ارتقای روابط صنعت و دانشگاه با نگرشی بر شاخص های توسعه یافتگی در سهمیه بندی کنکور
(کاربرد تکنیک داده کاوی)

* نرجس سرعتی آشتیانی ** سمیّه علیزاده *** علی مبصّری
* موسسه مطالعات بین المللی انرژی، دانشجوی دکترای مدیریت منابع انسانی دانشگاه تهران
** عضو هیئت علمی دانشکده صنایع، دانشگاه خواجه نصیرالدّین طوسی، تهران
*** دانش آموخته کارشناسی ارشد مهندسی صنایع، دانشگاه علم وصنعت ایران، تهران
تاریخ دریافت: ۱۳۹۳/۴/۳ تاریخ پذیرش: ۱۳۹۴/۱۰/۲

چکیده

تعداد زیادی از فارغ‌التحصیلان دبیرستان‌ها در سیستم آموزشی ایران خواهان ورود به دانشگاه‌ها می‌باشند و رقابت اصلی برای ورود به مراکز دانشگاهی معتبر می‌باشد. از سویی دیگر تسهیلات آموزشی، بهداشتی و ... در تمامی شهرها توزیعی مناسب ندارند. مدیران سازمان‌های ذیربط، تخصیص سهمیه را راه‌کاری مناسب برای حل این مسأله می‌دانند و به دنبال استفاده از دانش نهفته در داده-های موجود در این حوزه هستند. با منطقه‌بندی کلیه بخش‌های کشور، داوطلبان هر منطقه با هم مقایسه می‌شوند و در واقع با این روش از اینکه درصد پذیرفته‌شدگان یک شهر چند برابر شهر دیگری باشد، جلوگیری می‌شود. تعیین میزان سهمیه کنکور برای بخش‌های کشور در سال های اخیر، برمبنای میزان توسعه‌یافتگی مناطق با استفاده از روش تاکسونومی صورت گرفته است که خروجی حاصل از این روش نوعی رتبه‌بندی مناطق می‌باشد که در آن امکان تحلیل گروهی مناطق وجود ندارد، همچنین تعداد مناطق بصورت نظری تعیین می شود. برای رفع این مسائل بخش‌بندی می‌تواند به عنوان یک راهکار مناسب مورد استفاده قرار گیرد. تحقیق حاضر برای اولین بار در حوزه توسعه‌یافتگی، با استفاده از تکنیک‌های داده‌کاوی و روش کریسپ و در قالب متدولوژی پیشنهادی، بر روی داده‌های مرتبط، در وزارت آموزش و پرورش، وزارت کشور، وزارت بهداشت و درمان، مرکز آمار و سازمان سنجش، صورت گرفته است.

پس از شناسایی استانداردها و شاخص‌های اثرگذار در این زمینه، آماده‌سازی داده‌ها انجام شده و به ساخت انباره‌داده و ترکیب شاخص‌ها جهت استخراج عوامل جدید پرداخته شده است. در گام بعدی با بکارگیری الگوریتم K-means بخش‌های شبیه به هم در خوشه‌های مربوطه قرار گرفته و سپس با استفاده از روش پیش‌بینی شبکه های عصبی و درخت تصمیم امکان اختصاص بخش-های جدید به هر کلاس (خوشه‌های ایجاد شده) فراهم شده و جهت ارزیابی مدل‌های ایجاد شده، دقت خروجی با سایر روش‌ها مقایسه شده است. دستاوردهای این تحقیق عبارتند از: تعیین تعداد بهینه بخش‌ها، بخش‌بندی مناطق، تحلیل هر بخش، استخراج قواعد تصمیم‌گیری، امکان پیش‌بینی سریع‌تر و دقیق‌تر برچسب کلاس برای مناطق جدید، فراهم نمودن امکان تدوین راهبردهای مناسب برای هر بخش.

واژه‌های کلیدی: فرهنگ سازمانی، مدیریت منابع انسانی استراتژیک، رفتار شهروندی سازمانی، مدل معادلات ساختاری

مقدمه

امروزه شناخت و درک صحیح توانایی‌ها، نیازها و خواسته‌های خدمت‌گیرندگان یا مشتریان هر سازمان از موارد قابل تأمل می‌باشد و سازمان‌ها می‌باید به این موارد آگاه بوده و حتی آنها را پیش‌بینی کنند و با تجهیز شدن به این اطلاعات و دانش سلامت کاری خود را بهبود داده، امکان اتخاذ تصمیم درست را فراهم نمایند. باتوجه به

نویسندهٔ عهده‌دار مکاتبات: نرجس سرعتی آشتیانیnara.sorati@yahoo.com

اهمّیت کلیدی موضوع آموزش، بسترهای مورد نیاز آن و نحوه‌ی توزیع تسهیلات و سهمیه‌های مرتبط با آن در هر کشوری، بررسی و تحقیق در این حوزه بسیار حیاتی می باشد.

فن‌آوری داده‌کاوی امروزه به موضوعی داغ برای تصمیم-گیران تبدیل شده است، زیرا داده‌کاوی کسب و کارهای مخفی و باارزش را از مقادیر بزرگی از داده‌های تاریخی ارائه می‌دهد. اساساً داده‌کاوی فن‌آوری جدیدی نیست، موضوع استخراج اطلاعات و دانش از رکوردهای داده مفهومی کاملاً جاافتاده است، آنچه‌که جدید محسوب می‌گردد، تجمّع و پکپارچگی چندین رشته و فن‌آوری مرتبط می‌باشد که فرصتی یکتا برای کاوش داده‌ها در فضایی علمی و واحد را خلق نموده است (.کانتراتزیک ۲۰۰۳)

این تحقیق سعی داشته است تا با استفاده از مفاهیمی همچون بخش‌بندی مناطق (بخش‌های تقسیمات کشوری، به‌عنوان خدمت‌گیرندگان سازمان سنجش) با استفاده از داده‌کاوی و نیز استانداردهای موجود در راستای توسعه-یافتگی، ویژگی‌های مؤثر در این زمینه را انتخاب کند و پس از آن تعداد بهینه مناطق شبیه را تعیین نموده و مناطق شبیه را در خوشه‌های مربوطه قرار دهد، تا شناخت بهتری از مناطق ایجاد گردد. همچنین با استفاده از الگوریتمهای پیش‌بینی امکان تحلیل بهتر بخشها و تعیین کلاس بخشهای جدید را فراهم نموده است.

مبانی نظری پژوهش

الف) شاخص‌های توسعه‌یافتگی

توسعه به مفهوم ارتقاء مستمر کل جامعه و نظام اجتماعی به سوی زندگی بهتر و یا انسانی تر با استفاده بهینه از منابع موجود است (تودارو، ۱۳۷۰) توسعه به هدف و وسیله تغییرات اشاره داشته و به طور همزمان دور نمای نوعی زندگی بهتر که از نظر مادی مرفه‌تر، جدید تر، دارای غنای معنوی بیشتر و از نظر تکنولوژیکی "کاراتر" است را بصورت مجموعه ای از وسایل لازم برای رسیدن به این دورنما، ترسیم می‌کند. به طور کلی باید اذعان داشت توسعه فرایندی پیچیده و چند بعدی است که مستلزم تغییر در ساخت اجتماعی، طرز تلقی مردم و نهادهای ملی و نیز تسریع رشد اقتصادی، کاهش نابرابری و ریشه کن کردن فقر و برقراری عدالت اجتماعی و پایداری محیط می باشد. (تودارو، ۱۳۷۸)

توسعه و توسعه نیافتگی مناطق از جمله مباحث توسعه بوده که در بین اقتصادانان و برنامه ریزان مطرح است. در همین راستا وجود نابرابری‌ها و تفاوت های منطقه‌ای که علاوه بر ویژگی‌های طبیعی، اقتصادی، اجتماعی، متاثر از سیاست‌ها و برنامه‌ریزی‌های گذشته، حال و آینده است، برنامه ریزان را بر آن داشته که تکنیک‌ها و روش‌هایی را ابداع کنند تا از طریق تعیین درجه توسعه یافتگی و رتبه بندی مناطق بتوانند به شناخت و تحلیل علل یا عوامل نابرابریها و تفاوت‌های منطقه‌ای دست یابند. تعیین شاخصهای توسعه خاصه شاخصهای مرتبط با توسعه همه جانبه مهمترین قدم در مطالعات توسعه منطقه‌ای است. شاخصهای توسعه در واقع بیان آماری پدیده‌های موجود در منطقه بوده و برای بیان اهمیت شاخصهای توسعه و نقش آن در بیان آماری پدیده‌ها، ضروری است تا مفاهیم مربوط به متغیر و شاخص بطور عمیق‌تر بررسی قرار گرفته و تفاوت بین آنها مشخص شود.

الف-۱) شاخص‌های توسعه انسانی

شاخص توسعه انسانی یک سنجه خلاصه برای توسعه انسانی است. این شاخص متوسط دستاوردهای یک کشور را در سه بعد از توسعه انسانی محاسبه می‌کند:

- زندگی طولانی و سالم، که بر اساس امید به زنگی در بدو تولد محاسبه می‌شود.
- دانش، که بر اساس نرخ باسوادی بزرگسالان (با ضریب دوسوم) و نسبت ترکیبی ثبت‌نام در مدارس ابتدایی، متوسطه و عالی (با ضریب یک سوم) محاسبه می‌شود.
- استاندارد شایسته زندگی، که بر اساس سرانه تولید ناخالص داخلی (برحسب برابری قدرت خرید دلار آمریکا) محاسبه می‌گردد.

الف-۲) شاخص فقر انسانی[1] برای کشورهای در حال توسعه

در حالیکه، شاخص توسعه انسانی، متوسط دستاوردها را محاسبه می‌کند، شاخص فقر انسانی، محرومیت‌ها را در سه

[1]. The Poverty Human Index (HPI)

بعد از ابعاد توسعه انسانی اندازه‌گیری می‌کند و در دوحالت، برای کشورهای در حال توسعه و کشورهای پردرآمد بطور مجزا محاسبه می‌شود.

- زندگی طولانی و سالم: آسیب‌پذیری در برابر مرگ در سنین نسبتاً پایین، که بر اساس احتمال نرسیدن به سن ۴۰ سالگی (در بدو تولد) محاسبه می‌شود.
- دانش: محرومیت از دنیای ارتباطات و خواندنی‌ها که بر اساس نرخ بی‌سوادی بزرگسالان محاسبه می‌شود.
- استاندارد شایسته زندگی: عدم دسترسی به تسهیلات اقتصادی که بر اساس متوسط دو معیار محاسبه می‌شود، درصدی از جمعیت که به منابع آب سالم، دسترسی دائمی ندارند و درصدی از کودکان که نسبت به سنشان کم-وزن هستند.(وانتزکینز۲۰۰۸)

ب) داده‌کاوی

اساساً داده‌کاوی فن‌آوری جدیدی نیست، داده‌کاوی دستیابی به اطلاعات و دانش، و کشف مدلها و الگوهای پنهان از بانکهای اطلاعاتی حجیم و پیچیده می‌باشد. این فن‌آوری شامل حوزه‌های متنوعی از علوم پایگاه داده‌ها، آمار، بصری-سازی، علوم اطلاعاتی، یادگیری ماشین، تشخیص الگو، بازیابی اطلاعات، هوش مصنوعی و بعضی علوم دیگر می-باشد.(هان، ۲۰۰۶)

در داده‌کاوی معمولاً به کشف الگوهای مفید از میان داده‌ها اشاره می‌شود، منظور از الگوی مفید، مدلی در داده‌هاست که ارتباط میان یک زیر مجموعه از داده‌ها را توصیف می-کند و معتبر، ساده، قابل فهم و جدید است.(هند،۱۹۹۸)روش‌های مورد استفاده برای انجام داده-کاوی به صورت زیر هستند:

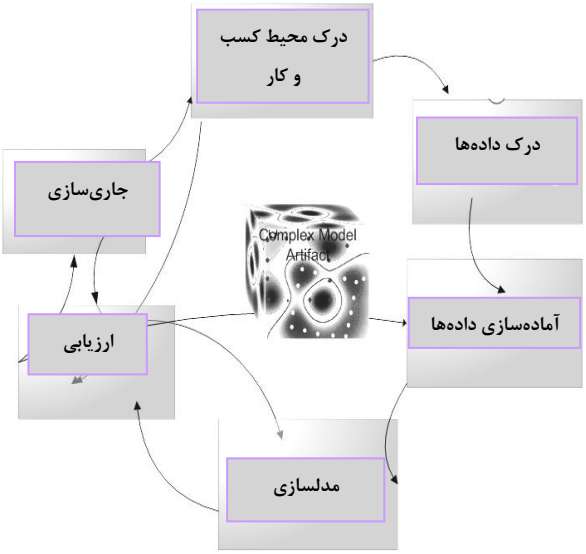
- دسته‌بندی[۲]: این مدل یک قلم داده را به یکی از چند طبقه موجود تخصیص می‌دهد. از جمله تکنیک‌های دسته-بندی، پس‌انتشار[۳]،تکنیک شبکه عصبی[۴]، دسته‌بندهای درخت تصمیم[۵] و دسته‌بندهای بیزین[۶] می‌باشند.
- رگرسیون[۷]:هدف این تابع مدل، نگاشت یک قلم داده به یک متغیر پیش‌بینی، با ارزش حقیقی یا پیوسته است.
- خوشه‌بندی[۸]: این مدل، یک قلم داده را به یکی از چند خوشه موجود، نگاشت می‌کند، جائیکه خوشه‌ها گروه‌های طبیعی از قلم‌های داده بر اساس استانداردهای شباهت یا مدل‌های تراکم احتمالی هستند.
- تولید قوانین[۹]: این روش قوانینی را در داده‌ها کاوش می‌کند. قواعد انجمنی که بدنبال کشف ارتباطات در میان ویژگی‌های مختلف هستند، زیرمجموعه این حوزه می‌باشند.
- خلاصه‌سازی و چگالش[۱۰]: این تابع، توصیفی فشرده برای یک زیرمجموعه از داده‌ها را فراهم می‌کند و نقشی مهم در فشرده‌سازی داده‌ها بویژه داده‌های چندرسانه‌ای، با کاهش تعداد بیتها و افزایش پهنای باند حافظه، ایفا می‌کند.
- تحلیل توالی[۱۱]: این روش الگوهای متوالی، همچون تحلیل سری‌های زمانی و ترتیب ژن‌ها، را مدل می‌کند و هدف آن مدل نمودن مراحل فرایند تولید توالی و یا استخراج و گزارش انحراف‌ها در طول زمان می‌باشد(بری ۲۰۰۴ و چیو ۲۰۰۹)

ب-۱)روش کریسپ برای داده‌کاوی

روش CRISP-DM [۱۲] که توسط کمیته اروپایی ارائه شده، شده، ازجمله متدولوژی‌های مطرح برای انجام پروژه‌های داده‌کاوی است که چارچوب واحدی را پیشنهاد می‌کند تا کیفیت نتایج همراه با کاهش هزینه و زمان تضمین گردد.(هیل دبرانت، ۲۰۰۸) این روش چشم‌اندازی از چرخه زندگی یک پروژه داده‌کاوی را ارائه می‌کند و شامل فازهای یک پروژه، وظایف مربوط به هر فاز و ارتباط میان این وظایف می‌باشد. تعیین دقیق روابط میان وظایف امکان‌پذیر نیست و بطور خاص، روابط وابسته به اهداف، پیش‌زمینه، علایق کاربران و بویژه براساس داده‌ها می‌باشد. راه‌حل داده‌کاوی[۱۳] همیشه به مرحله استقرار نمی‌رسد، بلکه درس-

2. Classification
3. Backpropagation
4. Neural network
5. Decision tree classifiers
6. Bayesian classifiers
7. Regression
8. Clustering
9. Rule generation
10. Summarization or condensation
11. Sequence analysis
3. CRoss-Industry Standard Process for Data Mining
13. Data mining solution

های آموخته شده[14] از هر فاز خود می‌توانند برای سؤالات سازمان مفید واقع شوند.(میلوتینوی ۲۰۰۲) گام‌های این متدولوژی عبارتند از: درک محیط کسب وکار، درک داده‌ها، آماده‌سازی داده‌ها، مدلسازی، ارزیابی و جاری‌سازی(هیل دبرانت ۲۰۰۸)



شکل (۱): روش CRISP-DM

ب-۲) الگوریتم k-means

K-means یکی از معروف‌ترین الگوریتم‌ها برای خوشه-بندی می‌باشد و در اصل بعنوان روش فورجی شناخته می-شود و در بسیاری از حوزه‌های مختلف شامل داده‌کاوی، تحلیل آماری داده‌ها و دیگر کاربردهای کسب و کار استفاده شده است.(MacQueen, 1967) واژه K-means را برای این الگوریتم که هر قلم داده را به خوشه‌ای تخصیص می‌دهد که دارای نزدیکترین فاصله به مرکز ثقل (میانگین) آن خوشه باشد، پیشنهاد نمود.(چنگ ۲۰۰۹)

ب-۳)شبکه‌های عصبی

یک شبکه عصبی از یک شبکه لایه‌ای شده، پیش‌رو و نرون های (نودهای) کاملاً متصل تشکیل شده است. ماهیت پیش‌رو بودن شبکه مانع بوجود آمدن حلقه و اتصالات تکی می‌گردد. این شبکه متشکل از سه لایه ورودی، پنهان و خروجی می‌باشد که لایه مخفی ممکن است بیشتر از یکی باشد. شبکه کاملاً متصل شبکه‌ای است که در آن تمامی نودهای یک لایه به کلیه نودهای لایه بعدی متصلند و هر اتصال دارای وزن مربوط به خود می‌باشد. در ابتدا این وزن‌ها بطور تصادفی و در فاصله ۰ و ۱ به نرون‌ها منصوب می‌شوند. نودهای ورودی معرف ویژگی‌ها در مجموعه داده هستند. تعداد لایه‌های مخفی و تعداد نودهای آنها توسط کاربر قابل پیکربندی بوده و در لایه خروجی ممکن است تعداد نودها بیشتر از یکی باشد.

با استفاده از ورودی‌های هر گره و وزن‌های مربوطه، طبق رابطه زیر، یک ترکیب خطی برای آن گره ایجاد می‌شود که نت گره نام دارد و در آن x_{ij}، iامین ورودی نود jام و w_{ij}، نیز وزن iامین ورودی نود jام است.

$$net_j = \sum_i W_{ij}x_{ij} = W_{0j}x_{0j} + W_{1j}x_{1j} + \cdots + W_{Ij}x_{Ij} \quad (۱)$$

در مرحله بعد مقدار نت محاسبه شده برای هر نود بعنوان x وارد تابع سیگموید[15] می‌شود که معروف‌ترین تابع فعال سازی[16] می‌باشد و فرمول آن بصورت ذیل می‌باشد:

$$y = \frac{1}{1+e^{-x}} \quad (۲)$$

خروجی این تابع به‌عنوان ورودی نرون‌ها درلایه(های) بعدی می‌باشد. برای یادگیری شبکه باید از الگوریتم پس‌انتشار خطا استفاده نمود.(لاروس ۲۰۰۵)

ب-۴)درخت تصمیم

در داده‌کاوی درخت تصمیم، مدلی است که می‌تواند برای هر دو منظور پیش‌بینی و دسته‌بندی مورد استفاده قرار گیرد. از سوی دیگر تصمیم‌گیران از این درخت برای تعیین راهبردی که به هدف موردنظرشان منتهی می‌شود، استفاده می‌کنند. درخت تصمیم دسته‌بندی برای دسته‌بندی نمودن یک شیء یا نمونه (نظیر بیمه) به مجموعه دسته‌های از قبل تعریف شده (نظیر پرریسک و کم ریسک)، بر اساس ویژگی-ها (نظیر سن و جنسیّت) بکار می‌رود. درخت تصمیم درختی است مستقیم با گره‌ای بنام ریشه که این گره[17] دارای یال‌های ورودی نمی‌باشد و تمامی گره‌های دیگر دارای یک یال ورودی هستند، هر گره‌ای که دارای یال‌های خروجی است، گره داخلی یا تست نامیده می‌شود. گره‌های

14. Lessons learned

15. Sigmoid
16. Activation Function
17. Node

باقیمانده در انتهای مسیر برگ‌ها هستندکه به عنوان پایانه یا تصمیم نیز شناخته می‌شوند. هر مسیر از ریشه یک درخت تصمیم تا هر یک از برگ‌های آن با در نظر گرفتن گره‌های تست، بعنوان مسیر میانی به صورت یک قانون ترجمه می‌شود. (روکاخ ۲۰۰۸)برای انتخاب نقطه انشعاب[18] معیارهای متعددی وجود دارد که از جمله معروف‌ترین آنها، شاخص جینی[19] می‌باشد که جزء روش‌های مبتنی بر ناخالصی[20] بوده و برای انشعاب دودویی بکار می‌رود، نحوه محاسبه شاخص جینی به‌صورت زیر است:

$$Gini = 1 - \sum_{i=1}^{m} (\frac{C_i}{n})^2$$

$$Gini_{Split}(S) = \frac{n_1}{n} Gini(s_1) + \frac{n_2}{n} Gini(s_2)$$

که در آن:

n: تعداد رکوردهای موجود در مجوعه S m: تعداد کلاس‌ها

C i: تعداد رکوردهای متعلق به کلاس I

و S به دو زیرمجموعه S1 و S2 با تعداد رکوردهای n1 و n2 تقسیم می‌شود. به ازای تمامی متغیرهای گره‌ها، این شاخص را محاسبه و کمترین مقدار به‌عنوان نقطه انشعاب، انتخاب می‌شود.(لیو،۲۰۰۱)

روش شناسی

پژوهش حاضر به لحاظ هدف در زمره‌ی تحقیقات کاربردی بوده و هدف آن توسعه‌ی دانش کاربردی در زمینه میزان توسعه‌یافتگی می‌باشد. افزون بر این، از لحاظ تقسیم بندی های روش شناسی، روش به کار رفته در پژوهش توصیفی و از نوع مطالعه موردی و از نظر سطح و قلمرو بررسی نیز در محدوده کشور جمهوری اسلامی ایران می‌باشد.

جامعه آماری و نمونه آماری

با توجه به عنوان پژوهش، پیداست که جامعه مورد پژوهش داده های مرتبط با شاخص های توسعه یافتگی است. در پژوهش حاضر با توجه به گستردگی و عدم امکان بررسی تمامی شاخص های جهانی مربوط به توسعه‌یافتگی که سالانه توسط سازمان ملل جهت تعیین میزان توسعه‌یافتگی مورد استفاده قرار می‌گیرد، بر اساس مصاحبه با صاحب-نظران و مطالعه مدارک سازمان‌های داخلی مرتبط،و نیز داده‌های موجود در کشور، شاخص‌های بومی در این زمینه استخراج گردیدند. این داده‌ها مربوط به وزارت آموزش و پرورش، وزارت کشور، وزارت بهداشت و درمان، مرکز آمار و سازمان سنجش و در بازه زمانی سال‌های ۱۳۸۴ تا ۱۳۸۶ می‌باشند.

مدل مفهومی پژوهش

در این بخش فرایند پیشنهادی تشریح گردیده است که گام-های آن مطابق مدل ارائه شده در شکل (۲) می‌باشد:

۱- استخراج شاخص‌های مؤثر به‌روش مصاحبه و مطالعه منابع

با در نظر گرفتن شاخص‌های توسعه انسان سازمان ملل و پس از مطالعه اولیه مستندات سازمان‌هایی که در این حوزه فعالند، همچنین مصاحبه با افراد خبره و داده‌های در دسترس تعداد ۴۰ شاخص طبق جدول (۱) حاصل شد:

۲- آماده‌سازی داده‌ها

نرمال‌سازی داده‌ها با روش Min-Max

در الگوریتم‌هایی که از سنجه‌های تعیین فاصله، مانند فاصله اقلیدسی استفاده می‌کنند، ممکن است داده‌هایی که دارای مقیاس بزرگ هستند نتایج را بسوی خود منحرف کنند، برای جلوگیری از این مسأله و بهبود کارایی و دقّت، داده‌ها را قبل از استفاده نرمال می‌نماییم. در روش Min-Max یک تبدیل خطی روی داده‌های اصلی انجام می‌شود که این تبدیل طبق رابطه زیر صورت می‌گیرد:

$$v' = \frac{v - Min_A}{Max_A - Min_A}(New\_Max_A - New\_Min_A) + New\_Min_A$$

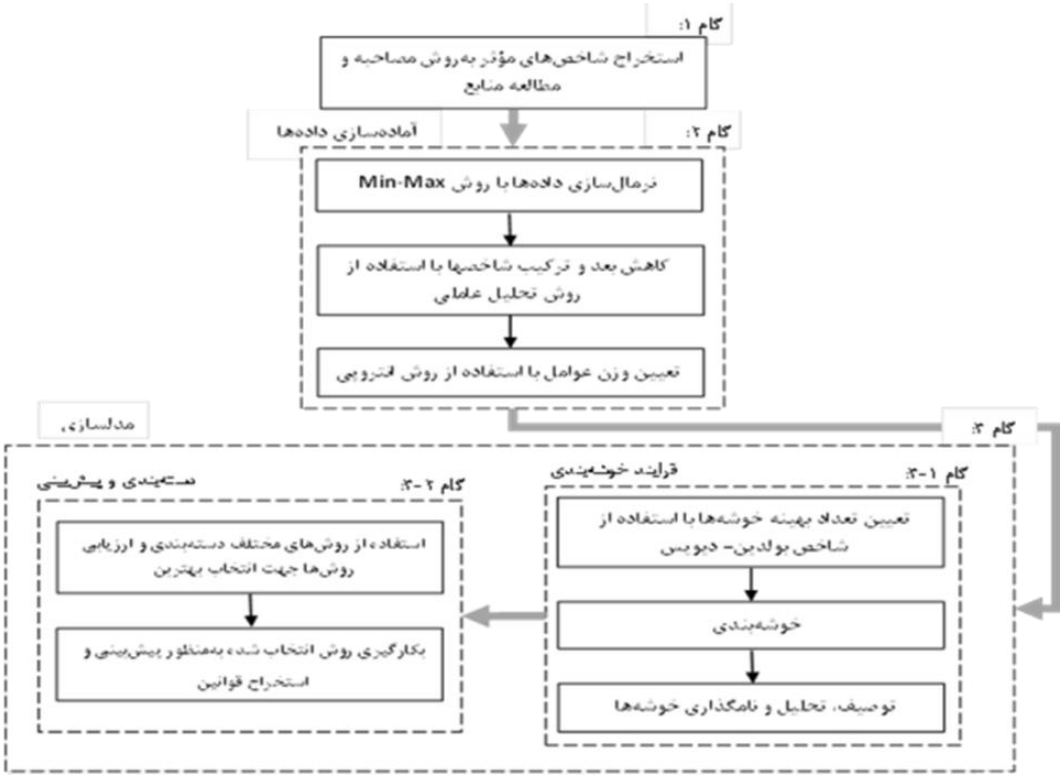
در این پژوهش ابتدا داده‌های مربوط به شاخصهای جدول ۱ جمع‌آوری و مورد پیش‌پردازش قرار گرفتند، که این پیش‌پردازش شامل، حذف نقاط مغشوش، حذف نقاط پرت، حذف داده‌های ناسازگار، تجمیع داده‌ها و ساخت انباره‌داده می‌باشد که در نتیجه ۷۱۹ رکورد داده‌های بخشهای تقسیمات کشوری مربوط به ۴۰ شاخص آماده گردید. سپس داده‌های که معرف شاخصهایی دارای ماهیت هزینه بودند، مانند نرخ بیکاری، معکوس شده تا به جنس سود تبدیل شوند. داده‌های بدست آمده باتوجه به فرمول ارائه شده در گام قبل در بازه [0,1] نرمال گردیدند.

18. Split selection
19. Gini index
20. Impurity based

کاهش بعد و ترکیب شاخص‌ها با استفاده از روش تحلیل عاملی

عامل متغیر جدیدی است که شامل ترکیب خطی مقادیر مشاهده شده متغیرها، مطابق معادله زیر است

$$F_j = \sum W_{ji} X_i = w_{j1}x_1 + w_{j2}x_2 + \ldots + w_{jp}x_p$$



کد شاخص	عنوان شاخص	کد شاخص	عنوان شاخص
1,5,9, 13,17	میانگین نمرات دروس عمومی شرکت‌کنندگان گروه‌های آزمایشی A، B، C، D و E (۵ شاخص)	29	فاصله مرکز بخش از مرکز شهرستان
2,6,10, 14,18	میانگین نمرات دروس اختصاصی شرکت‌کنندگان گروه‌های آزمایشی A، B، C، D و E (۵ شاخص)	30	فاصله مرکز بخش از مرکز استان
3,7,11, 15,19	میانگین نمرات دروس عمومی پذیرفته‌شدگان گروه‌های آزمایشی A، B، C، D و E (۵ شاخص)	31	فاصله مرکز بخش از اولین ایستگاه راه آهن
4,8,12, 16,20	میانگین نمرات دروس اختصاصی پذیرفته‌شدگان گروه‌های آزمایشی A، B، C، D و E (۵ شاخص)	32	تراکم خانواردر واحد مسکونی
21	درصد قبولی کل	33	نسبت مراکز بهداشتی درمانی به تعداد نقاط
22	درصد قبولی روزانه	34	نسبت پزشک به تعداد نقاط جمعیتی(شهر+روستا)
23	نسبت قبولی روزانه و شبانه به کل شرکت کنندگان	35	نسبت مدرسه راهنمایی به تعداد نقاط
24	تعداد پذیرفته شده	36	نسبت دبیرستان و هنرستان به تعداد نقاط
25	تعداد شرکت کننده	37	نسبت کتابخانه و کانون پرورش فکری کودکان به تعداد نقاط جمعیتی(شهر+روستا)
26	نرخ بیسوادی	38	نسبت تاسیسات ورزشی به تعداد نقاط
27	نرخ بیکاری	39	نسبت شعب بانک به تعداد نقاط جمعیت
28	نسبت تعداد شرکت کننده به جمعیت بالای ۶ سال	40	

در این رابطه x_i بیانگر متغیر i ام، w_{gi} ضریب نمره عاملی متغیر i ام و از نظر عامل j ام، p تعداد متغیرها و Fj عامل j است.
پس از ورود داده‌ها جهت بررسی کفایت تعداد داده‌ها، باید مقدار شاخص KMO از رابطه زیر محاسبه می‌شود که در آن r_{ij} ضریب همبستگی بین متغیرهای i و j و aij ضریب همبستگی جزئی بین آنهاست.

$$KMO = \frac{\sum\sum r_{ij}^2}{\sum\sum r_{ij}^2 + \sum\sum a_{ij}^2}$$

در پژوهش حاضر با توجه به اینکه تفسیر و تحلیل ۴۰ شاخص دشوار بوده و دارای پیچیدگی می‌باشد، با استفاده از روش تحلیل عاملی و محاسبه مقدار معیار KMO، همچنین ساخت ماتریس‌های مربوطه با استفاده از نرم‌افزار SPSS، پس از انجام ۱۰ مرتبه تکرار و حذف ۹ شاخص که دارای تأثیرگذاری کمتری بودند، نهایتاً ۸ عامل که نمایانگر ۳۱ شاخص باقیمانده بود، با دقت تبیین ۸۳٪ باقی ماند که در جدول (۲) ارائه شده‌اند:

تعیین وزن عوامل با استفاده از روش انتروپی

با توجه به اینکه همه عوامل دارای اهمیت یکسانی نمی‌باشند و برخی دارای اهمیت بیشتری نسبت به دیگر عوامل می‌باشد، می‌توان با استفاده از روش انتروپی میزان وزن عوامل را تعیین نمود(ماکویی،۲۰۰۷)
با فرض اینکه D ماتریس تصمیم‌گیری باشد، ستون‌ها معرف ویژگی‌ها و سطرها معرف اشیاء می‌باشند، به‌ترتیب با محاسبه P، K، E، d و W طبق روابط زیر وزن هریک از عوامل بدست خواهد آمد.

$$D = \begin{bmatrix} X_{11} & \cdot & \cdot & X_{1n} \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ X_{m1} & \cdot & \cdot & X_{mn} \end{bmatrix} \qquad P_{ij} = \frac{X_{ij}}{\sum_{i=1}^{m} X_{ij}} \qquad K = \frac{1}{\ln m}$$

$$E_j = -K \sum_{i=1}^{m} P_{ij} \ln P_{ij} \qquad d_j = 1 - E_j \qquad W_j = \frac{d_j}{\sum_{j=1}^{n} d_j}$$

در این پژوهش برای محاسبه میزان اهمیت هر یک از عوامل ، با استفاده از Oracle forms builder 10g بسته نرم افزاری ایجاد گردید که مقادیر Pij، K، Ej، dj و Wj توسط این بسته محاسبه و در نتیجه وزن عوامل مطابق جدول (۳) بدست آمد. همچنین پس از ضرب نمودن وزن در مقادیر هر یک از عوامل، عوامل وزندار حاصل گردید که بازه آنها مطابق جدول (۴) است.

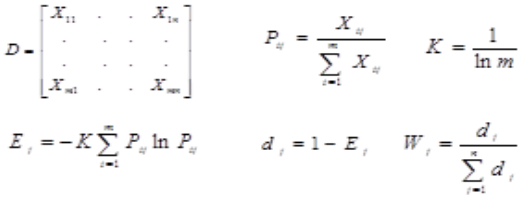
۳- مدلسازی

تعیین تعداد بهینه خوشه‌ها با استفاده از شاخص بولدین- دیویس

این شاخص برای یافتن تعداد بهینه خوشه‌ها در الگوریتم-هایی که برای بخش‌بندی نیازمند تعیین تعداد اولیه خوشه-ها می‌باشند مورد استفاده قرار می‌گیرد و از این منطق استفاده می‌کند که خوشه‌بندی مناسب، اولاً در آن میانگین فاصله عناصر درون هر خوشه از مرکز آن خوشه حداقل بوده و ثانیاً فاصله مراکز خوشه‌ها از یکدیگر حداکثر باشند که برای پیاده شدن این مفهوم از روابط زیر استفاده می-شود(دیویس،۱۹۹۷)

تعیین وزن عوامل با استفاده از روش انتروپی

با توجه به اینکه همه عوامل دارای اهمیت یکسانی نمی-باشند و برخی دارای اهمیت بیشتری نسبت به دیگر عوامل می‌باشد، می‌توان با استفاده از روش انتروپی میزان وزن عوامل را تعیین نمود(ماکویی،۲۰۰۷)
با فرض اینکه D ماتریس تصمیم‌گیری باشد، ستون‌ها معرف ویژگی‌ها و سطرها معرف اشیاء می‌باشند، به‌ترتیب با محاسبه P، K، E، d و W طبق روابط زیر وزن هریک از عوامل بدست خواهد آمد.



در این پژوهش برای محاسبه میزان اهمیت هر یک از عوامل ، با استفاده از Oracle forms builder 10g بسته نرم افزاری ایجاد گردید که مقادیر Pij، K، Ej، dj و Wj توسط
این بسته محاسبه و در نتیجه وزن عوامل مطابق جدول (۳) بدست آمد. همچنین پس از ضرب نمودن وزن در مقادیر هر یک از عوامل،
عوامل وزندار حاصل گردید که بازه آنها مطابق جدول (۴) است.

۳- مدلسازی

تعیین تعداد بهینه خوشه‌ها با استفاده از شاخص بولدین- دیویس

این شاخص برای یافتن تعداد بهینه خوشه‌ها در الگوریتم‌هایی که برای بخش‌بندی نیازمند تعیین تعداد اولیه خوشه‌ها می‌باشند مورد استفاده قرار می‌گیرد و از این منطق استفاده می‌کند که خوشه‌بندی مناسب، اولاً در آن میانگین فاصله عناصر درون هر خوشه از مرکز آن خوشه حداقل بوده و ثانیاً فاصله مراکز خوشه‌ها از یکدیگر حداکثر باشند که برای پیاده شدن این مفهوم از روابط زیر استفاده می‌شود(دیوایس،۱۹۹۷).

$$d_{ij} = \left\| z_i - z_j \right\| \qquad s_i = \frac{1}{C_i} \sum_{x \in C_i} \{ \| x - z_i \| \}$$

$$DB_{nc} = \frac{1}{nc} \sum_{i=1}^{nc} R_i$$

$$R_{ij} = \frac{s_i + s_j}{d_{ij}} \qquad R_i = \max_{i=1,\ldots,nc, i \neq j} R_{ij}$$

جدول(۲): عوامل، نام‌های پیشنهادی و شاخص‌های متعلق به هر عامل

شناسه عامل	نام عامل	کد شاخص‌ها	شناسه عامل	نام عامل	کد شاخص‌ها
F1	میزان برخورداری از تسهیلات	61,63,64, 65,66,67	F5	نمرات گروه D و E در کنکور	15,16,19, 20
F2	نمرات گروه A در کنکور	1,2,3,4	F6	میزان جذب جمعیت	24,25,70
F3	نمرات گروه B در کنکور و نرخ باسوادی	5,6,7,8, 27,29	F7	درصد قبولی در کنکور	21,22,23
F4	نمرات گروه C در کنکور	9,10,11,12	F8	نزدیکی مرکز بخش به مرکز استان	30

جدول(۳): وزن عوامل

شناسه عامل	F1	F2	F3	F4	F5	F6	F7	F8
وزن	0.071	0-656	0.008	0.003	0.103	0.121	0.005	0.033

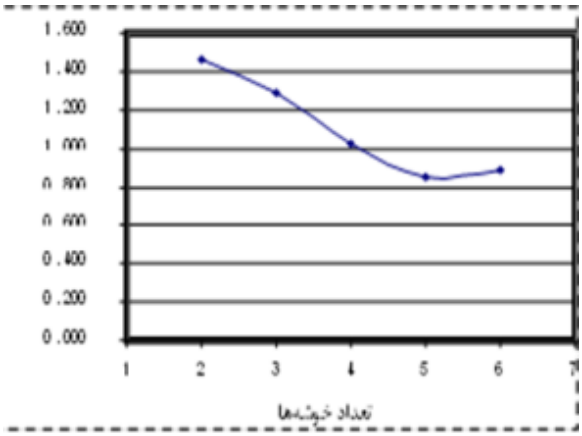
جدول(۴) بازه عوامل وزن‌دار

شناسه عامل	F1	F2	F3	F4	F5	F6	F7	F8
بازه	0-68	0-656	0-8	0-3	0-103	0-121	0-5	0-33

جدول(۵): تعداد نمونه‌ها در هر خوشه‌ها

شماره خوشه	1	2	3	4	5
تعداد نمونه‌های هر خوشه‌ها	379	25	8	194	113

برای بدست آوردن تعداد بهینه خوشه‌ها، الگوریتم K-means در بازه‌ی [۶ و ۲] که توسط افراد خبره تعیین شده بود، اجرا گردید و با انجام محاسبات مربوطه تعداد ۵ به‌عنوان تعداد خوشه بهینه بدست آمد که نتایج طبق شکل ۳ است.



شکل(۳): تعداد بهینه خوشه‌ها - شاخص بولدین- دیویس

خوشه‌بندی با استفاده از الگوریتم K-means

در این مرحله دو ماتریس باید تشکیل شود:

- ماتریس داده که نمایانگر n شیئ نظیر بخش‌های تقسیمات کشوری و p ویژگی یا عامل نظیر نرخ بیکاری است:
- ماتریس فاصله که میزان نزدیکی یا دوری هر زوج از اشیاء را نمایش می‌دهد که یک ماتریس n*n بوده و d(i,j) نمایانگر فاصله شیء i از شیء j است.

فاصله میان اشیاء که دارای مقیاس فاصله‌ای هستند، معمولاً از **روش اقلیدسی بر اساس رابطه زیر محاسبه می‌گردد:**

برای انجام این گام در پژوهش، الگوریتم K-means با استفاده از نرم‌افزار Clementine و با مقدار اولیه ۵ بر روی عوامل میزان برخورداری از تسهیلات، نمرات گروه ریاضی در کنکور، نمرات گروه تجربی در کنکور و نرخ باسوادی، نمرات گروه انسانی در کنکور، نمرات کروه زبان و هنر در کنکور، میزان جذب جمعیت، درصد قبولی در کنکور و نزدیکی مرکز بخش به مرکز استان اجرا گردید. تعداد اعضا هر خوشه مطابق جدول (۵) و فاصله میان مراکز خوشه‌ها در جدول (۶) ارائه گردیده‌اند.

I	j	d (i,j)
1	2	0.55
1	3	0.74
1	4	0.32
1	5	0.47
2	3	1.02
2	4	0.67
2	5	0.50
3	4	0.62
3	5	0.95
4	5	0.60

جدول (۶): فاصله میان مراکز خوشه‌ها

توصیف، تحلیل و نامگذاری خوشه‌ها

در این مرحله با استفاده از نمودارهای مربوطه، می‌توان خوشه‌های بدست آمده را با یکدیگر مقایسه نمود. میانگین و انحراف معیار مربوط به هر عامل در هریک از خوشه‌ها برای تحلیل ارائه می‌شود. همچنین فاصله مراکز خوشه‌ها از یکدیگر و فاصله هر عنصر از مرکز خوشه‌ی خود تولید شده است. بدین ترتیب بینشی در مورد هر خوشه ایجاد می‌گردد که در نتیجه می‌توان نام مناسبی برای خوشه‌ها انتخاب نمود و با دانش بدست آمده استراتژی مناسب برای هر خوشه را تعیین کرد.

طبق نتایج بدست آمده از خوشه‌بندی در این پژوهش، بخش‌هایی از کشور که در خوشه سوم قرار گرفته‌اند در کلیه عوامل، بجز عامل نمرات گروه زبان و هنر از دیگر خوشه‌ها برترند، بویژه این نسبت برتری در دو عامل میزان جذب جمعیت و میزان برخورداری از تسهیلات دارای اختلاف زیادی با دیگر خوشه‌هاست. بنابراین می‌توان خوشه سوم را با برچسب "بخش‌های توسعه‌یافته" شناخت، که البته این خوشه شامل مراکز بعضی استان‌های بزرگ مانند اصفهان، تبریز و شیراز می‌باشد. با توجه به میزان فاصله مراکز خوشه‌ها از خوشه سوم و همچنین بررسی وضعیت عوامل در هرخوشه می‌توان وضعیت دیگر خوشه‌ها را بدین ترتیب ذکر نمود:

- خوشه چهارم در تمامی ۸ عامل، بعد از خوشه سوم از دیگر خوشه‌ها برتر است. می‌توان برچسب کلاس این خوشه را "بخش‌های در حال توسعه" در نظر گرفت. بیشتر بخش‌های مرکزی شهرستان‌های کشور، زیرمجموعه این خوشه هستند.
- خوشه اول رتبه سوم فاصله از مرکز خوشه سوم را داراست، این خوشه از نظر عوامل درصد قبولی در کنکور و میانگین نمرات گروه‌های مختلف شرکت‌کننده، بجز گروه زبان و هنر در سطح خوبی قرار دارد. الیته عامل میزان جذب جمعیت آن در سطح پایینی قرار دارد. برچسب "کمتر توسعه‌یافته" برای این خوشه درنظر گرفته شده است.
- در خوشه‌هایی که عامل جذب جمعیت پایین و درصد قبولی بالا می‌باشد، باید به این موضوع توجه نمود که ممکن است در این خوشه‌ها بخش‌هایی وجود داشته باشند که از تعداد شرکت‌کننده مثلاً ۲ نفر، یک نفر آنها پذیرفته شده باشد و در نتیجه عامل درصد قبولی ۵۰ درصد خواهد شد که میزان بالایی است.اما با توجه به پایین بودن تعداد شرکت‌کننده نمی‌تواند دلیلی بر مناسب بودن درصد قبولی در آن بخش تلقی گردد.
- در خوشه دوم و پنجم وضعیت عوامل نسبت به سه خوشه دیگر پایین‌تر می‌باشد و می‌توان عناصر این خوشه را جزء "بخش‌های محروم" در نظر گرفت. .با مقایسه این دو خوشه با یکدیگر، عامل جذب جمعیت، میزان برخورداری از تسهیلات، همچنین نمرات گروه‌های مختلف شرکت‌کننده در کنکور، در خوشه دوم نسبت به خوشه پنجم وضعیت بهتر می‌باشد، اما عناصر خوشه پنجم به مراکز استان‌های خود نزدیکترند.

دسته‌بندی و پیش‌بینی

با توجه به آنکه هر ساله بخش‌های جدیدی به تقسیمات کشوری اضافه می‌شوند، بکارگیری کلیه مراحل مدل پیشنهادی جهت خوشه‌بندی نیازمند صرف زمان زیادی می‌باشد، لذا استفاده از الگوریتم‌های دسته‌بندی جهت پیش‌بینی کلاس مربوط به هر بخش جدید می‌تواند راه‌کاری بهینه محسوب گردد. همچنین ایجاد درخت تصمیم و استخراج قوانین مربوط به‌آن نگرشی مفید برای تصمیم-گیران ایجاد خواهد نمود. در این گام روش‌های مناسب برای پیش‌بینی را با توجه به نوع داده‌ها انتخاب نموده و سپس برای ارزیابی و انتخاب روش مناسب، با تقسیم داده‌ها به دو دسته‌ی آموزش و آزمون دقت مدل‌ها بررسی می‌شود.

برای این مرحله در پژوهش حاضر از سه الگوریتم شبکه‌های عصبی، C&RT و CHAID بر روی داده‌های آموزشی و آزمایشی استفاده شده است و ارزیابی نتایج بدست آمده در جدول ۷ ارائه شده است:

برای انتخاب روش مناسب باتوجه به درصد پیش‌بینی صحیح الگوریتمی که دارای بالاترین مقدار است را انتخاب می‌کنیم.

یافته های پژوهش

بر اساس نتایج حاصل از ارزیابی مدل‌های مختلف جدول حاصل از مرحله قبل، از الگوریتم C&RT برای ساخت درخت و استخراج قوانین و از شبکه عصبی برای پیش‌بینی برچسب کلاس بخش‌های جدید استفاده شده است. از درخت بدست آمده تعداد **۹** قانون استخراج شده است که این قوانین طبق شکل (۴) و ذیلاً توصیف شده‌اند:

۱. اگر (مقدار عامل نمرات گروه زبان و هنر کوچکتر یا مساوی ۲۱,۹) و (مقدار عامل نمرات گروه ریاضی در کنکور کوچکتر یا مساوی ۱۶۹,۸) و (مقدار عامل نزدیکی مرکز بخش به مرکز استان کوچکتر یا مساوی ۱۹,۴) باشد، آنگاه (بخش با این ویژگی‌ها با قطعیت ۹۰٪ متعلق به خوشه مناطق محروم خواهد بود).

۲. اگر (مقدار عامل نمرات گروه زبان و هنر کوچکتر یا مساوی ۲۱,۹) و (مقدار عامل نمرات گروه ریاضی در کنکور کوچکتر یا مساوی ۱۶۹,۸) و (مقدار عامل نزدیکی مرکز بخش به مرکز استان بزرگتر از ۱۹,۴) و (مقدار عامل نمرات گروه تجربی و نرخ باسوادی کوچکتر یا مساوی ۴,۲) باشد، آنگاه (بخش با این ویژگی‌ها با قطعیت ۹۷٪ متعلق به خوشه پنجم خواهد بود).

۳. اگر (مقدار عامل نمرات گروه زبان و هنر کوچکتر یا مساوی ۲۱,۹) و (مقدار عامل نمرات گروه ریاضی در کنکور کوچکتر یا مساوی ۱۶۹,۸) و (مقدار عامل نزدیکی مرکز بخش به مرکز استان بزرگتر از ۱۹,۴) و (مقدار عامل نمرات گروه تجربی و نرخ باسوادی بزرگتر از ۴,۲) باشد، آنگاه (بخش با این ویژگی‌ها با قطعیت ۷۵٪ متعلق به خوشه اول خواهد بود).

۴. اگر (مقدار عامل نمرات گروه زبان و هنر کوچکتر یا مساوی ۲۱,۹) و (مقدار عامل نمرات گروه ریاضی در کنکور بزرگتر از ۱۶۹,۸) و (مقدار عامل درصد قبولی در کنکور کوچکتر یا مساوی ۱,۴۶) و (مقدار عامل نزدیکی مرکز بخش به مرکز استان کوچکتر یا مساوی ۲۱,۹۸) باشد، آنگاه (بخش با این ویژگی‌ها با قطعیت ۱۰۰٪ متعلق به خوشه دوم خواهد بود).

۵. اگر (مقدار عامل نمرات گروه زبان و هنر کوچکتر یا مساوی ۲۱,۹) و (مقدار عامل نمرات گروه ریاضی در کنکور بزرگتر از ۱۶۹,۸) و (مقدار عامل درصد قبولی در کنکور کوچکتر یا مساوی ۱,۴۶) و (مقدار عامل نزدیکی مرکز بخش به مرکز استان بزرگتر از ۲۱,۹۸) باشد، آنگاه (بخش با این ویژگی‌ها با قطعیت ۶۷٪ متعلق به خوشه پنجم خواهد بود).

۶. اگر (مقدار عامل نمرات گروه زبان و هنر کوچکتر یا مساوی ۱۶) و (مقدار عامل نمرات گروه ریاضی در کنکور بزرگتر از ۱۶۹,۸) و (مقدار عامل مقدار عامل درصد قبولی در کنکور بزرگتر از ۱,۴۶) باشد، آنگاه (بخش با این ویژگی‌ها با قطعیت ۹۶٪ متعلق به خوشه اول خواهد بود).

۷. اگر (مقدار عامل نمرات گروه زبان و هنر کوچکتر یا مساوی ۲۱,۹ و بزرگتر از ۱۶) و (مقدار عامل نمرات گروه ریاضی در کنکور بزرگتر از ۱۶۹,۸ و کوچکتر یا مساوی ۳۳۶) و (مقدار عامل درصد قبولی در کنکور بزرگتر از ۱,۴۶) باشد، آنگاه (بخش با این ویژگی‌ها با قطعیت ۸۹٪ متعلق به خوشه اول خواهد بود).

۸. اگر (مقدار عامل نمرات گروه زبان و هنر کوچکتر یا مساوی ۲۱,۹ و بزرگتر از ۱۶) و (مقدار عامل نمرات گروه ریاضی در کنکور بزرگتر از ۳۳۶) و (مقدار عامل درصد قبولی در کنکور بزرگتر از ۱,۴۶) باشد، آنگاه (بخش با این ویژگی‌ها با قطعیت ۹۳٪ متعلق به خوشه چهارم خواهد بود).

۹. اگر (مقدار عامل نمرات گروه زبان و هنر بزرگتر از ۲۱,۹) باشد، آنگاه (بخش با این ویژگی‌ها با قطعیت ۹۳٪ متعلق به خوشه چهارم خواهد بود).

بحث و نتیجه گیری

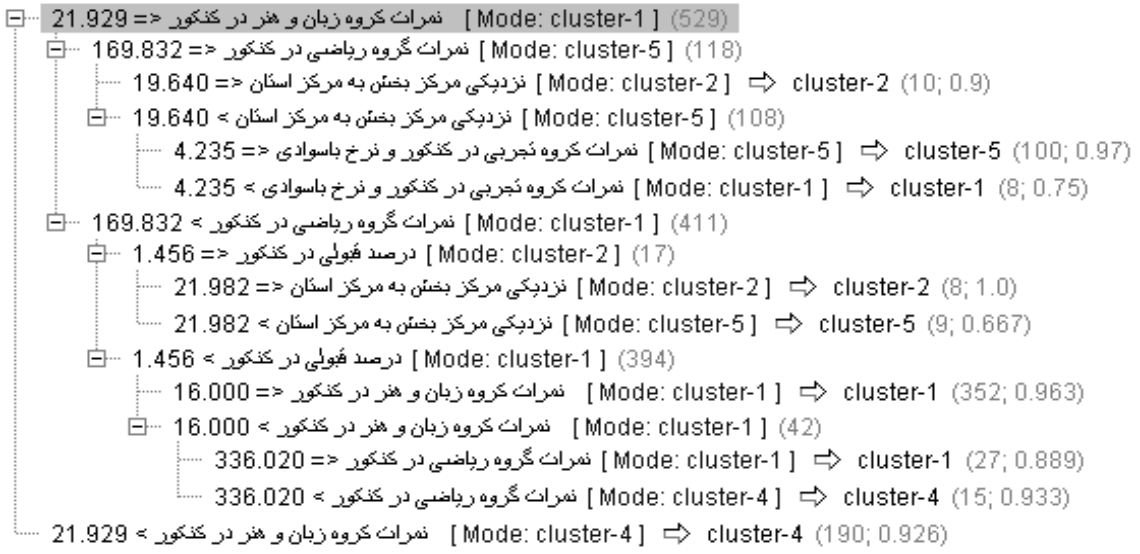
با توجه به اینکه طبق بررسی‌های صورت‌گرفته توسط محققین این طرح، تاکنون کار منتشر شده‌ای در زمینه‌ی بخش‌بندی مناطق و پیش‌بینی در راستای توسعه‌یافتگی صورت نگرفته است و تحقیقات و کارهای انجام شده صرفاً رتبه‌بندی[21] بوده‌اند، در نتیجه برای از بین بردن ضعف رتبه بندی که امکان تحلیل گروهی بخش‌ها وجود ندارد و همچنین در روش‌های قبلی برای تعیین رتبه مناطق جدید باید کل فرایند از ابتدا صورت می‌گرفت، تحقیق حاضر صورت گرفته است. بدین منظور با بررسی استانداردهای بین-المللی موجود در این زمینه و استخراج شاخص‌های بومی و نیز بکارگیری تکنیک‌های داده‌کاوی مدل جدیدی که حاصل ترکیب روش‌های مختلف آماری (تحلیل عاملی)، پایگاه‌داده‌ها، یادگیری ماشین (شبکه عصبی)، تصمیم‌گیری (انتروپی) و داده‌کاوی می‌باشد، ارائه شده است.

پس از جمع‌آوری داده‌های مرتبط با شاخص‌ها، از سازمان-های ذیربط و ساخت انباره داده مربوطه، داده‌ها در قالب مدل پیشنهادی بکار گرفته شده‌اند تا در نهایت خروجی‌های بدست آمده جهت تدوین استراتژی در اختیار تصمیم‌گیران قرار گیرد. دستاوردهای این تحقیق عبارتند از: تعیین تعداد بهینه بخش‌ها، بخش‌بندی مناطق، تحلیل هر بخش، استخراج قواعد تصمیم‌گیری، امکان پیش‌بینی سریع‌تر و دقیق‌تر برچسب کلاس برای مناطق جدید، فراهم نمودن امکان تدوین راهبردهای مناسب برای هر بخش و تخصیص میزان سهمیه مناسب به دانش‌آموزان هر منطقه جهت ورود به سازمان‌هایی مانند دانشگاه آزاد، دانشگاه دولتی و وزارت کار. با توجه به اینکه مناطق مختلف جغرافیایی می‌توانند دارای خصوصیات منحصر بفرد خود باشند و ممکن است شاخصی که در یک منطقه دارای اهمیت بالایی است، در منطقه‌ای دیگر از اهمیت کمی برخوردار باشد تقسیم نمودن اولیه کشور به چند ناحیه و تعیین وزن شاخص‌ها در هرناحیه بطور مجزا و سپس بکارگیری فرایند مدل پیشنهادی، جهت کسب نتایج دقیق‌تر و منطقی‌تر، می‌تواند بعنوان تحقیقات آتی در نظر گرفته شود

21. Ranking

جدول (۷): ارزیابی نتایج دسته‌بندی

نام روش	تعداد پیش‌بینی صحیح	درصد پیش‌بینی صحیح	تعداد پیش‌بینی نادرست	درصد پیش‌بینی نادرست
C&RT	۶۷۹	٪۹۴,۴۴	۴۰	٪۵,۵۶
CHAID	۶۵۹	٪۹۱,۶۶	۶۰	٪۸,۳۴
شبکه‌های عصبی	۶۹۸	٪۹۶,۸	۲۳	٪۳,۲



شکل (۴): قوانین استخراج شده از درخت تصمیم

منابع:

۱. توداور، مایکل ، ۱۳۷۸، *توسعه اقتصادی در جهان سوم*، ترجمه غلامعلی فرجادی، سازمان برنامه و بودجه ، ص ۲۵).

2. Berry, M. J. A., & Linoff, G. S. (2004). *Data Mining Techniques: For Marketing, Sales And Customer Support*, John Wiley And Sons.

3. Cheng, C.-H., & Chen, Y.-S. (2009). *Classifying the segmentation of customer value via RFM model and RS theory*, Expert Systems with Applications, 36, 4176–4184.

4. Chiu, C. Y., Chen, Y. F., Kuo, I. T., & Ku, H. C. (2009). *An Intelligent Market Segmentation System Using K-Means And Particle Swarm Optimization*. Expert Systems with Applications, 36, 4558-4565.

5. Davies, D.L., & Bouldin, D.W. (1979), *A Cluster Separation Measure. IEEE Transaction on Pattern Analysis and Machine Intelligence*, 224-227.

6. Han, J., & Kamber, M. (2006). *Data Mining: Concepts And Techniques*, (2nd ed.). San Francisco: Morgan Kaufmann Publishers.

7. Hair, J. F., Anderson, R. E., Tatham, R. L., & Black, W. C. (1998). *Multivariate Data Analysis*. Prentice Hall, (Chapter 3).

8. Hand, D. (1998). *Data Mining: Statistics And More?*. The American Statistician, Vol. 52.

9. Hildebrandt, M., & Gutwirth, S. (2008). *Profiling The European Citizen: Cross-Disciplinary Perspectives*, Springer Publishing Company.

10. Kantardzic, M. (2003). *Data Mining: Concepts, Models, Methods, And Algorithms*. John Wiley & Sons Inc.
11. Larose, D. T. (2005). *Discovering In Knowledge DataAn Introduction To Data Mining*, John Wiley And Sons.
12. Liu H., & Motoda H. (2001). *Instance Selection and Construction for Data Mining* , Publisher: Kluwer Academic Publishers Norwell, MA, USA.
13. Makui, A. (2007). *Decision Making Techniques*. Mehr va mahe no Pubublier, (Chapter 3 in persian).
14. Milutinovi, V., & Patricelli, F. (2002). *E-Business And E-Challenges*, Publisher: Ios Press Inc.
15. Myatt, G. J. (2006), *Making Sense Of Data: A Practical Guide To Exploratory Data Analysis And Data Mining*, John Wiley And Sons Publication, (Chapter 3).
16. Parr, R. O. (2001). *Data Mining Cookbook Modeling Data for Marketing, Risk, and Customer Relationship Management*. John Wiley & Sons Inc., Wiley computer publishing, (Chapter 3).
17. Pregibond. (2001). *a statistical odyssey, proceedings of the fifth acm sigkdd,* international conference on knowledge discovery and data mining.
18. Rokach L., & Maimon O.(2008). *Data Mining with Decision Trees: Theory and Applications*, World Scientific Publishing.
19. Soukup, T., & Davidson, I. (2002). *Visual Data Mining: Techniques and Tools for Data Visualization and Mining*. John Wiley & Sons, (Chapters 4, 5 and 6).
20. Watkins, K. (2007/2008), *Human Development Report,* United Nations Development Programme, http://hdr.undp.org.
21. Ye, N. (2003), *The Handbook Of Data Mining*. Arizona State University, Lawrence Erlbaum Associates Publishers, (Chapter 14).