تعیین عوامل موثر در هزینه و زمان حمل‌ونقل جاده‌ای و ریلی با قواعد انجمنی

مقاله علمی – پژوهشی

*مهدی یوسفی نژاد عطاری (نویسنده مسئول)، گروه مهندسی صنایع، واحد بناب، دانشگاه آزاد اسلامی، بناب، ایران
انسیه نیشابوری جامی، گروه مهندسی صنایع، دانشگاه بناب، بناب، ایران
*پست الکترونیکی نویسنده مسئول: 1689597631@iau.ir

دریافت: ۱۴۰۵/۰۳/۲۱- پذیرش: ۱۴۰۵/۰۵/۲۵
صفحه ۵۳۸-۵۱۳

چکیده

حمل‌ونقل به جابه‌جایی انسان، حیوان یا کالا از یک مکان به مکان دیگر گفته می‌شود. به بیان دیگر، حمل‌ونقل به‌عنوان حرکتی مشخص از نقطه A به نقطه B تعریف می‌شود. از انواع روش‌های حمل‌ونقل می‌توان به حمل‌ونقل هوایی، زمینی (ریلی و جاده‌ای)، دریایی، کابلی، خط لوله و فضایی اشاره کرد. در حمل‌ونقل زمینی، به دلیل اهمیت این کسب‌وکار و رقابت میان سازمان‌های فعال در این حوزه، به‌کارگیری فناوری‌ها و تکنولوژی‌های نوین در مدیریت و اتخاذ تصمیمات بهتر می‌تواند بسیار سودمند باشد. در این مقاله، نخست به معرفی داده‌کاوی، تکنیک‌های آن و نرم‌افزارهای مرتبط با داده‌کاوی می‌پردازیم. سپس، با استفاده از جدولی از حمل‌های صادراتی انجام شده در بازه زمانی یک سال و شش ماه گذشته که از یک شرکت حمل‌ونقل بین‌المللی تهیه شده است، قوانین انجمنی را با استفاده از نرم‌افزار کلمنتاین در دو مدل استخراج کرده و قوانین به‌دست‌آمده را مورد ارزیابی قرار داده شود.

واژه‌های کلیدی: داده‌کاوی، زمان، حمل‌ونقل جاده‌ای و ریلی، خوشه‌بندی، هزینه مقدمه

۱-مقدمه

در دنیای امروز داده‌ها و اطلاعات بعنوان ثروت سازمانی محسوب گشته و همواره شرکت‌ها و سازمان‌های بزرگ و موفق دنیا به دنبال استفاده مناسب‌تر و تجاری‌تر از این منابع مجازی می‌باشند. از سوی دیگر با پیچیده شدن محیط‌های کسب‌وکار، ماهیت و حجم داده‌های سازمانی بسیار متفاوت گشته و نگاه یکپارچه و مدیریتی به آنها ضروری می‌گردد. یکی از راه‌حل‌هایی که سازمان‌های موفق دنیا در این زمینه اتخاذ می‌نمایند، ایجاد یک سیستم جامع داده‌ای و آماری و بکارگیری صحیح تکنیک‌های داده‌کاوی و کشف دانش می‌باشد (اولسون، ۲۰۰۳). یکی از راه‌حل‌هایی که می‌تواند تصمیم‌گیری در چنین فعالیت‌هایی را تسریع بخشیده و دقت تصمیم‌گیری را افزایش دهد، بکارگیری صحیح تکنیک‌های داده‌کاوی بر روی داده‌های موجود، جهت کشف دانش‌های سازمانی و یا پیش بینی وضعیت آینده است. حمل‌ونقل نقش مهمی را در رشد اقتصادی و جهانی شدن ایفا می‌کند.

۲-پیشینه تحقیق

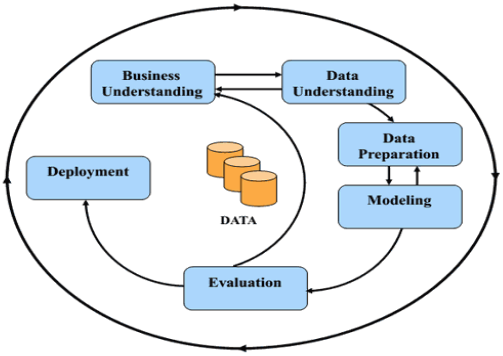
تا کنون تعاریف متعددی برای داده‌کاوی ارائه شده است. در برخی از این تعاریف داده‌کاوی در حد ابزاریست که کاربران را قادر به ارتباط مستقیم با حجم عظیم داده‌ها می سازد و در برخی تعاریف دقیق‌تر به کاوش در داده‌ها نیز توجه شده است. در ادامه به برخی از این تعاریف اشاره می‌کنیم: اصطلاح داده‌کاوی به فرایند نیم‌خودکار تجزیه‌وتحلیل پایگاه داده‌های

بزرگ به منظور یافتن الگوهای مفید اطلاق می‌شـــود (هان و کمبار،۲۰۱۱). به طور خلاصه می‌توان داده‌کاوی را جستجو در مجموعه‌ای از داده‌ها برای یافتن الگوهایی در بین آن‌ها نامید و یا داده‌کاوی اســتخراج دانش جدید و قابل اســتناد از مجموعه پایگاه داده‌های بزرگ نیز معرفی شـــده اســـت. (دیوید و هاند، ۲۰۰۱). از منظر داده‌کـاوی یعنی تجزیـه‌وتحلیـل مجموعـه داده‌های قابل مشـــاهده برای یافتن روابط مطمئن بین داده‌ها تعریف می‌شود(هربرت، ۱۹۹۶).

۲-۱-فرآیند داده‌کاوی

بر اساس فرایند استاندارد داده‌کاوی کریسپ می‌توان پنج فاز را برای پروژه‌های داده‌کاوی در نظر گرفت که عبارتند از: شناخت کسب‌وکار، شناخت داده، آماده‌سازی داده‌ها، مدلسازی، ارزیابی و بکارگیری(جانگ، ۲۰۱۶).

این فرایند ا ستاندارد که شمای آن در شکل ۱ موجود ا ست فرایند بصـــورت خطی حرکت نمی کند بلکه در هر مرحله امکان بازگشـــت به عقب وجود دارد و این روند آنقدر تکرار می‌شود تا در نهایت به بهترین نتیجه برسیم.



شکل ۱. مدل استاندراد غیر خطی داده کاوی (جانگ، ۲۰۱۶)

۲-۲-الگوریتم کا-میانگین

خو شه‌بندی یکی از شاخه‌های یادگیری بدون نظارت ا ست و فرآیند خودکاری ا ست که در طی آن، نمونه‌ها به د سته‌هایی که اعضای آن م شابه یکدیگر می‌با شند تقسیم می شوند که به این دسـته‌ها خوشـه گفته می‌شـود. بنابراین خوشـه مجموعه‌ای از ا شیاء ا ست که در آن ا شیاء با یکدیگر م شابه بوده و با ا شیاء موجود در خو شه‌های دیگر غیرم شابه می‌با شند. برای م شابه بودن می‌توان معیارهای مختلفی را در نظر گرفت مثلاً می‌توان معیار فا صله را برای خو شه‌بندی مورداستفاده قرارداد و ا شیائی را که به یکدیگر نزدیک‌تر هستند را به‌عنوان یک خوشه در نظر گرفت که به این نوع خو شه‌بندی، خو شه‌بندی مبتنی بر فا صله نیز گفته می شود. یکی از این روش‌ها خو شه‌بندی k-means است (هان و کمبار،۲۰۱۱).

در نوع ساده‌ای از این روش ابتدا به تعداد خو شه‌های موردنیاز نقاطی به صورت ت صادفی انتخاب می شود. سپس در داده‌ها با توجه با میزان نزدیکی (شباهت) به یکی از این خوشه‌ها نسبت داده می‌شـــوند و بدین ترتیب خوشـــه‌های جدیدی حاصـــل می‌شـــود. با تکرار همین روال می‌توان در هر تکرار با میانگین گیری از داده‌ها مراکز جدیدی برای آن‌ها محا سبه کرد و مجدداً داده‌ها را به خوشـــه‌های جدید نســـبت داد. این روند تا زمانی ادامه پیدا می‌کند که دیگر تغییری در داده‌ها حاصـــل نشـــود (کرمیک و سوبای، ۲۰۱۵).

۲-۳-ارزیابی خوشه‌بندی (شاخص دیویس بولدین)

برای ارزیابی الگوریتم‌های خوشه‌بندی معیارهای متفاوتی وجود دارد که می‌توانیم آن‌ها را به دو دسته بدون ناظر و با ناظر تقسیم کنیم. معیارهای ارزیابی بدون ناظر که گاهی در متون علمی از آن با نام معیارهای داخلی نیز یاد می‌شـــود آن دســـته عملیاتی هســـتند که با توجه به اطلاعات موجود در مجموعه داده به کار می‌روند. مهم‌ترین وظیفه یک الگوریتم خوشـــه‌بندی این اســـت که بتواند به بهترین شـــکل فاصـــله درون خوشـــه‌ای را کمینه و فا صله بین خو شه‌ها را بی شینه کند. ازاین‌رو دو فاکتور مهم که در تمامی معیارهای ارزیابی به کار می‌رود: تراکم خوشـــه‌ای و جدایی خوشـــه‌ای اســـت؛ و برآورده شـــدن اهداف کمینه‌ســـازی

فا صله درون خو شه‌ای و بی شینه سازی فا صله میان خو شه‌ای به ترتیب در گروه بیشـــینه نمودن تراکم هر خوشـــه و نیز بیشینه‌سازی جدایی میان خوشه‌ها است.

یکی از این معیارها دیویس بولدین اســت که بر اســاس این دو فاکتور مهم فرمول آن را به این صورت تعریف می‌شود.

فاکتور جدایی یا Sep در این معیار به‌صـــورت $dis(C_i,C_{j,})$ تعریف می‌شـــود که در آن dis فاصـــله اقلیدســـی بین مرکز دو

(۱)

۲-٤-الگوریتم FP-Growth

برای تولید قوانین انجمنی الگوریتم‌های زیادی وجود دارند دو تـا از پرکـار برد تر ین روش‌هـا ا لگور یـتم Apriority و FP-Growth اســت. که در این تحقیق از روش دوم اســتفاده می‌کنیم. که در آن بدون تولید مجموعه اقلام کاندید و با کمک یک ســاختمان داده‌های درختی الگوهایی مکرر را شـناسـایی می‌کند. این الگوریتم که اســتراتژی تقســیم و غلبه را دنبال می‌کند، در ابتدا مجموعه داده را به‌صورت یک درخت موسوم به FP-tree تبدیل می‌کند. پس‌ازآن به‌صـــورت مســـتقیم به استخراج الگوهای مکرر از این درخت می‌پردازد.

به‌طور خلاصه می‌توان گفت یک FP-tree نمایش فشرده‌ای از داده‌های مجموعه تراکنش اســت. برای ســاخت این درخت الگوریتم، اقلام تراکنش‌ها را یک‌به‌یک خوانده و به‌صورت یک مســیر بر روی درخت نگاشـــت می‌کند. ازآنجاکه معمولاً تراکنش‌ها از اقلام مشـــترکی اســـتفاده می‌کنند، لذا مســـیرهای تراکنش‌ها بر روی درخت همپوشـــانی دارند و به همین دلیل درخت نسبت به داده‌های اولیه فشرده‌تر خواهد بود. در اولین اســکن پایگاه داده مجموعه آیتم‌های یک عضـــوی و پشـــتیبانی آن‌ها مشـــخص می‌شـــود و نیز مجموعه اقلام پرتکرار به ترتیب نزولی پشتیبانی شان مرتب می شوند. سپس یک درخت به این صورت ساخته می شود که: ابتدا ریشه درخت با برچسب null ســـاخته می‌شـــود. بعدازآن مجموعه داده برای بار دوم اســـکن می شود. اقلام هر تراکنش به ترتیب L پردازش می شوند و یک شاخه برای هر تراکنش ایجاد می شود. به‌منظور تسهیل پیمایش درخت، یک جدول ســـاخته می‌شـــود که هر قلم در آن به محل خودش در درخت اشـــاره می‌کند. درخت پس از اســـکن همه تراکنش‌ها کامل می‌شود(کرمیک و سوبای، ۲۰۱۵).

خوشه است.

فاکتور تراکم یا Coho به صـــورت $\frac{1}{m}\ \sum_{x\in C_i,} dis(x,C_{i,})$ تعریف می شود. که در آن x نقطه داخل خو شه و c مرکز همان خوشه است.

فرمول کلی دیویس بولدین به صـــورت زیر اســـت: مقدار NC برای تمام نقاط است.

$$\frac{1}{nc}\sum_{i=1}^{i\ nc}\left\{\frac{coh(i)+coh(j)}{Sep(i,j)}\right\}$$

۲-۵-الگوریتم دسته بندی

تئوری بیز یکی از روش‌های آماری دسته‌بندی به شمار می‌آید. در این روش رده‌های مختلف، هرکدام به شـــکل یک فرضـــیه دارای احتمال در نظر گرفته می‌شـــوند. هر رکورد آموزشـــی جدید، احتمال در ست بودن فر ضیه‌های پیشین را افزایش و یا کاهش می‌دهد و درنهایت، فرضـیاتی که دارای بالاترین احتمال شوند، به‌عنوان یک کلاس در نظر گرفته‌شده و برچسبی بر آن‌ها زده می‌شـــود. این فن با ترکیب تئوری بیز و رابطه ســـببی بین داده‌ها، به طبقه‌بندی می‌پردازد (هان و کمبار،۲۰۱۱).

جنگل تصادفی درخت تصمیم‌های زیادی را تولید می‌کند. برای طبقه‌بندی یک شیء جدید، برداری ورودی را در انتهای هر یک از درختان جنگل تصادفی قرار می‌دهد. هر درختی ک طبقه‌بندی را ارائه می‌کند و اصـــطلاحاً گفته می‌شـــود این درخت به آن کلاس «رأی» می‌دهد. جنگل طبقه‌بندی را که بیشـــترین رأی داشته باشد (بین همه درخت‌های جنگل) انتخاب می‌کند. (لی و چن، ۲۰۱٤)

هر درخت به‌صورت زیر تشکیل می‌شود:

-اگر N تعداد حا لت ها در مجمو عه داده های آموزشـــی (مجموعه‌ی کار) باشـــد، N حالت را به‌صـــورت تصـــادفی با جایگذاری از داده‌های اصـــلی، نمونه‌گیری می‌کنیم. این نمونه مجموعه‌ی کار برای این درخت است.

-اگر M متغیر داشـــته باشـــیم و m را کوچک‌تر از M در نظر بگیریم به‌طوری‌که در هر گره، m متغیر به‌صـــورت تصـــادفی از M انتخاب می‌شـــوند و بهترین جداســـازی روی این m متغیر برای جداسازی گره استفاده می شود. مقدار m در طول ساخت جنگل ثابت در نظر گرفته می‌شود.

– هر درخت به‌اندازه‌ی ممکن بزرگ می‌شود. هیچ هرسی وجود ندارد.

۲-٦-داده کاوی در حمل‌ونقل

در حوزه مهندسی حمل‌ونقل مقدار زیادی از داده‌ها در زمینه ترافیک، تحلیل حوادث و تصادفات، شرایط آسفالت، دوام جاده‌ها، مدیریت پل‌ها و غیره ا ستخراج شده ا ست. بر ا ساس این داده ها، تصمیم‌گیرندگان تصمیم می‌گیرند تا یک مشکل مربوط را حل کنند. تصمیم‌گیرندگان همیشه به دنبال روش‌هایی برای د ستر سی راحت به داده‌ها و ا ستفاده از داده‌های متفاوتی هستند. توانایی شناسایی داده‌های در دسترس، تعیین ویژگی‌های داده‌ها، ا ستخراج داده‌های مورد نیاز، تبدیل داده‌ها به فرمت‌های لازم جزو الزامات لازم برای برنامه‌ریزی هستند. در دنیای واقعی حمل‌ونقل اطلاعات مختلف زیادی باید با هم ادغام شده تا به راه حل بر سیم. تحقیقات اخیر در زمینه داده‌کاوی در حمل‌ونقل افق جدیدی را برای تصمیم‌گیرندگان مهندسان حمل‌ونقل ایجاد کرده است.

در سال ۲۰۱٤ لی و چن با ادغام سه روش داده‌کاوی، خو شه بندی کا-میانگین، درخت‌های تصمیم‌گیری، و شبکه‌های عصبی، به پیش‌بینی زمان سفر آزادراه با ازدحام غیر مکرر پرداختند. (لی و چن، ۲۰۱٤).

احمدوند و همکاران با استفاده از الگوریتم اپریوری، قوانین انجمنی موجود در بازه‌های زمانی/مکانی رخداد مشکلات به دست آورده‌اند. همچنین با استفاده از یک تکنیک ۲ مرحله‌ای از این الگوریتم، مشکلاتی که با هم و به صورت همزمان در یک بازه زمانی مشخص اتفاق می‌افتند را شنا سایی کرده و در نهایت نشان داده‌اند که از قوانین بدست آمده می‌توان برای شناسایی نیازهای شهروندان و بهبود مدیریت خدمات شهری استفاده نمود (احمدوند، ۱۳۸۸).

پیتر ژانگ، در تحقیق خود در زمینه پیش‌بینی سری‌های زمانی از مدل ترکیبی م یانگین متحرک خودهمبسته یکپار چه و شبکه‌های عصبی ا ستفاده کرد و به این نتیجه ر سید که ا ستفاده از مدل‌های سری زمانی به تنهایی از دقت و قدرت شبکه‌های عصبی برخوردار نیست. لذا ترکیب مدل‌های شبکه عصبی مصنوعی و ARIMA را به عنوان یک راه‌کار بهینه در پیش‌بینی سری‌های زمانی مطرح می‌کند و مدعی می‌شود که ترکیبی از مدل‌های شبکه ع صبی م صنوعی به عنوان یک مدل غیرخطی و مدل ARIMA به عنوان یک مدل خطی در پیش‌بینی سری‌های زمانی، عملکرد بهتری ن سبت به پیش‌بینی صورت گرفته با استفاده از هر یک از مدل‌ها به تنهایی دارد. او برای اثبات این ادعا با استفاده از چند داده تجربی، ابتدا مجموعه داده را با استفاده از مدل ARIMA مدل‌سازی و خروجی را با خروجی مدل فرضیه خود مقایسه کرده و بر اساس آن، فرضیه خود را به اثبات رسانید (لاتیسا و همکاران، ۲۰۱۰)

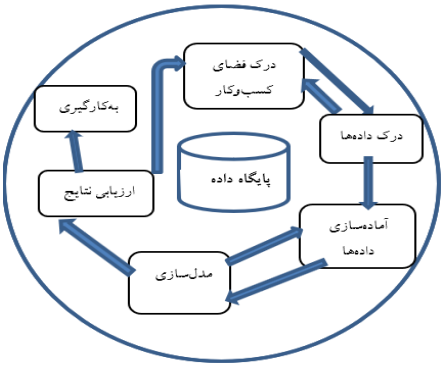
لاتیا، فرولیچ و کاپرا در مقاله‌ای تحت عنوان داده‌کاوی سی ستم حمل‌ونقل عمومی و هوشمند نمودن سیستم‌های حمل‌ونقل، از ابزارها و تکنیک‌های داده‌کاوی استفاده نموده و مدت زمان سفر مسافران برای ر سیدن به مقصد را کاهش داده و حتی الگوهای سفر م سافران را پیش‌بینی نمودند. مطالعه موردی این تحقیق با استفاده از داده‌های موجود در پایگاه داده سیستم حمل‌ونقل متروی شهر لندن صورت گرفته و مدت زمان سفر هر مسافر را تخمین زده و همچنین ایستگاه‌هایی که بیشتر تو سط هر مسافر مورد استفاده قرار می‌گیرد را استخراج نموده است و با ارائه مدلی، مسیری که کمترین مدت زمان سفر را دارد به مسافر پیشنهاد می‌دهد. از اهداف دیگر این مقاله، تعیین ایستگاه‌هایی بود که بیشتر مورد استفاده قرار می‌گیرد زیرا این اطلاعات در جهت بهبود و تغییر برنامه‌های سفر آن ایستگاه موثر است (فیضا، ۲۰۱۸)

آدامز، آکانو و آسموتا در مقاله خود به پیش‌بینی میزان برق مورد نیاز برای تولید در ک شور نیجریه پرداختند. هدف آن بود که با در اختیار دا شتن میزان برق تولیدی در فا صله سال‌های ۱۹۷۰ تا ۲۰۰۹ ، میزان برق مورد نیاز تولیدی این کشور را در ۱۰ سال آینده پیش‌بینی نمایند. علت نیاز به این تحقیق آن بود که در سال ۲۰۰۵ به علت عدم تولید برق کافی و همچنین کاهش میزان بارندگی، تنها ٤۰ درصد برق مورد نیاز این کشور تامین گردید. لذا استفاده از مدلی که بتواند این پیش‌گویی را انجام دهد هدف این گروه قرار گرفت. این گروه در این تحقیق از مدل ARIMA برای پیش‌بینی استفاده کرده و دلیل آن را اینگونه بیان نموده‌اند که مدل ARIMA مدلی عمومی ا ست و توانایی مدل‌سازی داده‌ها و سری‌های با دوره و بدون دوره فصلی را دارا می‌باشد و جهت تعیین درجه این مدل از نمودارهای ACF و PACF ا ستفاده نمودند. این گروه پس از مدل سازی داده‌های ت ست با درجه‌های مختلف اتورگر سیون،

میانگین متحرک و تفاضل‌گیری و ارزیابی خروجی‌ها بر اساس پارامترهای Stationary R2 و Bayesian Information Criteria(BIC)، مدل برتر را انتخاب و بر اساس آن میزان برق مورد نیاز برای تولید در ۱۰ سال بعد را پیش‌بینی نمودند (قصری، ۲۰۱۷)

۳-فرآیند کریسپ

بزرگ‌ترین شرکت الکترونیکی نرم‌افزار و سخت‌افزار آمریکا، به‌عنوان بخشی از هدفش، برای دادن ارزش بیشتر به مشتریان انبار داده خود، گروه‌هایی از مشاوران و متخصصین فنی داده‌کاوی را برای سرویس‌دهی به نیازمندی‌های مشتریانش تأسیس نمود (کیانیان، ۱۳۹۵) فرآیند داده‌کاوی در استاندارد کریسپ داده‌کاوی، به روش کشف دانش که توسط اسامه فیاد و همکارانش ارائه شد، در طول شش مرحله اجرا می‌شود؛ که در ادامه، روش اجرای پژوهش بر اساس این مراحل تشریح خواهد شد (فیاد و همکارانش، ۲۰۰۸)



شکل ۲. مدل فرآیندی کریسپ

۳-۱-درک فضای کسب و کار

شناخت اهداف کسب‌وکار: مرحله اول هر پروژه داده‌کاوی این است که مشخص شود چه هدفی از اجرای پروژه مورد انتظار است.

درک اهداف: اهداف تجاری در محدوده کسب‌وکار و یا حوزه سلامت، مطرح هستند و اهداف داده‌کاوی باید در حوزه داده‌کاوی و تکنیک‌های آن بررسی شوند. مهم‌ترین خروجی در این مرحله لیست اهداف داده‌کاوی است. از طرفی ضوابط موفقیت داده‌کاوی نیز باید تعیین گردد.

تهیه برنامه پروژه: در انتهای مرحله نخست کریسپ باید برنامه مانند هر پژوهش دیگری تنظیم گردد. در این برنامه علاوه بر مشخص کردن جزییات فعالیت‌های پژوهش و زمان اجرای آن، باید روش‌ها و ابزارهای موردنیاز نیز در پروژه انتخاب گردد. خروجی این بخش از فرایند به‌صورت مستندات اجرا و زمان‌بندی پژوهش است. برنامه پژوهش موردنظر در قالب پروپوزال، پیش‌ازاین ارائه شده است.

۳-۲-درک داده‌ها

جمع‌آوری داده‌های اولیه: در این مرحله بر اساس نوع پروژه، داده‌ها را منابع مختلف با پرسشنامه و یا سیستمی جمع‌آوری می‌کنند.

تشریح داده‌ها: تشریح داده‌ها را می‌توان بررسی سطحی و مقدماتی داده‌ها و استخراج مشخصات اولیه مجموعه داده‌ها، تعریف کرد. در این پژوهش در این مرحله فرمت داده، تعداد متغیرها، نوع متغیرها (پیوسته و گسسته) داده و مفهوم متغیرهای موردبررسی قرار می گیرد. هم‌چنین در این بخش، با توجه تعداد رکوردها مشخص گردید، آیا نوع و حجم داده‌ها ملزومات پژوهش را برآورده می‌سازد.

اکتشاف داده: پس از تشریح مقدماتی داده‌ها، باید دید دقیق‌تری به داده‌ها پیدا کرد. در این مرحله باید با استفاده از روش‌های آماری مشخصات اصلی داده‌ها را استخراج و به کمک ابزارهای مصورسازی خصوصیات داده‌ها را به تصویر کشیده شد.

بررسی کیفیت داده‌ها: خروجی این مرحله معمولاً به‌صورت گزارشی هست که در آن اشتباهات داده‌ها، مقادیر ازدست‌رفته، وجود رکوردهای تکراری، مشخص می‌گردد. منظور از اشتباه‌های داده‌ها، مقادیر بی‌معنی و داده دورافتاده و پرت است که در نتایج تأثیر می‌گذارد.

۳-۳-آماده‌سازی داده‌ها

انتخاب داده‌ها: در برخی مسائل به علت حجم بسیار زیاد رکوردهای موجود، از روش‌های نمونه‌گیری به منظور کاهش حجم داده استفاده می‌شوند.

پاک‌سازی داده‌ها: در این بخش باید با استفاده از گزارش خروجی بررسی کیفیت داده‌ها در مرحله قبل، ابتدا مقادیر ازدست‌رفته را جایگزین کرد. هم‌چنین در مرحله دوم باید مقادیر بی‌معنی موجود در متغیرها با توجه به گزارش مفاهیم متغیرها در مرحله قبل، حذف و یا با مقادیر بامعنی جایگزین شوند.

کاهش متغیرها: در این مرحله پس از پاک سازی داده‌ها، باید از روش‌های موجود انتخاب ویژگی جهت یافتن متغیرهای تأثیرگذار و حذف متغیرهای کم تأثیر و بی تأثیر استفاده کرد. این مرحله بسیار کاربردی است و در مواقعی که تعداد متغیرها بالا باشد، زمان اجرا و کارایی مدل‌سازی را پایین می‌آورد، ازاین‌رو مرحله بسیار مهمی است.

٤-مدلسازی

انتخاب روش مدل‌سازی: در این مرحله ابتدا چندین تکنیک دسته‌بندی یا خوشه‌بندی بر اساس نوع داده‌ها انتخاب خواهد شد. با توجه به مرحله پیش‌پردازش ممکن است در این مرحله بر اساس پارامترهای ارزیابی نیاز به بازگشت به عقب داریم و تکنیک‌های مختلف دسته‌بندی و پارامترها را بررسی کنیم.

٤-١-ارزیابی و به کارگیری

ارزیابی نتایج: در این گام می‌توان با توجه حوزه پژوهش و مطالعه مقالات معتبر و پژوهش‌های کار شده روی داده‌ها، شاخص‌های ارزیابی مدل را انتخاب نمود؛ و بهترین روش و دسته‌بند در هر مرحله بر اساس این ویژگی انتخاب شد. در مواردی که هزینه کم باشد و بودجه کافی در اختیار باشد، می‌توان مدل ساخته‌شده را بر روی مجموعه داده‌های واقعی دیگر آزمود و صحت و دقت مدل ساخته‌شده را بیشتر بررسی نمود.

٤-٢-جمع‌آوری داده‌های مدل

مجموعه داده جمع آوری شده شامل ۱۳۸۹ اطلاعات مرتبط به حمل نقل انواع کالا یک شرکت حمل‌ونقل در انواع ریلی و جاده‌ای است.که کالاهای شامل شمش روی، و شمش آلومینیومی را به کشور ترکیه صادر می کند.

داده‌ها شامل ١٤ ستون عددی و گسسته است. ویژگی‌ها به شرح جدول ۱ تعریف شده‌اند.

جدول ۱. داده‌ها و نوع آن‌ها

عنوان	نوع متغیر
تاریخ تقاضا	تاریخ شمسی
تعداد تقاضا	عددی
تناژ متقاضی (به کیلوگرم)	عددی
مبدا	گسسته
مقصد	گسسته
نوع بار	گسسته ZING INGOT: شمش روی CONTAINER: کانتینر ALUMINIUM INGOT: شمش آلومینومی
تعداد عرضه	عددی
تناژ بارگیری (به کیلوگرم)	عددی
زمان حمل تخمینی (به روز)	عددی
زمان حمل از روز تقاضا تا روز تخلیه	عددی
هزینه حمل (به ریال)	عددی
نوع حمل	گسسته دو مقداری حمل ریلی: ۰ حمل کامیونی: ۱
وضعیت آب و هوایی	گسسته دو مقداری معتدل: ۰ نامعتدل : ۱
نیاز به مجوز	گسسته دو مقداری دارد: ۱ ندارد: ۰

٤-٣-پیش‌پردازش: پاکسازی و آماده‌سازی داده‌ها

در این گام پس از تشریح داده با نمودار، در صورت وجود داده از دست رفته، برمبنای تعداد داده‌های میسینگ رویکرد حذف و یا جایگزینی داده از دست رفته انتخاب می‌شود.

هم چنین وجود سطرهای تکراری و متغیرهای اضافه را در نرم‌افزار بررسی می‌کنیم.

٤-٤-خوشه‌بندی

جهت تحلیل داده‌های حمل نقل، گام اول خوشه‌بندی و گروه‌بندی اطلاعات بار است. جهت مدلسازی تحلیل از خوشه‌بندی کامینز k-means استفاده کردیم. ابتدا تعداد پارامترها را به‌طور د ستی وارد کرده و با پارامتر ارزیابی دیویس بولدین ابتدا تعداد بهینه خوشه‌ها را مشخص می‌کنیم.

در مرحله تحلیل، با بهینه ترین تعداد خوشه‌ها را با خروجی جدول مرکزی ارزیابی می‌کنیم.

استخراج دانش قوانین انجمنی

فرآیند کشف قوانین وابستگی، یکی از رویکردهای مهم در علم نوین داده‌کاوی برای یافتن قواعد و الگوها در پایگاه داده است که بسیار موردتوجه تحقیقگران قرارگرفته است. در گام استخراج دانش، هدف تولید قواعد انجمنی از مجموعه داده است تا روابط بین ویژگی‌ها در حوزه حمل‌ونقل استخراج شوند.

فرضیات ابتدایی از این قرارند: مجموعه $I = I_1, I_2, I_3, \ldots$ را به‌عنوان مجموعه اقلام در نظر بگیرید. مجموعه D را به‌عنوان مجموعه داده‌ها یعنی تراکنش‌های پایگاه داده‌ها در نظر بگیرید بطوریکه هر تراکنش شامل مجموعه‌ای از اقلام با شد، یعنی هر تراکنش T زیرمجموعه‌ای از I است ($T \subseteq I$). هر تراکنش شناسه‌ای بنام TID دارد. اگر A یک مجموعه از اقلام باشد، می‌گوییم. تراکنش T شامل A است اگر و فقط اگر $A \subseteq T$ باشد. یک *قانون انجمنی* گزاره‌ای است به‌صورت $A \Rightarrow B$، در حالی که $A \subseteq I$، $B \subseteq I$ و $A \cap B = \emptyset$. معیار مقبولیت قوانین انجمنی دو پارامتر مهم یعنی پشتیبانی و قابلیت اطمینان قوانین را معرفی می‌کنیم.

قانون $A \Rightarrow B$ در مجموعه تراکنش‌های D دارای پشتیبانی برابر با s است اگر s در صد از تراکنش‌های D شامل $A \cup B$ باشند و این قانون دارای قابلیت اطمینان برابر c است اگر c درصد از تراکنش‌هایی که شامل A هستند شامل B نیز باشند.

یا به قول دیگر:

$$Support(A \Rightarrow B) = P(A \cup B)$$
$$Confidence(A \Rightarrow B) = P(B|A)$$

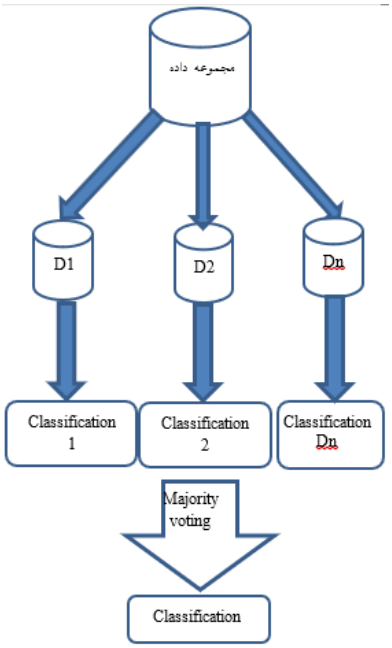
قوانینی که حد پایین Support یا min-sup و حد پایین Confidence یا min-conf را دارا باشند قوانین انجمنی *قوی* نامیده می‌شوند و هدف تمام الگوریتم‌ها چنین قوانینی است (یا چنانکه خواهیم دید در بعضی موارد زیرمجموعه‌ای از این قوانین). مثلاً قانون زیر بیانگر این ا ست که ۷۰ در صد ک سانی که تلویزیون خر یداری می‌کن‌ند دست گاه پخش و یدئو نیز می‌خرند و ۸۰ درصد از کل تراکنشات شامل هر دو قلم تلویزیون و دستگاه پخش ویدئو است.

$$Television \implies VCR$$

$$[Support = 80\%\ Confidence = 70\%$$

استخراج دانش (الگوریتم‌های دسته‌بندی)

در گام دسته‌بندی پس از گروه‌بندی داده‌های بار و تحلیل خوشه‌ها، از الگوریتم‌های دسته‌بندی درختی جمعی مانند جنگل تصادفی و روش احتمالی بیزساده استفاده کنیم و نتایج و قوانین خروجی را با هم مقایسه کنیم.



شکل ۲. مدل الگوریتم‌های تجمعی

جهت اجرای د سته‌بندها باید داده‌ها را به دو د سته آموز شی و آزمون تقسیم کنیم (شکل ۲). با استفاده از روش اعتبارسنجی ضربدری مجموعه داده‌ها را به دودسته آموزشی و آزمون تقسیم می‌کنیم. درروش اعتبارسنجی ساده در تمام مراحل اجرا می‌توان مثلاً ۷۰ در صد مجموعه داده آموز شی و مابقی را برای آزمایش و محاسبه شاخص‌ها بکار گرفت؛ اما درروش هوشمندانه‌تر برای اینکه کل مجموعه داده یک‌بار هم در آموزش و هم در آزمایش با شد، از روش اعتبار سنجی ضربدری ا ستفاده می‌کنیم که در این پژوهش نیز استفاده شد.

در این نوع اعتبارسنجی داده‌ها به K زیرمجموعه افراز می‌شوند؛ از این K زیرمجموعه، هر بار یکی برای اعتبارسنجی و K-1 تای دیگر برای آموزش بکار می‌رو ند. این روال K بار تکرار می‌شود و همه داده‌ها دقیقاً یک‌بار برای آموزش و یک‌بار برای اعتبارسنجی بکار می‌روند. درنهایت میانگین نتیجه این K بار اعتبارسنجی، به‌عنوان یک تخمین نهایی برگزیده می‌شود.

معیارهای ارزیابی الگوریتم‌های طبقه‌بندی

این ماتریس چگونگی عملکرد الگوریتم طبقه‌بندی را با توجه به مجمو عه داده ورودی به تفکیک انواع رده های مسئله طبقه‌بندی نشان می‌دهد. با توجه به جدول ۲، مفاهیم مثبت درست، مثبت غلط، منفی درست و منفی غلط به شرح ذیل است.

جدول ۲. ماتریس درهم‌ریختگی

		رکوردهای تخمینی	
		رده منفی (–)	رده مثبت (+)
رکوردهای واقعی	رده منفی (–)	TN	FP
	رده مثبت (+)	FN	TP

-مثبت درســـت: این مقدار بیانگر تعداد رکوردهایی اســـت که رده واقعی آن‌ها مثبت بوده و دســـته‌بند نیز رده آن‌ها را به درســـتی مثبت تشخیص داده است.

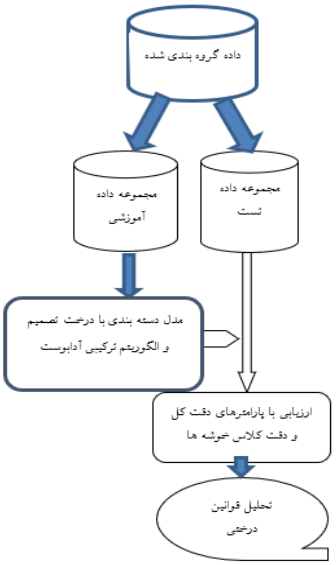
-مثبت غلط: این مقدار بیانگر تعداد رکوردهایی ا ست که رده واقعی آن‌ها منفی بوده و د سته‌بند رده آن‌ها را به نادر ستی مثبت ت شخیص داده است.

-منفی درست: این مقدار بیانگر تعداد رکوردهایی است که رده واقعی آن‌ها منفی بوده و دسته‌بند نیز رده آن‌ها را به‌درستی منفی تشخیص داده است.

-منفی غلط: این مقدار بیانگر تعداد رکوردهایی اســت که رده واقعی آن‌ها مثبت بوده و دســته‌بند رده آن‌ها را به نادرســتی منفی تشــخیص داده است.

اکنون به تشـــریح انواع مهم معیارهای ارزیابی الگوریتم‌های طبقه‌بندی می‌پردازیم. مهم‌ترین معیار، دقت یا نرخ طبقه‌بندی اســـت و به این مفهوم ا ست که طبقه بند، چند درصد رکوردهای آزمایشی را به‌درستی طبقه‌بندی کرده است. این دقت بر اساس مفاهیم ماتریس با توجه به رابطه قابل‌محاسبه می‌شود.

$$Accuracy = \frac{Tp+TN}{Tp+TN+Fp+FN} \quad (۲)$$



شکل ۳. مدل دسته‌بندی (استخراج دانش)

ارزیابی

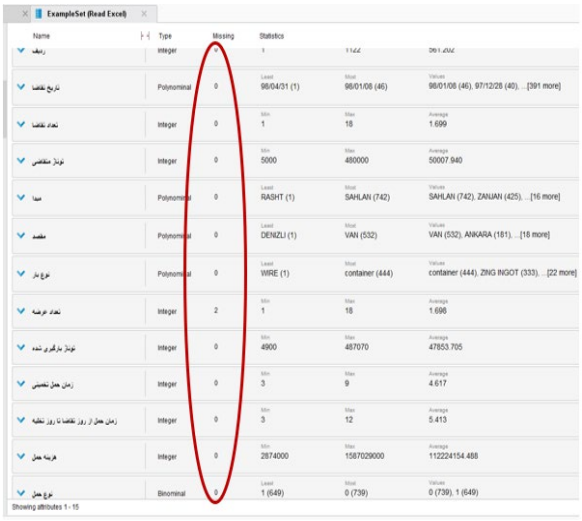
در ادامه بر اســـاس تعریف ماتریس درهم‌ریختگی و پارامترهای ارزیابی دقت کل، نتایج را از خروجی نرم‌افزار گزارش داده و ســـپس تحلیل می‌کنیم. در نهایت با نتیجه‌گیری و تحلیل خوشه‌بندی، استراتژی‌های حوزه حمل‌ونقل را گزارش خواهیم داد.

نتایج تحقیق

در این بخش بر اساس مدل فرآیند کریسپ، مراحل درک داده، مصورسازی داده، آماده سازی و پاکسازی داده را با اجرا در نرم افزار تشریح می‌کنیم. سپس خوشه بندی و دسته بندی نتایج را در قالب جدول گزارش می دهیم. نتایج تحلیل خوشـــه‌های بهینه، جدول مرکزی خوشه ها را در انتها تشریح می‌کنیم. سپس در بخش دسته بندی دو روش جنگل تصادفی، و بیزساده روی مجموعه داده خو شه بندی شده اعمال کرده و دقت مدل ها را گزارش داده و سپس ارزیابی می‌نماییم.

درک داده‌ها

اطلاعات آماری مجموعه داده با توجه به شکل ٤ به شرح زیر می‌باشد.



شکل ٤. اطلاعات آماری داده

با توجه به شـــکل ٤ در بخش اطلاعات آماری ســـتون اول نام داده، ستون دوم نوع داده، ستون سوم تعداد داده‌های از د ست رفته هر متغیر، ستون چهارم و پنجم ماکزییم و مینیمم متغیرهای عددی و کمترین و بی‌شترین مقدار متغیرهای گ سسته و ستون آخر مقدار میانگین و انحراف معیار متغیرهای عددی و مقدار مد متغیرهای گسسته را نشان می‌دهد.

تنها در ستون عرضه ۲ سطر داده از دسته رفته وجود دارد. متغیر تاریخ تقا ضا از نوع تاریخ شم سی ا ست که در نرم افزار از نوع گسسته در نظر گرفته می شود.

تعداد عرضه و تقاضا از رنج ۱ تا ۱۸ و میانگین ۱,٦۹۸ مشخص شـــده اســـت. تناژ بارگیری از ٤۹۰۰ کیلوگرم تا ٤۸۷۰۷۰ کیلو تعریف شده است.

Index	Nominal value	Absolute count	Fraction
1	SAHLAN	742	0.535
2	ZANJAN	425	0.306
3	QAZVIN	104	0.075
4	TABRIZ	69	0.050
5	ISFAHAN	13	0.009
6	YAZD	13	0.009
7	QOM	4	0.003
8	TEHRAN	4	0.003

Index	Nominal value	Absolute count	Fraction
11	SALAFCHEGAN	2	0.001
12	ARAK	1	0.001
13	BANDAR EMAM	1	0.001
14	ESFAHAN	1	0.001
15	ESHTEHARD	1	0.001
16	HAMEDAN	1	0.001
17	KASHAN	1	0.001
18	RASHT	1	0.001

شکل ۵. فراوانی مبدا شهرها در حمل و نقل

با توجه به شـــکل ۵ ، ۵۳ درصـــد حمل و نقل از واحد مبدا از شهر سهلان، ۳۰ درصد از زنجان و مابقی با درصد بسیار کمتر در شهرهای قزوین، اصفهان، یزد، قم، تهران، ایلام، سلفچگان و اراک و ... تعریف شده است.

Index	Nominal value	Absolute count	Fraction
1	VAN	532	0.383
2	ANKARA	181	0.130
3	MERSIN	141	0.102
4	GAZIANTEP	94	0.068
5	MURATBEY	93	0.067
6	KONYA	80	0.058
7	ADANA	75	0.054
8	ISKENDERUN	53	0.038

Index	Nominal value	Absolute count	Fraction
10	DERINCE	33	0.024
11	IZMIR	18	0.013
12	KAYSERI	18	0.013
13	BURSA	6	0.004
14	ERENKOY	5	0.004
15	GEBZE	5	0.004
16	HATAY	4	0.003

شکل ٦. فراوانی مقصد شهرها در حمل و نقل

با توجه به شکل ٦، ۳۸ درصد مقصد بارها به شهر وان ترکیه، ۱۳ درصد آنکارا، ۱۰ درصد مرسین ترکیه، ۰,۰٦ درصد Gaziantep، ۰,۰٦ درصد Muratbey، ۰,۰۵ درصد Konya و مابقی شهرهای Adana، Izmir و سایر شهرهای مرزی ترکیه است.

-زمان حمل تخمین زده شده، از مینیمم ۳ روز تا ۹ روز و میانگین ٤,٦١ تعریف شده است.

-زمان حمل از روز تقاضا تا روز تخلیه از مینیمم ۳ روز تا ۱۲ روز تعریف شده است.

-وضعیت آب و هوایی در ۲۳۲ سطر در حالت نامعتدل و ۱۱٥٦ سطر در حالت معتدل تعریف شده است.

-٤٦ درصد حمل و نقل‌ها در حالت کامیونی و ۵۳ درصد در حالت ریلی هستند.

-۹٦ درصد بارها در حمل و نقل نیاز به مجوز ندارند.

-اطلاعات از سال ۹٦ تا ۹۸ تعریف شده است.

-ماکزیمم رنج تناژ متقاضی و تناژ بارگیری شده، برای حمل کامیونی تا میانگین ٤۷٥۰۰۰ تعریف شده و برای حمل ریلی تنها از رنج ۲٥۰۰۰ تا ٥۰۰۰۰ کیلوگرم تعریف شده است.

-برای حمل ریلی تنها شهر سهلان مبدا حمل و نقل بوده و سایر بارها در حمل کامیونی در سایر شهرها تعریف شده‌اند.

-در حمل ریلی تنها شهرهای وان، قاضی انتپ، مرسین و اسکندرون مقصد حمل و نقل و مابقی مربوط به حمل کامیونی هستند.

-بیشترین زمان حمل از روز تقاضا تا روز تخلیه، در فاصله زمانی سه روز تا ۱۲ روز برای حمل کامیونی و برای حمل ریلی تا میانگین ۵ روز و نهایت ۸ روز فاصله زمانی داریم.

-بیشترین هزینه حمل و نقل مربوط به حمل کامیونی تا ۱,٥۰۰,۰۰۰,۰۰۰ ریال است و تفاوت هزینه در حمل ریلی و کامیونی مشهود است.

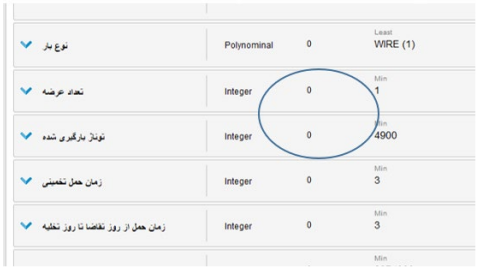
پیش پردازش و آماده سازی داده‌ها

متغیرهای اضافی

در بخش تحلیل داده های حمل و نقل، گام اول در بخش خوشه‌بندی نیاز به متغیرهای توصیفی عددی و پیوسته داریم. به دلیل وجود متغیرهای زمان تخمینی و فاصله زمانی تاریخ حمل تا تخلیه، وجود ویژگی تاریخ تقاضا که از نوع شمسی است در مدلسازی و استخراج دانش اضافی و از داده کنار گذاشته می‌شود.

حذف مقدار ازدست‌رفته

همان‌طور که در بخش تشریح داده اشاره کردیم تنها در ستون تعداد عرضه ۲ سطر داده از دست رفته داریم، یکی از راه‌های حذف سطرهای با داده میسینگ است که ایراد این روش حذف اطلاعات مفید در ویژگی‌های دیگر است که دارای مقدار است. یکی از بهترین روش‌ها جایگزینی مقدار از دست رفته با مقدار میانگین همان ستون است تا انحراف معیار داده حفظ شود. از این رو در نرم‌افزار برای متغیر عددی عرضه از جایگزین داده از دست رفته استفاده می‌کنیم.



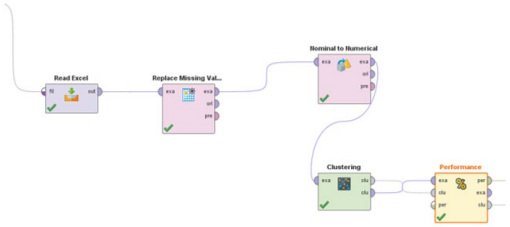
شکل ۷. جایگزینی داده از دست رفته

مدلسازی(خوشه بندی)

در این گام الگوریتم خوشـــه بندی کامیانگین را به داده‌های پاکسازی شده اعمال کرده و در هر مرحله تعداد k (خو شه‌ها) شبیه سازی و براساس پارامتر دیویس بولدین ارزیابی می‌نماییم و براســـاس مقدار ارزیابی و در بخش تحلیل نتایج بهینه‌ترین تعداد خوشه را ارزیابی می‌کنیم.

خوشه‌بندی کامیانگین

جهت خوشـــه‌بندی باید داده‌ها را به مقدار عددی تبدیل کنیم و مدل به شرح شکل ۸ است.



شکل ۸. مدل خوشه‌بندی

نتایج پارامتر ارزیابی دیویس بولدین را به ازای تعداد خوشه‌های ۳ و ٤ گزارش می‌دهیم و در بخش مقایسه تحلیل می‌کنیم.

جدول ۳. مقایسه دو روش خوشه‌بندی

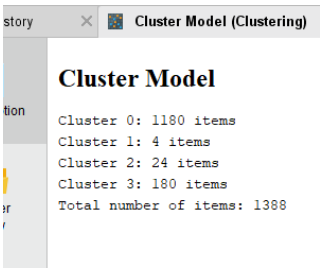
Medel(K)	Davis Bouldin
Kemans(k=3)	0.007
Kmeans(k=4)	0.005

ارزیاب دیوس بولدین نسبت فا صله درون خو شه‌ای به فا صله بین خوشـــه‌ها را گزارش می‌دهد و عددی بین صـــفر و یک را خروجی می‌دهد در این صورت مدلی موفق ا ست که نمونه‌ها در یک خوشـــه به یکدیگر نزدیک‌تر و متراکم‌تر و از نمونه‌های خوشـــه‌های دیگر دورتر باشـند. یعنی هر چه دیویس بولدین به صفر نزدیکتر باشد مدل موفق‌تر و تعداد خوشه‌ها بهینه‌تر است. با توجه به جدول ٤ کمترین مقدار ارزیاب با تعداد ٤ خوشـــه و الگوریتم کامینز است. از این رو در بخش بعد نتایج خوشه‌بندی 4-means را برای گروه‌بندی، گزارش می‌دهیم.

تحلیل خوشه‌بندی 4-means

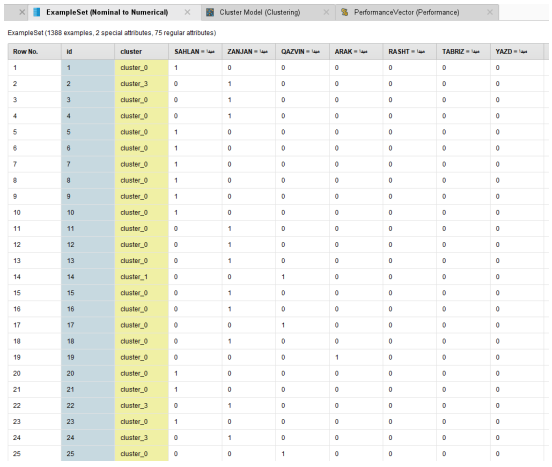
در این مرحله با توجه به نتایج نرم‌افزار به تحلیل خوشـــه‌ها، تعداد نمونه‌ها در هر خوشه می‌پردازیم.

با توجه به شکل ۸ تعداد نمونه‌ها به تفکیک خو شه‌ها مشخص است. در خوشه صفر ۱۱۸۰ آیتم یعنی بیشتر نمونه‌ها، در خوشه ۱ تنها ٤ آیتم، در خوشـــه دو ۲٤ آیتم ودرخوشـــه ســـه ۱۸۰ آیتم وجود دارد.



شکل ۸. تفکیک نمونه خوشه

همچنین خروجی زیر (شکل ۹) مشخص می‌کند که هر تراکنش حمل بار در چه خوشه‌ای قرار دارد.



Row No.	id	cluster	SAHLAN = صفا	ZANJAN = صفا	QAZVIN = صفا	ARAK = صفا	RASHT = صفا	TABRIZ = صفا	YAZD = صفا
1	1	cluster_0	1	0	0	0	0	0	0
2	2	cluster_3	0	1	0	0	0	0	0
3	3	cluster_0	0	1	0	0	0	0	0
4	4	cluster_0	0	1	0	0	0	0	0
5	5	cluster_0	1	0	0	0	0	0	0
6	6	cluster_0	1	0	0	0	0	0	0
7	7	cluster_0	1	0	0	0	0	0	0
8	8	cluster_0	1	0	0	0	0	0	0
9	9	cluster_0	1	0	0	0	0	0	0
10	10	cluster_0	1	0	0	0	0	0	0
11	11	cluster_0	0	1	0	0	0	0	0
12	12	cluster_0	0	1	0	0	0	0	0
13	13	cluster_0	0	1	0	0	0	0	0
14	14	cluster_1	0	0	1	0	0	0	0
15	15	cluster_0	0	1	0	0	0	0	0
16	16	cluster_0	0	1	0	0	0	0	0
17	17	cluster_0	0	0	1	0	0	0	0
18	18	cluster_0	0	1	0	0	0	0	0
19	19	cluster_0	0	0	0	1	0	0	0
20	20	cluster_0	1	0	0	0	0	0	0
21	21	cluster_0	1	0	0	0	0	0	0
22	22	cluster_3	0	1	0	0	0	0	0
23	23	cluster_0	1	0	0	0	0	0	0
24	24	cluster_3	0	1	0	0	0	0	0
25	25	cluster_0	0	0	1	0	0	0	0

شکل ۹. خوشه‌بندی نمونه‌ها

این جدول با روش 4-means به‌صورت جدول ٤ می‌باشد.

جدول ٤. جدول مرکزی خوشه‌ها بر اساس شاخص‌های عددی و کیفی حمل‌ونقل

شاخص های حمل و نقل	امتیاز cluster0	امتیاز cluster1	امتیاز cluster2	امتیاز cluster3
هزینه حمل	۷۲۰٤۸٤٦۳,۲	۱٤۷۱٤٤۹۷۵۰	٦٦۷۱۳۱۸٤۷,۵	۲۷۲۲۷۷۱۳۱,۵
زمان حمل از روز تقاضا تا روز تخلیه	۵,۲۳۱	٦,۷۵۰	٦,۳۵۰	٦,٦٦۱
زمان حمل تخمینی	٤,٤٤۵	٦,۲۵۰	۵,٦۰۰	۵,۲۷۱
تناژ بارگیری	۳۷۷۵۱	۱۹۵٦۲۷	۱٤۸۲۷۸,۱	۸۲۳۳۸,۲
تعداد عرضه	۱,۲۵٤	۸,۲۵	٦,۱۵۰	۳,۳۲۷
تناژ متقاضی	٤۰۹۹۷,۵	۱۹۸۰۰۰	۱٤۷۰۷۵	۷۹۸۳۳,۳
تعداد تقاضا	۱,۲۵٦	۸,۲۵۰	٦,۱۵۰	۳,۳۲۱
نیاز به مجوز: ندارد	۰,۹۷۰	۱	۰,۹۵۰	۰,۹۳۹
نیاز به مجوز: دارد	۰,۰۳۰	۰	۰,۰۵۰	۰,۰٦۱
وضعیت آب و هوایی: معتدل	۰,۸٤٤	۰,۷۵۰	۰,٦۰۰	۰,۷۷۰
وضعیت آب و هوایی: نامعتدل	۰,۱۵٦	۰,۲۵۰	۰,٤۰۰	۰,۲۳۰
نوع حمل: ریلی	۰,٦۲۵	۰	۰	۰,۰۰٦
نوع حمل: کامیونی	۰,۳۷۵	۱	۱	۰,۹۹٤

خوشه دو (۱,۷۲ درصد): بعد از خوشه ۱ هزینه های نسبتا بالای حمل، مربوط به بارهای این گروه است. بیشترین تناژ بارگیری بعد از خوشه ۱، و بیشترین تعداد عرضه و تقاضا با میانگین ۶,۱ در مورد آیتم های این گروه است. ۰,۰۵ درصد بارها برای حمل و نقل در این گروه نیاز به مجوز دارند و ۴۰ درصد حمل و نقل ها در وضعیت آب و هوایی نامعتدل است. درصورتیکه در خوشه های دیگر بیشتر جمل و نقل‌ها در وضعیت معتدل بوده است. ۴۱ درصد بارها از شهر رشت، ۲۹ در صد قزوین و ۲۰ در صد قزوین ارسال شده و مقصد با درصد ۳۳ درصد به وان و مابقی آدانا و آنکارا بوده است. ۱۰۰ درصد بارها از نوع کامیونی است.

خوشه ۳ (۱۳,۰۷ درصد): هزینه های حمل و نقل در این گروه متوسط، از خوشه ۱ و ۲ کمتر از خوشه صفر بیشتر است. تفاوت شاخص این گروه نسبت به گروه های دیگر در تفاوت فاصله زمانی نسبت زمان حمل تخمینی تا زمان حمل از روز تقاضا و روز تخلیه است که نسبت به گروه های دیگر بیشتر است. تناژ بارگیری، تعداد عرضه و تقاضا در این خوشه متوسط و ۰,۰۶ در صد بارها نیاز به مجوز دارد. ۲۳ در صد در وضعیت آب و هوای نامعتدل بوده است. در ۰,۰۶ بارها از نوع ریلی است. ۵۶ درصد بارها از زنجان و به طور میانگین به تمامی شهرهای ترکیه ارسال شده است.

متغیرهای تاثیرگذار در تحلیل بارها

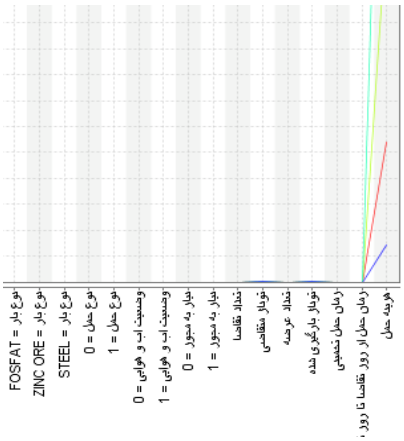
در بحث تحلیل و گروه بندی خوشه ها، با نمودار deviation (شکل ۱۰) فاکتورهای تاثیر گذار در طبقه بندی بارهای در سیستم حمل و نقل را بررسی می‌کنیم.

با توجه به جدول ٤ تحلیل هر خوشه به شرح زیر است.

خوشه ۱ (۰,۲۸درصد): بارهای حمل و نقل شده در این گروه از نظر ویژگی حمل پرهزینه ترین بارها نسبت به خوشه های دیگر (تا مبلغ ۱ میلیارد بیشتر)، است و میانگین تناژ بارگیری بیشترین مقدار، هم چنین بیشترین تعداد عرضه و تقاضا در این گروه است. ۱۰۰ در صد بارهای حمل و نقلی این گروه نیاز به مجوز ندارد و نوع حمل تمام تراکنش‌ها ۱۰۰ درصد از نوع کامیونی است و حمل ریلی را شامل نمیشود. مبدا تمامی بارها از شهر قزوین و ۷۵ در صد به شهر آدانا و ۲۵ در صد به شهر MURATBAY است. نوع بارها در ۷۵ در صد از نوع میلگرد و ۲۵ درصد از نوع شمش روی است. ماکزیمم تایم زمانی حمل از روزتقاضا تا تخلیه در بارهای این خوشه است.

تنها ٤ آیتم بار با ویژگی های خیلی متفاوت و ویژه در این گروه قرار دارند.

خوشه صفر(فراوانی ۸۵,۰۱ درصد): بیشترین بارها (۱۱۸۰آیتم) در این گروه هستند که به طور میانگین کم هزینه ترین بارها نسبت به خوشه های دیگر، کمترین زمان حمل از روز تقاضا تا روز تخلیه در مورد بارهای این گروه است. ۹۹ در صد بارهای ریلی در این خوشه هستند و در گروه های دیگر به جز درصد بسیار کم در خوشه ۳، بار از نوع ریلی نداریم. اما ۳۷ در صد نیز بار کامیونی داریم. با تفاوت زیادی کمترین مقدار عرضه و تقاضا در این خوشه است. مبدا شهرها ٦٢ درصد از سهلان و ٢٦ درصد از زنجان و مابقی با درصد کم از قزوین، تبریز و یزد است و تقریبا به تمامی شهرها مقصد داریم اما بیشترین مکان خالی شده بار در شهر وان ترکیه (٤٢ درصد) است. ۳۷ درصد نوع بارها کانتینر و ۲۲ درصد از نوع شمش روی است.



شکل ۱۰. نمودار deviation خوشه‌ها

در این شـــکل برای هر خوشـــه خطی بارنگ مجزا در نظر گرفته‌شده است؛ که میانگین امتیازات داده را برای هر ستون در خوشــه مربوطه مشــخص می‌کند. پایین بودن امتیاز خوشــه آبی (خوشه صفر) کم هزینه ترین بارها مشخص است.

-خط قرمز مربوط به خوشه ۳ (بارهای با هزینه متوسط و تعداد عرضه و تقاضای متوسط) است.

-خط سبز مربوط به خوشه (بارهای ۲ با هزینه نسبتا بالا است.

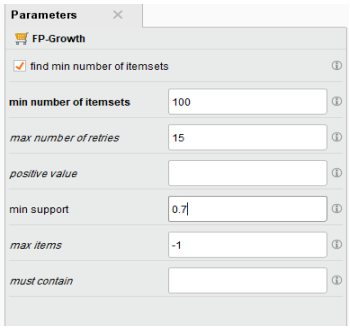
-خط نیلی رنگ بالاترین امتیاز مربوط به خوشـــه ۱(بارهای با هزینه بالا و تناژ بالا و ...) است.

اما تحلیل اصلی این نمودار در جداسازی خطوط از بخش زمان حمل و هزینه حمل و تا حدودی تناژ بار ا ست که نشان دهنده این ا ست که این سه فاکتور نسبت به فاکتورهای وضعیت هوا، نوع بار و مبدا و مقصـــد تاثیرگذاری بیشـــتری در جداســـازی خوشه‌ها دارند.

مدلسازی (قواعد انجمنی)

جهـت تولیـد قواعـد انجمنی و اقلام مکرر، بـا الگوریتم fp-growth در رپیدماینر، فیچرهای گسسته و باینری را برای تولید قوانین مکرر و آیتم هایی که با هم تکرار می‌شـــوند را انتخاب و نتایج به شرح زیر است.

Size	Support	Item 1	Item 2	Item 3
1	0.535	SAHLAN = مبدا		
1	0.468	نوع حمل		
1	0.383	VAN = مقصد		
1	0.320	container = نوع بار		
1	0.306	ZANJAN = مبدا		
1	0.240	ZING INGOT = نوع بار		
1	0.167	وضعیت آب و هوایی		
2	0.343	SAHLAN = مبدا	VAN = مقصد	
2	0.320	SAHLAN = مبدا	container = نوع بار	
2	0.306	نوع حمل	ZANJAN = مبدا	
2	0.240	نوع حمل	ZING INGOT = نوع بار	
2	0.320	VAN = مقصد	container = نوع بار	
2	0.193	ZANJAN = مبدا	ZING INGOT = نوع بار	
3	0.320	SAHLAN = مبدا	VAN = مقصد	container = نوع بار
3	0.193	نوع حمل	ZANJAN = مبدا	ZING INGOT = نوع بار



شکل ۱۱. اقلام مکرر حمل و نقل

شکل ۱۱ بخشی از آیتم‌های پرتکرار را مشخص می‌کند که براساس مقدار support=۷۰٪ پرتکرارترین آیتم‌ها، به شرح زیر است.

۱.۵۳ درصد سیستم حمل و نقل مبدا از شهر سهلان و ۳۸ درصد ا=مقصد بارها شهر وان بوده است.

۲.٤٦,٥ درصد بارها از نوع حمل کامیونی است.

۳.۳۲ درصد سیستم حمل نوع بار کانتینر و ٢٤ درصد شمش روی بوده است.

٤.در ٣٤ درصد حمل و نقل، مبدا از سهلان و مقصد به شهر وان بوده است.

٥.در ۳۲ درصد حمل و نقل با مبدا از شهر سهلان، نوع بار کانتینر است.

٦.در ۳۰ درصد نوع حمل کامیونی، بار از نوع شمش روی بوده است.

۷.در ۳۰ درصد نوع حمل کامیونی، مبدا از شهر زنجان است.

۸.در ۳۲ مقصد به شهر وان، بار از نوع کانتینر است.

۹.در ۳۲ درصد مبدا سهلان، مقصد شهر وان و نوع بار کانتینر است.

با confidence=70% قوانین انجمنی؛ اگر آنگاه؛ به شرح زیر است.

No.	Premises	Conclusion	Support	Confidence
4	VAN = مقصد	container = نوع بار	0.320	0.835
5	VAN = مقصد	container = نوع بار ,SAHLAN = مبدا	0.320	0.835
6	VAN = مقصد	SAHLAN = مبدا	0.343	0.895
7	VAN = مقصد ,SAHLAN = مبدا	container = نوع بار	0.320	0.933
8	container = نوع بار	SAHLAN = مبدا	0.320	1
9	ZANJAN = مبدا	نوع حمل	0.306	1
10	ZING INGOT = نوع بار	نوع حمل	0.240	1
11	container = نوع بار	VAN = مقصد	0.320	1
12	container = نوع بار	VAN = مقصد ,SAHLAN = مبدا	0.320	1
13	container = نوع بار ,SAHLAN = مبدا	VAN = مقصد	0.320	1
14	container = نوع بار ,VAN = مقصد	SAHLAN = مبدا	0.320	1
15	ZING INGOT = نوع بار ,ZANJAN = مبدا	نوع حمل	0.193	1

شکل۱۲. قواعد انجمنی

۱.با اطمینان ۸۳,۵ درصد اگر مقصد بار شهر وان است، آنگاه نوع بار از نوع کانتینر است.
۲.با اطمینان ۸۳,۵ درصد اگر مقصد بار شهر وان است، آنگاه نوع بار از نوع کانتینر و مبدا از شهر سهلان است.
۳.با اطمینان ۸۹,۵ درصد اگر مقصد بار شهر وان است ، آنگاه مبدا از شهر سهلان است.
٤.با اطمینان ۹۳,۳ درصد، اگر مقصد بار شهر وان و مبدا از شهر ، نوع بار کانتینر است.
۵.با اطمینان ۱۰۰ درصد، اگر مبدا شهر زنجان است ، نوع حمل کامیونی است.
٦.با اطمینان ۱۰۰ درصد، اگر نوع بار شمش روی باشد، نوع حمل کامیونی است.
۷.با اطمینان ۱۰۰ درصد، اگر نوع بار کانتینر باشد، مقصد شهر وان است.
۸.با اطمینان ۱۰۰ درصد، اگر نوع بار کانتینر باشد، مقصد شهر وان و مبدا شهر سهلان است.
۹.با اطمینان ۱۰۰ درصد، اگر نوع بار شمش روی باشد، مبدا زنجان و نوع حمل کامیونی است.

مدلسازی (دسته‌بندی: استخراج دانش)

در این بخش دو دسته بند روش احتمالی بیزساده و روش جمعی درختی جنگل تصادفی به مجموعه داده با برچسب خوشه‌ها اعمال می‌کنیم. و نتایج هر دسته بند را گزارش می‌دهیم.

نتایج بیزساده

مجموعه داده (خروجی خوشه بند) را با روش اعتبارسنجی ضربدری در ده مرحله به داده‌های آموزشی و تست تقسیم می‌کنیم و متغیر برچسب (هدف) را cluster در نظر می‌گیریم که یک متغیر گسسته چند مقداری است. (Nominal).

بعد از تقسیم نمونه‌ها از بیزساده برای دسته بندی استفاده می‌کنیم.

در سمت راست نرم‌افزار پارامترهای دقت که در خروجی مشخص می‌شود را مشخص می‌کنیم. پارامترهای دقت کل (accuracy)، Recall پس از اجرا، نتایج به شرح زیر است.

Accuracy: 89.77%

نتیجه ماتریس درهم‌ریختگی به‌صورت زیر است.

accuracy: 89.77% +/- 3.00% (mikro: 89.77%)

	true cluster_0	true cluster_3	true cluster_1	true cluster_2	class precision
pred. cluster_0	1076	6	0	0	99.45%
pred. cluster_3	94	147	0	2	60.49%
pred. cluster_1	0	0	2	1	66.67%
pred. cluster_2	10	27	2	21	35.00%
class recall	91.19%	81.67%	50.00%	87.50%	

شکل ۱۳. نتایج دقت بیزساده و ماتریس درهم‌ریختگی

-دقت کلاس خوشه صفر : ۹۱,۱۹ درصد

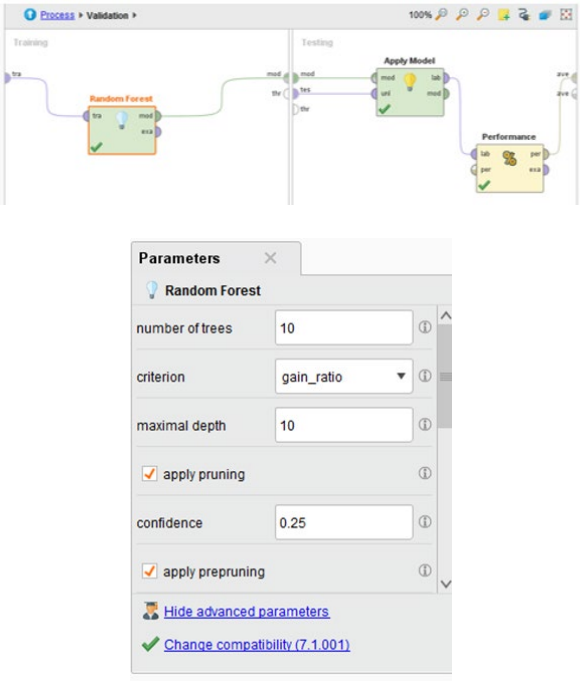
-دقت کلاس خوشه ۳: ۸۱,٦۷ درصد

-دقت کلاس خوشه ۲: ۵۰ درصد

-دقت کلاس خوشه ۱: ۸۷,۵۰ درصد.

نتایج جنگل تصادفی

جنگل تصادفی درخت تصمیم‌های زیادی را تولید می‌کند. برای طبقه‌بندی یک شـــیء جدید از برداری ورودی را در انتهای هر یک از درختان جنگل تصـــادفی قرار می‌دهد. هر درخت به ما یک طبقه‌بندی می‌دهد و می‌گوییم این درخت به آن کلاس «رای» می‌دهد. جنگل طبقه‌بندی ای که بیشـترین رای را داشـته باشد (بین همه درخت‌های جنگل) انتخاب می‌کند.



شکل ۱٤. عملگر جنگل تصادفی

الف) دقت خوشه ۰: ۹۹,۷۵ درصد

-ازاین‌رو تعداد نمونه‌هایی که واقعاً در خوشه صفر هستند و به‌درستی پیش‌بینی‌شده‌اند: ۱۱۷۷ نمونه

-ازاین‌رو تعداد نمونه‌هایی که واقعاً در خوشه ۰ هستند و به‌اشتباه در خوشه ۱ پیش‌بینی‌شده‌اند: ۰ نمونه
-ازاین‌رو تعداد نمونه‌هایی که واقعاً در خوشه ۰ هستند و به‌اشتباه در خوشه ۲ پیش‌بینی‌شده‌اند: ۰ نمونه
-ازاین‌رو تعداد نمونه‌هایی که واقعاً در خوشه ۰ هستند و به‌اشتباه در خوشه ۳ پیش‌بینی‌شده‌اند: ۳ نمونه

ب)دقت خوشه ۳: ۹۱,٦۷ درصد

-ازاین‌رو تعداد نمونه‌هایی که واقعاً در خوشه ۳ هستند و به‌درستی پیش‌بینی‌شده‌اند: ۱٦۵ نمونه
-ازاین‌رو تعداد نمونه‌هایی که واقعاً در خوشه ۳ هستند و به‌اشتباه در خوشه ۰ پیش‌بینی‌شده‌اند: ۱۵ نمونه
-ازاین‌رو تعداد نمونه‌هایی که واقعاً در خوشه ۳ هستند و به‌اشتباه در خوشه ۲ پیش‌بینی‌شده‌اند: ۰ نمونه
-ازاین‌رو تعداد نمونه‌هایی که واقعاً در خوشه ۳ هستند و به‌اشتباه در خوشه ۱ پیش‌بینی‌شده‌اند: ۰ نمونه

پ)دقت خوشه ۱: ۱۰۰ درصد

-ازاین‌رو تعداد نمونه‌هایی که واقعاً در خوشه ۱ هستند و به‌درستی پیش‌بینی‌شده‌اند: ٤ نمونه
-ازاین‌رو تعداد نمونه‌هایی که واقعاً در خوشه ۰ هستند و به‌اشتباه در خوشه ۱ پیش‌بینی‌شده‌اند: ۰ نمونه
-ازاین‌رو تعداد نمونه‌هایی که واقعاً در خوشه ۰ هستند و به‌اشتباه در خوشه ۲ پیش‌بینی‌شده‌اند: ۰ نمونه
-ازاین‌رو تعداد نمونه‌هایی که واقعاً در خوشه ۰ هستند و به‌اشتباه در خوشه ۳ پیش‌بینی‌شده‌اند: ۰ نمونه

ج)دقت خوشه ۲: ۵٤,۱۷ درصد

-ازاین‌رو تعداد نمونه‌هایی که واقعاً در خوشه ۲ هستند و به‌درستی پیش‌بینی‌شده‌اند: ۱۳نمونه
-ازاین‌رو تعداد نمونه‌هایی که واقعاً در خوشه ۲ هستند و به‌اشتباه در خوشه ۱ پیش‌بینی‌شده‌اند: ۰ نمونه
-ازاین‌رو تعداد نمونه‌هایی که واقعاً در خوشه ۲ هستند و به‌اشتباه در خوشه ۳ پیش‌بینی‌شده‌اند: ۸ نمونه
-ازاین‌رو تعداد نمونه‌هایی که واقعاً در خوشه ۲ هستند و به‌اشتباه در خوشه ۰ پیش‌بینی‌شده‌اند: ۳ نمونه

تحلیل نتایج

نتایج دو مدل درخت تصـــمیم و جنگل تصـــادفی در خروجی خوشه‌بندی را به‌طور خلاصه جهت تحلیل در جدول ۵ گزارش شده است.

با دقت در نتایج مشـــاهده می‌کنیم که دقت روش جمعی جنگل تصـــادفی که از ۱۰ درخت تصـــمیم برای پیش‌بینی اســـتفاده می‌کند، با دقت ٪۹۱,۹۷ دقت بالاتری از روش احتمالی بیزساده دارد. از آنجا که نسبت نمونه‌ها در خوشه های مختلف بسیار متفاوت است و عدم توازن در خوشه‌ها مشاهده می‌شود. دقت مدل‌ها در خوشه ها تفاوت زیادی دارد. در روش بیزساده روی خوشه ۱ که تنها ٤ آیتم دارد قطعیت پایین ۵۰ درصد دارد.

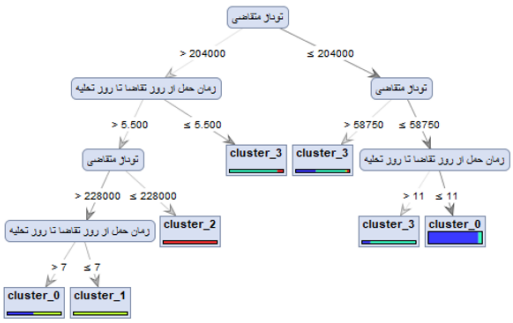
اما در خوشـــه ها دیگر نیز میانگین ۸۱ درصـــد داریم که خطای ۱۵ درصـــد بیشـــتر از جنگل تصـــادفی اســـت. به طور کلی در روشـــهای جمعی، به دلیل تقســـیم مجموعه داده، روی داده‌های نامتوزان دقت بهتری دارند و نتایج را بهبود می‌دهند.

جدول ۵. نتایج دسته بندی

Model	Recall (cluster0)	Recall (cluster3)	Recall (cluster2)	Recall (cluster 1)	Accuracy
NB	91.19%	81.67%	87.50%	50%	89.77%
Random forest	99.75%	91.67%	54.17%	100%	%97.91

تحلیل دانش درختان جنگل تصادفی

درخت تصمیم ازجمله روش‌هایی است که خروجی آن به‌صورت قوانین قابل‌تحلیل بررسی می‌شود، ازاین‌رو جزو روش‌های جعبه سفید است و در مقایسه با الگوریتم‌های جعبه سیاه مانند شبکه عصبی، نتایج آن را می‌توان به صورت قوانین «اگر آنگاه» بررسی کرد و ارتباط بین فاکتورها را در حوزه حمل و نقل تحلیل کرد. چند مدل درختان جنگل تصادفی که در بخش دسته‌بندی اشاره کردیم، به شرح زیر است.



شکل ۱۵. درخت ۱

Tree 1

تناژ متقاضی < ۲۰٤۰۰۰

| زمان حمل از روز تقاضا تا روز تخلیه < ۵,۵۰۰

| | تناژ متقاضی < ۲۲۸۰۰۰

| | | زمان حمل از روز تقاضا تا روز تخلیه < ۷: cluster_0 {cluster_0=1, cluster_1=1, cluster_2=0, cluster_3=0}

| | | زمان حمل از روز تقاضا تا روز تخلیه ≥ ۷: cluster_1 {cluster_0=0, cluster_1=3, cluster_2=0, cluster_3=0}

| | تناژ متقاضی ≥ ۲۲۸۰۰۰: cluster_2 {cluster_0=0, cluster_1=0, cluster_2=2, cluster_3=0}

| زمان حمل از روز تقاضا تا روز تخلیه ≥ ۵,۵۰۰: cluster_3 {cluster_0=0, cluster_1=2, cluster_2=2, cluster_3=17}

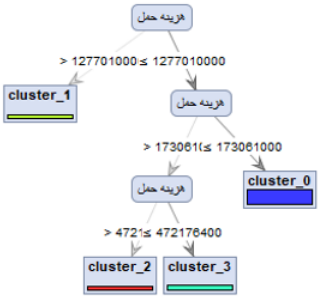
تناژ متقاضی ≥ ۲۰٤۰۰۰

| تناژ متقاضی < ۵۸۷۵۰: cluster_3 {cluster_0=47, cluster_1=4, cluster_2=5, cluster_3=66}

| تناژ متقاضی ≥ ۵۸۷۵۰

| | زمان حمل از روز تقاضا تا روز تخلیه < ۱۱: cluster_3 {cluster_0=1, cluster_1=0, cluster_2=0, cluster_3=5}

| | زمان حمل از روز تقاضا تا روز تخلیه ≥ ۱۱: cluster_0 {cluster_0=1157, cluster_1=0, cluster_2=0, cluster_3=77}



شکل ۱۶. درخت ۲

Tree 2

هزینه حمل < ۱۱۴۱۰۷۰۰۰۰: cluster_1 {cluster_0=0, cluster_3=0, cluster_1=6, cluster_2=0}

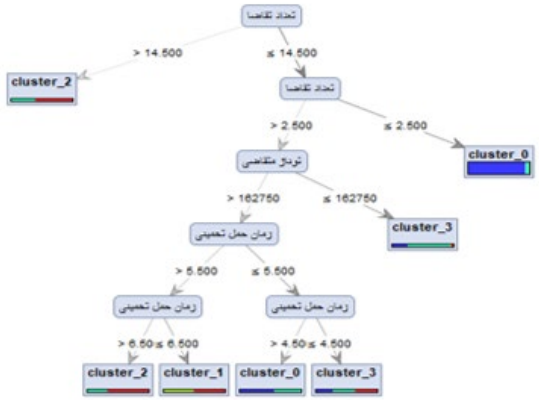
هزینه حمل ≥ ۱۱۴۱۰۷۰۰۰۰

| هزینه حمل < ۱۷۳۱۱۳۵۰۰

| | هزینه حمل < ۵۱۹۰۰۸۸۵۰: cluster_2 {cluster_0=0, cluster_3=0, cluster_1=0, cluster_2=13}

| | هزینه حمل ≥ ۵۱۹۰۰۸۸۵۰: cluster_3 {cluster_0=0, cluster_3=186, cluster_1=0, cluster_2=0}

| هزینه حمل ≥ ۱۷۳۱۱۳۵۰۰: cluster_0 {cluster_0=1183, cluster_3=0, cluster_1=0, cluster_2=0}



شکل ۱۷. درخت ۳

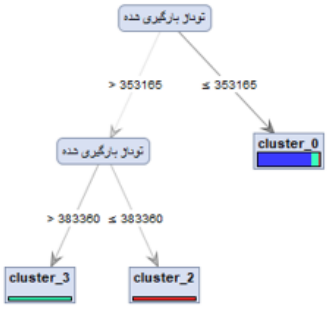
Tree 3

تناژ بارگیری شده < ۳۵۳۱۶۵

| تناژ بارگیری شده < ۳۸۳۳۶۰: cluster_3 {cluster_0=0, cluster_3=9, cluster_1=0, cluster_2=0}

| تناژ بارگیری شده ≥ ۳۸۳۳۶۰: cluster_2 {cluster_0=0, cluster_3=0, cluster_1=0, cluster_2=7}

تناژ بارگیری شده ≥ ۳۵۳۱۶۵: cluster_0 {cluster_0=1193, cluster_3=157, cluster_1=1, cluster_2=21}



شکل ۱۸. درخت ٤

Tree 4

تعداد تقاضا < ۱٤,۵۰۰: cluster_2 {cluster_0=0, cluster_3=2, cluster_1=0, cluster_2=3}

تعداد تقاضا ≥ ۱٤,۵۰۰

| تعداد تقاضا < ۲,۵۰۰

| | تناژ متقاضی < ۱٦۲۷۵۰

| | | زمان حمل تخمینی < ۵,۵۰۰

| | | | زمان حمل تخمینی > ٦,٥٠٠: {cluster_0=0, cluster_3=1, cluster_1=0, cluster_2=2} cluster_2
| | | | زمان حمل تخمینی ≥ ٦,٥٠٠: {cluster_0=0, cluster_3=0, cluster_1=3, cluster_2=3} cluster_1
| | | زمان حمل تخمینی ≥ ٥,٥٠٠
| | | | زمان حمل تخمینی > ٤,٥٠٠: {cluster_0=4, cluster_3=3, cluster_1=0, cluster_2=0} cluster_0
| | | | زمان حمل تخمینی ≥ ٤,٥٠٠: {cluster_0=4, cluster_3=5, cluster_1=0, cluster_2=5} cluster_3
| | تناژ متقاضی ≥ ١٦٢٧٥٠: {cluster_0=30, cluster_3=77, cluster_1=2, cluster_2=4} cluster_3
| تعداد تقاضا ≥ ٢,٥٠٠: {cluster_0=1158, cluster_3=79, cluster_1=0, cluster_2=3} cluster_0

تجزیه و تحلیل نتایج

مجموعه داده جمع آوری شده شامل ۱۳۸۸ اطلاعات مرتبط به حمل نقل انواع کالا یک شرکت حمل‌ونقل در انواع ریلی و جاده ای است.که کالاهای شامل شمش روی، و شمش آلومینیومی را به کشور ترکیه صادر می کند. داده ها شامل ١٤ ستون عددی و گسسته است. نوع ویژگی ها مبدا، مقصد، نوع بار، هزینه حمل، زمان حمل تخمینی و زمان انجام شده، نوع حمل، تعداد عرضه، تناژ متقاضی و تناژ بارگیری است.

با تشریح اطلاعات آماری و نمودارهای فراوانی اطلاعات زیر استخراج شد:

-تعداد عرضه و تقاضا از رنج ۱ تا ۱۸ و میانگین ۱,٦٩٨ مشخص شده است.

-تناژ بارگیری از ٤٩٠٠ کیلوگرم تا ٤٨٧٠٧٠ کیلو تعریف شده است.

-۵۳ درصد حمل و نقل از واحد مبدا از شهر سهلان، ۳۰ درصد از زنجان و مابقی با درصد بسیار کمتر در شهرهای قزوین، اصفهان، یزد، قم، تهران، ایلام، سلفچگان و اراک و ... تعریف شده است.

-زمان حمل تخمین زده شده، از مینیمم ۳ روز تا ۹ روز و میانگین ٤,٦١ تعریف شده است. زمان حمل از روز تقاضا تا روز تخلیه از مینیمم ۳ روز تا ۱۲ روز تعریف شده است.

-٤٦ درصد حمل و نقل‌ها در حالت کامیونی و ۵۳ درصد در حالت ریلی هستند.

-٩٦ درصد بارها در حمل و نقل نیاز به مجوز ندارند و اطلاعات از سال ٩٦ تا ۹۸ تعریف شده است.

-ماکزیمم رنج تناژ متقاضی و تناژ بارگیری شده، برای حمل کامیونی تا میانگین ٤٧٥٠٠٠ تعریف شده و برای حمل ریلی تنها از رنج ۲۵۰۰۰ تا ۵۰۰۰۰ کیلوگرم تعریف شده است.

-برای حمل ریلی تنها شهر سهلان مبدا حمل و نقل بوده و سایر بارها در حمل کامیونی در سایر شهرها تعریف شده‌اند.

-در حمل ریلی تنها شهرهای وان، قاضی انتپ، مرسین و اسکندرون مقصد حمل و نقل و مابقی مربوط به حمل کامیونی هستند.

-بیشترین زمان حمل از روز تقاضا تا روز تخلیه، در فاصله زمانی سه روز تا ۱۲ روز برای حمل کامیونی و برای حمل ریلی تا میانگین ۵ روز و نهایت ۸ روز فاصله زمانی داریم.

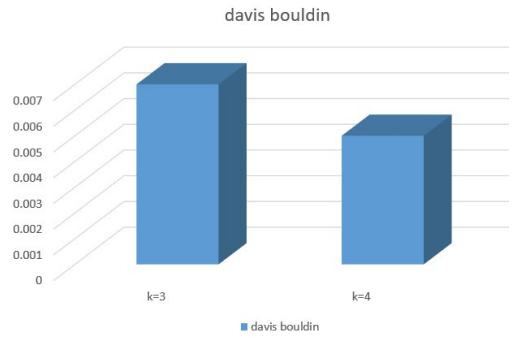
-بیشترین هزینه حمل و نقل مربوط به حمل کامیونی تا ۱,۵۰۰,۰۰۰,۰۰۰ ریال است و تفاوت هزینه در حمل ریلی و کامیونی مشهود است.

در گام پاکسازی داده تنها در ویژگی تعداد عرضه دو سطر داده از دست رفته داریم، که به جای حذف سطرها که اطلاعات مهم ستون های دیگر حذف شود، از جانشینی داده با مقدار میانگین ستون استفاده کردیم. متغیر تاریخ تقاضا از نوع شمسی است و به دلیل وجود فاصله زمانی تقاضا تا روز حمل، نیازی به این ویژگی نداشته و آن را حذف کردیم.

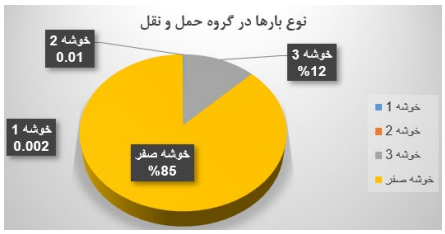
استخراج دانش (خوشه بندی)

جهت بهینه‌سازی تعداد خوشه‌ها و بهترین مقدار خوشه‌بندی از ارزیاب دیوس بولدین و مفاهیم خو شه‌بندی به د ست آمد؛ این ارزیاب هر چه کمتر با شد خو شه‌ها بهینه‌تر یعنی فا صله درون خو شه‌ای متراکم‌تر و بین خو شه‌ای بی شتر ا ست. که برا ساس نمودار ۱-۵ با مقدار k=٤ بهترین خوشه‌بندی با کامینز را داشتیم.

در ادامه نتایج گروه‌بندی م شتریان خو شه‌بندی کامینز را ازنظر تعداد نمونه‌ها در هر خوشه، جدول میانگین مرکزی خوشه‌ها بررسی کردیم؛ خوشه ها را از نظر مفاهیم حمل و نقل تحلیل کردیم؛ شکل ۲۰.



شکل ۱۹. مقایسه ارزیاب خوشه بندی با دیویس بولدین



شکل ۲۰. نمودار فراوانی گروه بارهای حمل و نقل

خوشه ۱ (بارهای پرهزینه کامیونی) که با تنها ٤ آیتم که از نظر هزینه حمل ، تناژ بارگیری به کیلو، تعداد عرضه و تقاضا، ماکزیمم تایم زمانی حمل از روزتقاضا تا تخلیه بیشترین مقدار را داشتند و متفاوت از خوشه‌های دیگر در یک گروه مجزا قرار داده شدند. همه بارهای کامیونی از شهر قزوین به دو شهر آدانای ترکیه و شهر MURATBAY ارسال شده‌اند.

خوشه صفر (کم هزینه ترین بار با بیشترین بار ریلی): بیشترین تعداد بارها (۸۵ در صد) که به طور میانگین از نظر هزینه حمل، تناژ بارگیری به کیلو، تعداد عرضه و تقاضا، ماکزیمم تایم زمانی حمل از روزتقاضا تا تخلیه کمترین مقدار را داشتند. مبدا بیشتر این بارها که عمدتا از نوع ریلی و ۳۷ درصد کانتینر و مابقی شمش روی هستند از شهر سهلان است و به طور میانگین به تمامی شهرهای ترکیه بار ارسال شده است.

خوشه ۲ (بارهای نسبتا پرهزینه کامیونی در وضعیت هوای نا معتدل): بعد از خوشه ۱ بیشترین هزینه حمل، بیشترین تناژ بارگیری بعد از خوشه ۱، و بیشترین تعداد عرضه و تقاضا با میانگین ٦,۱ در مورد آیتم‌های این گروه با ۲٤ نمونه است. وجه تمایز این خوشه با خوشه ۱ در این است که ٤۰ در صد بارها در وضعیت هوایی نامعتدل است که عمدتا به وان و آدانا و آنکارا ارسال شده است.

خوشه ۳ (شاخص‌های متوسط از نظر حمل و نقل با بیشترین فاصله زمانی حمل): هزینه های حمل و نقل در این گروه متوسط، از خوشه ۱ و ۲ کمتر از خوشه صفر بیشتر است. تفاوت شاخص این گروه نسبت به گروه‌های دیگر در تفاوت فاصله زمانی نسبت زمان حمل تخمینی تا زمان حمل از روز تقاضا و روز تخلیه است که نسبت به گروه‌های دیگر بیشتر است. در اصل ۱۸۰ بار عمدتا از نوع شمش روی، به طور متوسط از زنجان به تمامی شهرهای ترکیه ارسال شده است.

در بخش مصورسازی داده‌ها از نمودار deviation استفاده کردیم. با تحلیل این نمودار مشخص شد که زمان حمل، هزینه حمل و تا حدودی تناژ بار و وضعیت هوا تاثیرگذاری بیشتری نسبت به سایر فاکتورها در جداسازی نوع حمل و نقل دارد.

استخراج دانش (قواعد انجمنی)

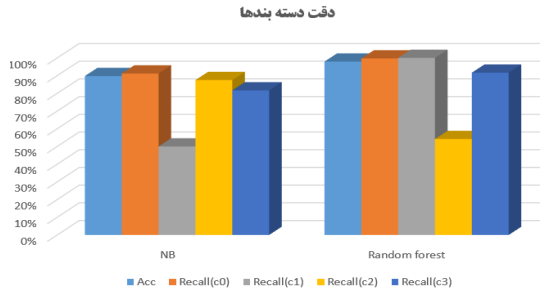
ابتدا با روش fp-growth و مینیمم اطمینان ۷۰ درصد الگوهای پرتکرار را ایجاد کردیم و براساس الگوهای پرتکرار ۹ قانون انجمنی را روی شاخص‌های کیفی استخراج کردیم که به شرح زیر است.

۱. با اطمینان ۸۳,۵ درصد اگر مقصد بار شهر وان است، آنگاه نوع بار از نوع کانتینر است.

۲. با اطمینان ۸۳,۵ درصد اگر مقصد بار شهر وان است، آنگاه نوع بار از نوع کانتینر و مبدا از شهر سهلان است.
۳. با اطمینان ۸۹,۵ درصد اگر مقصد بار شهر وان است ، آنگاه مبدا از شهر سهلان است.
٤. با اطمینان ۹۳,۳ درصد، اگر مقصد بار شهر وان و مبدا از شهر ، نوع بار کانتینر است.
۵. با اطمینان ۱۰۰ درصد، اگر مبدا شهر زنجان است، نوع حمل کامیونی است.
٦. با اطمینان ۱۰۰ درصد، اگر نوع بار شمش روی باشد، نوع حمل کامیونی است.
۷. با اطمینان ۱۰۰ درصد، اگر نوع بار کانتینر باشد، مقصد شهر وان است.
۸. با اطمینان ۱۰۰ درصد، اگر نوع بار کانتینر باشد، مقصد شهر وان و مبدا شهر سهلان است.
۹. با اطمینان ۱۰۰ درصد، اگر نوع بار شمش روی باشد، مبدا زنجان و نوع حمل کامیونی است.

استخراج دانش (دسته‌بندی)

نتایج زیر از تحلیل و مقایسه دقت دسته‌بندها در شکل ۲۱ گزارش می‌شود.



شکل ۲۱. دقت دسته‌بندها

با دقت در نتایج مشاهده می‌کنیم که دقت روش جمعی جنگل تصادفی که از ۱۰ درخت تصمیم برای پیش بینی استفاده می‌کند، با دقت ۹۱,۹۷ درصد دقت بالاتری از روش احتمالی بیز ساده دارد. از آنجا که نسبت نمونه‌ها در خوشه‌های مختلف بسیار متفاوت است و عدم توازن در خوشه‌ها مشاهده می‌شود. دقت مدل‌ها در خوشه‌ها تفاوت زیادی دارد. در روش بیز ساده روی خوشه ۱ که تنها ٤ آیتم دارد قطعیت پایین ۵۰ درصد دارد. اما در خوشه‌ها دیگر نیز میانگین ۸۱ درصد داریم که خطای ۱۵ درصد بیشتر از جنگل تصادفی است.

به دلیل متوازن نبودن داده‌ها از روش تولید قوانین، جعبه سفید و ترکیبی جنگل تصادفی استفاده شد. این روش به دلیل تقسیم مجموعه داده‌ها بین ۱۰ درخت تصمیم موجب توازن داده‌ها شده و دقت را روی تمام کلاس‌ها با مقدار بسیار بالا گزارش داد. ریشه و والد بودن متغیر هزینه حمل در اکثر درخت‌ها نشان دهنده تاثیرگذاری این متغیر در خوشه بندی است. تعدادی از قوانین تولیدی و مفید درختان تصمیم و جنگل تصادفی به شرح ادامه است.

۱. اگر هزینه حمل بیشتر از یک میلیارد باشد، بارها در خوشه ۱ (بارهای پرهزینه پرتقاضای کامیونی) هستند.
۲. اگر هزینه حمل کمتر از ۱۷۳ میلیون باشد بار در خوشه بارهای کم هزینه و کم تقاضای ریلی است.
۳. اگر هزینه حمل بیشتر از ۵۱۹ میلیون باشد در خوشه ۲ و کمتر از ۵۱۹ میلیون و بیشتر از ۲۰۰ میلیون در خوشه ۱۳ است. (تاثیر مستقیم ویژگی هزینه حمل)
٤. گر تناژ بارگیری کمتر از ۳٦۳۱٦۵ کیلو گرم باشد در خوشه صفر و اگر بیشتر از ۳۸۳۳٦۰ کیلو باشد در خوشه ۳ است.
۵. اگر تناژ متقاضی کمتر از ۵۸۷۵۰ کیلو و زمان حمل از روز تقاضا تا روز تخلیه کمتر از ۱۱ روز باشد در خوشه صفر است.
٦. اگر تناژ متقاضی بیشتر از ۲۰۳۰۰۰ کیلو و زمان حمل بیشتر از ۵ روز در خوشه ۲ است.
۷.ولی اگر تناژ بیشتر از ۲۲۸۰۰۰ و زمان حمل بیشتر از ۷ روز باشد در خوشه ۱ است.
۸.اگر تعداد تقاضا کمتر از ۲,۵ باشد در خوشه صفر است.

۵-نتیجه‌گیری

در این مقاله جهت بررسی فاکتورهای تاثیرگذار در هزینه و زمان حمل جاده و ریلی و استخراج دانش از سه تکنیک داده کاوی: خوشه بندی، قواعد انجمنی و دسته بندی استفاده کردیم. در گام اول با روش خوشه بندی مبتنی بر فاصله کامیانگین ۱۳۸۸ تراکنش حمل و نقل جاده و ریلی را به ٤ خوشه بهینه تقسیم کردیم. نتایج نشان داد با اختلاف، ویژگی‌های هزینه حمل، تناژ و تعداد عر ضه و تقا ضا تاثیرگذارترین ویژگی‌ها در جداسازی خوشه‌ها هستند. درصد بسیار کمی آب و هوا و نوع حمل و نوع بار تاثیر در گروه بندی بارهای حمل و نقل داشت. در خوشه صفر و یک هزینه حمل و مقدار عرضه و تقاضا تفاوت شاخص با گروه های دیگر داشت. در خوشه ۲ وضعیت هوای نامعتدل و در خوشه ۳ فاصله زمانی از روز تقاضا تا روز تخلیه موجب تفاوت خوشه ای با دیگر خوشه ها شده بود. در فاز قواعد انجمنی، ۹ قواعد با قطعیت بالای ۸۵ درصد گزارش روی متغیرهای کیفی نوع بار و مبدا و مقصد گزارش شد. در گام د سته بندی از دو روش بیز ساده (روش احتمالی و روش جمعی جنگل تصادفی) روی خروجی خوشه بندی استفاده شد. نتایج نشان داد جنگل تصادفی با دقت کل ۹۷,۹۱ درصد نتایج بسیار دقیق تری از بیزساده برای دسته بندی خو شه‌ها دارد. این روش هم چنین اهمیت الگوریتم‌های جمعی (رای گیری) را جهت بهبود دقت روی داده‌های نامتوزان اثبات کرد. با درختان جنگل تصادفی، ۸ قواعد مفید با استخراج شد. نتایج قوانین اثبات کرد که در والد بیشتر درختان هزینه حمل و تناژ بارگیری معیار مقایسه خوشه‌ها و ساخت درخت است که اهمیت این فاکتورها را در گروه بندی سیستم حمل و نقل جاده ای وریلی اثبات کرد. به‌طورکلی اگر محققان در آینده بخواهند تو سعه‌ای بر نتایج حا صل این پژوهش پیشنهاد دهند، می‌توانند موارد زیر را مدنظر قرار دهند: ۱) استفاده از روش‌های دیگر خوشه‌بندی مانند som و خوشه‌بندی فازی.۲) استفاده از روش‌های بهینه سازی تعداد خو شه‌ها مانند الگوریتم فرابتکاری ژنتیک و ازدحام ذرات.

٦-مراجع

-اسماعیلی، م، (۱۳۹۱). مفاهیم و تکنیک‌های داده کاوی، انتشارات نیاز دانش.

- ع. م. احمدوند. ب. مینایی بیدگلی. ا. اخوندزاده. (۱۳۸۸). تحلیل رضایتمندی شهروندان با استفاده از تکنیکهای داده کاوی: مورد کاوی شهرداری تهران. *سومین کنفرانس داده کاوی.*

- رستگار ، ن. ، علیزاده ، س. ، غضنفری ،م. ، وزیری،ف. (۱۳۸۷). کاربرد روش خو شه بندی در انتخاب ا ستراتژی های صادرات کالاها در ایران. *دومین کنفرانس داده کاوی ایران.*

- محمد بیگی، زهرا (۱۳۹۳). کاربرد RFID مه منظور کنترل و کاهش تصادفات درون شهری با رویکرد داده کاوی.

- کیانیان، علیرضا (۱۳۹۵). ارزیابی و بهبود کیفیت خدمات حمل و نقل عمومی با ا ستفاده از رویکرد تلفیقی داده کاوی و تو سعه عملکرد کیفیت (مطالعه موردی: اتوبوسرانی مشهد).

-Olson,D. (2003). Managerial Issues of Enterprise Resource Planning. *McGraw-Hill/Irwin,* 1 edition, 45.

-Adams S.O, Akano R.O, Asemota O, J. (2011). Forecasting Electricity Generation in Nigeria using Univariate Time Series Models. *European journal of Scientific Reaserch.*Vol.58. No.1. 30-37.

-Albert Y. Chen, Ting-Yi Yu. (2016). Network based temporary facility location for the Emergency Medical Services considering the disaster induced demand and the transportation infrastructure in disaster response. *Transportation Research Part B.* Approaches and Big Data Analytics in Smart Transportation System.

-C. S. Li and M.-C. Chen (2014). A data mining-based approach for travel time prediction in freeway with non-recurrent congestion. *Neurocomputing,* Vol. 133, 74-83.

-David J. HAND, (2001). Data Mining: Statistics and More? *The American Statistician,* Vol. 52, No. 2. 112-113

-Fayyad.U, Piatetsky-Shapiro.G, Padhraic.S, (2008). From Data Mining to Knowledge Discovery in Databases.

-Feyza Gürbüza, Fatma Turnab (2018). Rule extraction for tram faults via data mining for safe. *Transportation Research Part A.*

-Han.J, Kamber.M, (2011). Data Mining: Concepts and Techniques. San Diego Academic Press, 25-26.

Hana Ševcˇíková, Adrian E. Raftery, Paul A. Waddell (2011). Uncertain benefits: Application of

-Soyoung Jung, Xiao Qin, Cheol Oh. (2016). Jung, S., Qin, X., & Oh, C. (2016). Improving strategic policies for pedestrian safety enhancement using classification tree modeling. *Transportation Research Part A: Policy and Practice,* *85*, 53–64.

-Venkata Siva Reddy and R. Vasanth Kumar Mehta. (2019). Study on Computational Intelligence

-Wirth, R. and J. Hipp. CRISP-DM, (2000). Towards a standard process model for data mining. in Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining. *Citeseer.*

-Yuxuan Ji, Nikolas Geroliminis (2012). On the spatial partitioning of urban transportation networks. *Transportation Research Part B.*

-Zhang G Peter, Min Qi. (2005). Neural network forecasting for seasonal and trend time series. *European Journal of Operational Research.* Vol. 160. 501-515

-Zhang G Peter. (2003). Time series forecasting using a hybrid ARIMA and neural network model. *Neurocomputing 50*. 159-175.

Bayesian melding to the Alaskan. *Transportation Research Part A.*

-Herbert A. E. (1996). Introduction to Data Mining and Knowledge Discovery.*Two Crows Corporation,* Third Edition. 5-10.

-Hu R. (2010). Medical data mining based on association rules. *Computer and Information Science.* 3(4): 104-8.

-Jithin Raj, Hareesh Bahuleyan, Lelitha Devi Vanajakshi. (2016). Application of data mining techniques for traffic density estimation and prediction.

-Kremic, E. and A. Subasi (2015). Performance of Random Forest and SVM in Face Recognition. *The International Arab Journal of Information Technology.*

-Lathia.N, Froehlich.J, Capra.L, (2010). Mining Public Transport Usage for Personalised Intelligent Transport Systems. *IEEE ICDM.*

-Mehrdad Kargari, Mohammad Mehdi Sepehri. (2012). Stores clustering using a data mining approach for distributing automotive spare-parts to reduce transportation costs.

-Milad Ghasri (2017). Taha Hossein Rashidi, S. Travis Waller. Developing a disaggregate travel demand system of models using. *Transportation Research Part A.*

-Ramageri M. (2011). Data mining techniques and applications. *Indian Journal of Computer Science and Engineering.* 4(1): 301-5.

Determining Factors Affecting the Cost and Time of Road and Rail Transportation Using Association Rules

Mahdi Yousefi Nejad Attari, *Department of Industrial Engineering, Bon. C., Islamic Azad University, Bonab, Iran.*

Ensiyeh Neishabouri Jami, Department of industrial Engineering, University of Bonab, Bonab, Iran.

E-mail: 1689597631@iau.ir

Received: March 2026- Accepted: August 2026

ABSTRACT

Transportation or movement refers to the movement of people, animals, or goods from one place to another. In other words, the act of transportation is defined as a specific movement from point A to point B. Types of transportation methods include air, land (rail and road), sea, cable, pipeline, and space. In land transportation, due to the importance of this business and the competition between organizations active in this field, the use of new technology and technologies in management and making better decisions can be very useful. In this article, we will first introduce data mining and its techniques and data mining software. Then, considering a table of export shipments made in the past one year and six months for an international transportation company, we prepared and extracted association rules in two models using Rapid Miner software and evaluated the relevant rules. The results presented show the factors affecting time and cost and finally, future suggestions are expressed.

Keywords: Data Mining, Time, Road and Rail Transportation, Clustering, Cost