, B., Nalchigar, S., (2009). A new method for ranking discovered rules from data mining by DEA. Expert Systems with Applications. 36: 8503-8508.
[4] Coenen, F., (2011). Data mining: past, present and future. The Knowledge Engineering Review. 26(1): 25-29.
[5] Fayyad, U., Piatesky-Shapiro, G., Smyth, P., (1996). The KDD Process for Extracting Useful Knowledge from Volumes of Data. Communications of The ACM. 37: 27-34.
[9] Jingzhu, W., Xiangy, L., (2009). The Multiple Attributed Decision-Making VIKOR Method and Its Application. IEEE.
[10] Haleh, H.,Ghaffari, A., Meshki, A. K ., (2012). A Combined Model of MCDM and Data Mining forDetermining Question Weights in Scientific Exams .Applied Mathematical Sciences. 6: 173-196.
[11] Mastrogiannis, N., Boutsinasa, B., Giannikos, I.,(2009). Amethod for improving the accuracy of datamining classification algorithms. Computers&Operations Research. 36: 2829-2839.
[12] Choi, D. H., Ahn, B. S., Kim, S. H.,(2005). Prioritization of association rules in data mining: Multiple criteria decision approach. Expert Systems with Applications. 29: 867-878.
[13] Muata, K., Bryson, O., (2004). Evaluation of decision trees: a multi-criteria approach. Computers&Operations Research. 34: 1933-1945.
[15] Foroughi, A., (2011). A note on "A new method for ranking discovered rules from data mining by DEA," and a full ranking approach. Expert Systems with Applications. 38: 12913-12916.
[16] Toloo, M., Nalchigar, S., (2011). A new DEA method for supplier selection in presence of both cardinal and ordinal data . Expert Systems with Applications. 38: 14726-14731.
[17] Rad, A., Naderi, B., Soltani, M., (2011). Clustering and ranking university majors using data mining and AHP algorithms:A case study in Iran. Expert Systems with Applications. 38: 755-763.
[18] Liu, D., Shih, Y., (2005). Integrating AHP and data mining for product recommendation based on customer lifetime value. Information&Management. 42: 387-400.