ارایه الگوی دسته بندی مشتریان با رویکرد داده کاوی ترکیبی (مورد مطالعه صنعت محصولات بهداشتی وآرایشی)

امید بشردوست[۱]
عزت اله اصغری زاده[۲]
محمد علی افشار کاظمی[۳]

تاریخ دریافت مقاله : ۹۹/۱۲/۱۷ تاریخ پذیرش مقاله : ۱۴۰۰/۰۱/۱۰

چکیده

با توجه به حجم انباشته شده اطلاعات خریدمشتریان و پیچیدگی رقابت در عصر حاضر اهمیت ایجاد بستری برای تحلیل داده‌های به روز و دقیق مشتریان، با هدف ایجاد ارتباط‌های مؤثر با مشتریان فعلی و وفادار، بیش از پیش برای سازمان‌ها به عنوان یک مزیت رقابتی جلوه‌گر شده است. هدف این پژوهش بررسی الگوهای رفتاری خرید مشتریان محصول‌های بهداشتی به منظور دسته‌بندی آنها بر اساس مدل WRFMبا استفاده از روش‌های ترکیبی داده‌کاوی است. از میان مشتریان استان تهران که دربازه سال های۱۳۹۶-۱۳۹۷ازشرکت خرید داشته‌اند ازپایگاه داده‌های مشتریان ۶۵۵۳۴ نمونه، باروش نمونه‌گیری هدفمند در دسترس جمع آوری شده و به کمک SPSS مقدارWRFM با توجه به نظرخبرگان صنعت مشخص و سپس این فیلد به دیگر داده‌های پژوهش اضافه شده و توسط نرم افزار داده‌کاوی کلمنتاین بر اساس۷۰ درصد داده‌ها، خوشه‌بندی مشتریان صورت گرفته است؛ همچنین به منظور بررسی کیفیت خوشه‌بندی از معیارهای امتیازجینی، درصدخطا، اطلاعات متقابل نرمال‌شده (NMI)استفاده شده است. نتایج پژوهش حکایت از کارایی بالای روش خوشه‌بندی Kمیانگین با تعداد چهارخوشه با درصد خلوص (۰/۷۶۱)، برای بخش‌بندی مشتریان داشته است.

کلمات کلیدی

الگوهای رفتار خرید، داده‌کاوی، خوشه‌بندی، بخش‌بندی، **WRFM**

۱-گروه مدیریت، واحد رودهن، دانشگاه آزاد اسلامی، رودهن، ایران. dr.o.bashardoust@gmail.com
۲-گروه مدیریت، دانشکده مدیریت، دانشگاه تهران، تهران، ایران.(نویسنده مسئول) asghari@ut.ac.ir
۳-گروه مدیریت صنعتی، واحد تهران مرکزی،دانشگاه آزاد اسلامی ، تهران، ایران. M_afsharkazemi@iauec.ac.ir

مقدمه

دانش نوین داده‌کاوی از بسترهای در حال توسعه است که دهه حاضر را با انقلاب فناورانه مواجه ساخته و به همین علت است که در سال‌های اخیر در دنیا گسترش سریعی پیداکرده است. دانش داده‌کاوی فرآیند کشف دانش پنهان درون داده‌هاست که با برخورداری از دامنه وسیع زیر زمینه‌های تخصصی با توصیف، تشریح، پیش‌بینی وکنترل پدیده‌های گوناگون پیرامونی، امروزه دارای کاربرد وسیع در حوزه‌های مختلف است که دیگر نمی‌توان محدودیتی برای کاربرد آن در نظرگرفت. سازمان‌های امروزی داده‌های زیادی را در تصرف خود دارند و درحالی‌که گرسنه دانش درون داده‌ها بوده و با فقدان دانش پنهان درون داده‌ها مواجه‌اند. با توجه به تنوع زیاد مخاطبان، مشتریان، بازارها و پیچیدگی خدمات و محیط‌های کسب وکار، دسترسی به اطلاعات مناسب برای تصمیم‌گیری صحیح ضروری است. از این‌رو استفاده از راهکارهای مناسب برای طبقه‌بندی و تولیداطلاعات ازمیان انبوهی از داده‌ها برای سازمان امری ضروری و حیاتی است. داده‌کاوی پاسخی به‌این‌نیازباهدف کشف دانش پنهان درون‌داده‌ها و تأمین‌اطلاعات مورد نیاز است پس ضروری است برای به کاربستن این ابزار در سازمان‌ها اهمیت بیشتری قایل شده تا در نهایت به فرآیند تصمیم‌گیری بهینه مدیران منجر شود (شهرابی، ۱۳۹۴). در محیط کسب وکار امروز که پیچیدگی‌های فناورانه و رقابتی شدن آن روز افزون است، مشتری‌مداری و حفظ مشتری به عنوان یک استراتژی، برای ایجاد مزیت رقابتی محسوب می‌شود (محمدی و سهرابی، ۱۳۹۶).

در این پژوهش صنعت محصول‌های بهداشتی و آرایشی مورد تحلیل و بررسی قرارگرفته است که این صنعت نیز از این قاعده مستثنی نبوده؛ چراکه با حجم انبوهی از اطلاعات مشتریانی مواجه است که سلیقه و شیوه‌های خرید آنها با تغییرهای سریعی در حال دگرگونی است و همچنین تشخیص سریع و به موقع رفتارهای گوناگون خرید مشتریان برای مدیران ارشد سازمان به خصوص مدیران فروش بازاریابی بالتبع آن مدیران تولید شرکت حایز اهمیت است چراکه اگر این شیوه‌ها و روندها تداوم داشته باشند و به موقع شناسایی شوند نشان از قاعده مند بودن بعضی از خریدهای مشتریان دارد که برای امر برنامه‌ریزی و تصمیم‌گیری شرکت می‌تواند بسیار مفید باشد. پژوهشگر به دنبال بررسی این صنعت و رفتار مشتریان به کمک داده‌کاوی است تا مشخص نماید اگر داده‌کاوی ابزاری قدرتمند و کارا برای تحلیل رفتار مشتریان است؛ این ابزار و متدولوژی این پژوهش بتواند راهگشای مسایل پیش‌روی این صنعت برای مدیران فروش و بازاریابی شرکت باشد.

مبانی نظری و مروری بر مطالعه‌های گذشته

مبانی نظری

دانشمندان حوزه‌های علمی مختلف مانند: اقتصاد و بازاریابی، مدل‌های متعددی برای تحلیل رفتار مصرف کننده عرضه کرده‌اند. اقتصاددانان، اولین افرادی‌اند که با رویکرد علمی به رفتار مصرف‌کننده توجه کرده‌اند. از جمله مهمترین مدل‌های تحلیل رفتار مصرف‌کننده که اقتصاددانان عرضه کرده‌اند عبارت است از مدل منحنی بی‌تفاوتی و مدل خصوصیات (داگلاس، ۱۳۷۵، ۱۰۲-۱۳۷). این مدل‌ها از زیرمجموعه‌های علم اقتصاد خردند و به تحلیل این امر می پردازند که چگونه مصرف کننده منابع محدود خود را نسبت به نیازهای گوناگون خود تخصیص دهد. در مدل منحنی بی تفاوتی-که مدلی برای پیش‌بینی رفتار مصرف کننده است- فرض بر این است که مصرف‌کنندگان از مصرف کالاها و خدمات، مطلوبیت کسب می‌کنند. فرض می‌شود که مصرف‌کننده در دوره‌ای مشخص برای محصولی خاص، از مصرف متوالی کالا، مطلوبیت نزولی به دست می آورد. با توجه به اینکه فرض می‌شود مصرف کننده درجستجوی حداکثر مطلوبیت است، ترکیب‌های‌کالایی و خدماتی را که مطلوبیت کل بیشتری نسبت به ترکیب‌هایی که مطلوبیت کل کمتری دارند، انتخاب می‌کند و بین ترکیب‌هایی که مطلوبیت کل یکسانی دارند بی‌تفاوت است. در مدل خصوصیات، فرض بر آن است که مصرف‌کنندگان، مطلوبیت خود را از خود کالا به دست نمی آورند بلکه از خصوصیات و خواص کالا، رضایت کسب می‌کنند. محققان بازاریابی متأثر از اندیشه‌های اقتصادی رفتار مصرف‌کننده، به پژوهش و تحقیق در زمینه رفتار مصرف‌کننده پرداخته‌اند. در این رویکرد، رفتار مصرف‌کننده وسیله‌ای است برای تأمین نیازها و خواسته‌های فرد که دربردارنده فرآیندهای ذهنی و فیزیکی از قبل خرید تا بعد از مصرف کالاها و خدمات است به عبارت دیگر می‌توان گفت ادبیات رفتار مصرف‌کننده به توصیف و تحلیل نحوه تصمیم‌گیری در زمینه خرید و شیوه بهره‌گیری از کالاها و خدمات خریداری شده تمرکز دارد (دنیل و همکاران[1]،۲۰۰۳،۱۵۲)؛ بدین منظور محققان بازاریابی همچون: انگل، کاتلر، آرمسترانگ، ساندرز و وانگ[2](۲۰۰۱)، مدل‌هایی برای تحلیل رفتار مصرف کننده عرضه کرده‌اند. مدل انگل و همکاران[3](۱۹۶۸) شامل مراحل تشخیص نیاز، جستجوی اطلاعات، ارزیابی و بررسی راه‌حل‌های جایگزین، خرید و نتایج است. مدل هوارد و شث[4](۱۹۶۹) دربردارنده چهارمرحله: داده‌ها (آگاهی از محصول‌ها و ویژگی‌های آنها به طور مستقیم از طریق رسانه‌های گروهی)، ساخت ادراکی (فعالیت ذهنی بر روی داده‌ها)، ساخت یادگیری (فعال شدن مواردی نظیر انگیزه، بینش و معیارهای ارزشی مصرف‌کننده) و ستاده‌ها (خرید) است. مدل فرآیندی ویلک[5](۲۰۰۰) متشکل از فعالیت‌های پیش ازخرید (تشخیص مسأله و جمع‌آوری اطلاعات و ارزیابی گزینه‌ها)، فعالیت‌های ضمن

خرید (تصمیم‌گیری و خرید) و فعالیت‌های پس از خرید (مصرف و ارزیابی محصول‌ها و فرآیندهای تصمیم‌گیری آتی) است. مدل کاتلر و همکاران (۲۰۰۱) درباره رفتار مصرف‌کننده، فرآیندی متشکل از: ۱- محرک‌ها: شامل محرک‌های بازاریابی (کالا، قیمت، مکان و تبلیغات) و سایر محرک‌ها (اقتصادی، فناورانه، سیاسی و فرهنگی)؛ ۲- جعبه سیاه خریدار: شامل فرآیند تصمیم خریدار و مشخصه‌های خریدار و ۳- واکنش‌های خرید: مانند انتخاب کالا، انتخاب برند تجاری کالا، انتخاب فروشنده، زمانبندی خرید و مبلغ خرید است (مال امیری به نقل از کاتلر و آرمسترانگ، ۱۳۹۲). مدل‌های مصرف به دو دسته مدل مصرف منزلت‌گرا و مدل مصرف بدون توجه به نقش نیز تقسیم شده است. افراد منزلت‌گرا در خرید و مصرف محصولات توجه زیادی به نظرها و عقاید دیگران دارند(شائو و شور[۶]، ۱۹۹۸؛ اوکاس[۷]، ۲۰۰۱ ؛ فریمن و لو[۸] ، ۲۰۰۷) و به دنبال ارتقای مقام و موقعیت اجتماعی خود،کسب وجهه و هویتی متمایز با دیگرانند. در مدل مصرف بدون توجه به نقش، مصرف‌کننده در خرید و مصرف محصول متمرکز برکارکرد کالاها و خدمات است (گلداسمیت و همکاران[۹]،۲۰۰۷). در این مدل مصرف‌کنندگان معمولاً اعتماد به نفس زیادی دارند، پس در خرید و مصرف محصولات به کارکرد و مفیدبودن محصول‌ها توجه می‌کنند (کهله[۱۰]، ۱۹۹۵) و در فرآیند خرید و مصرف محصول‌ها نسبت به نظرها و انتظارهای دیگران توجه چندانی چندانی ندارند (گلد اسمیت و کلارک[۱۱]، ۲۰۰۸).

ایده‌های اصلی داده‌کاوی[۱۲]برای مدیریت ارتباط با مشتری[۱۳]این است که داده‌های قدیمی حاوی اطلاعاتی اند که در آینده مفید واقع می‌شوند چراکه رفتار مشتریانی که در داده‌های شرکت نشان داده شده، تصادفی نیستند بلکه نیازهای متفاوت، تمایل‌ها، ترجیح‌ها و عملکردهای مشتریان را نشان می‌دهد. هدف داده‌کاوی یافتن الگوهایی در داده‌های پیشین است که آن نیازها، تمایل‌ها را روشنتر می‌نماید. تفکیک علایم مفید از موارد غیر مفید یعنی تشخیص الگوهای اساسی در بطن متغیرهای به ظاهر تصادفی یکی از نقش‌های مهم داده‌کاوی است. داده‌کاوی بررسی، تجزیه و تحلیل مقادیر عظیمی از داده به منظورکشف الگوها و قوانین پنهان و معنی‌دار درون داده‌ها اطلاق می شود. داده‌کاوی به دو گروه هدایت شده و هدایت نشده تقسیم می شود که در داده‌کاوی هدایت شده پژوهشگر دارای متغیر هدف خاص و از پیش تعیین‌شده بوده درحالی‌که در داده‌کاوی هدایت نشده[۱۴]پژوهشگر به دنبال یافتن الگوها یا شباهت‌هایی بین گروه‌های داده‌ها، بدون داشتن متغیر هدف خاص و یا مجموعه‌ای از دسته‌ها و الگوهای از پیش‌تعیین شده است. داده‌کاوی با ساختن مدل‌ها ارتباط دارد. یک مدل اساساً به الگوریتم‌ها یا مجموعه‌ای از قوانین گفته می‌شود که مجموعه‌ای از ورودی‌ها را که معمولاً به شکل زمینه‌هایی در پایگاه داده‌های سازمان است با هدف یا مقصد خاصی مرتبط می‌نماید. برای تبدیل یک مسأله کسب و کار و

تجارت به یک مسأله داده‌کاوی باید آنرا به یکی از شش فعالیت اصلی داده‌کاوی تبدیل کرد: ۱- دسته بندی[۱۵]، ۲- تخمین، ۳- پیش بینی[۱۶]، ۴-گروه‌بندی شباهت، ۵- خوشه‌بندی[۱۷]، ۶- توصیف و نمایه سازی[۱۸]؛ که سه دسته اول از نوع داده‌کاوی هدایت شده بوده و دسته چهار و پنج داده‌کاوی غیر هدایت شده است و دسته آخر به داده‌کاوی هدایت شده و هدایت نشده تقسیم می‌شود. از جمله خصوصیات مهم فرآیند داده‌کاوی، پویا و دینامیک بودن آن بوده چراکه این فرآیند به هیچ وجه ایستا نیست. سازمان‌ها با انجام فرآیند داده‌کاوی دینامیک همواره پویا مانده و به صورت پیوسته از نتایج دانش داده‌کاوی بهره‌مند می‌شوند (شهرابی،۱۳۹۴، ۷۸- ۷۹).

خوشه‌بندی یک روش یادگیری بدون نظارت است که روی دسته‌های از قبل تعریف شده به عنوان هدف تکیه ندارد. در واقع خوشه‌بندی شکلی از یادگیری به‌وسیله مشاهده‌هاست. در روش k میانگین ابتدا تعداد خوشه‌ها (k) توسط پژوهشگر تعیین می‌شود. آنگاه از میان داده‌های N مشتری، k مشتری به عنوان مراکز ابتدایی خوشه‌ها انتخاب شده و بقیه مشتریان برحسب نزدیکی اقلیدسی به این مراکز تخصیص می‌یابند. سپس مراکز جدید خوشه‌ها به صورت میانگین مقادیر هر خوشه محاسبه شده و هر مشتری مطابق با فاصله اقلیدسی‌اش به این مراکز جدید تخصیص می یابد. این فرآیند هنگامی‌که هیچ انتساب جدیدی وجود نداشته باشد متوقف می شود (بلت برگ و همکاران[۱۹] ، ۲۰۰۸).

شبکه عصبی خودسازمانده[۲۰]بر اساس ساختار و عملکرد مدل کوهونن[۲۱]که مدلی بدون نظارت است بنا نهاده شده است. علت استفاده از این شیوه در این پژوهش، تأکید برخی از پژوهش‌ها بر کارایی بهتر آن نسبت به روش رایج kمیانگین است (حنفی زاده و میرزازاده، ۲۰۱۱؛ چیو و همکاران[۲۲]، ۲۰۰۹؛ بلتبرگ و همکاران، ۲۰۰۸).

خوشه‌بندی دو مرحله‌ای[۲۳]بر پایه خوشه‌بندی سلسله مراتبی و شامل دو مرحله است در مرحله اول که پیش خوشه‌بندی نامیده می‌شود، چندین خوشه کوچک شکل گرفته و در مرحله دوم که خوشه‌بندی سلسله مراتبی گفته می‌شود، خوشه‌های کوچک مرحله اول دوباره در هم ادغام و چند خوشه بزرگتر شکل می‌گیرد (چیو و همکاران [۲۴]، ۲۰۰۱).

شاخص‌هایی که معمولاً برای تعیین کیفیت و میزان خلوص[۲۵]خوشه‌بندی و دسته‌بندی داده‌ها استفاده می‌شود عبارتند از: ۱- خلوص و پراکندگی،۲- امتیاز جینی[۲۶](پراکندگی جمعیت)، ۳- آنتروپی[۲۷](بهره اطلاعاتی)، ۴- نسبت بهره اطلاعاتی[۲۸]، ۵- آزمون مربع کای(χ^2). انتخاب روش مناسب اندازه‌گیری خلوص بستگی به دسته‌ای یا عددی بودن متغیر هدف دارد.

امتیازجینی یا پراکندگی جمعیت؛ احتمال قرارگیری دو مورد انتخاب شده تصادفی از یک جمعیت یکسان را در یک دسته نشان می‌دهد. برای یک جمعیت خالص این احتمال برابر یک است. اندازه گیری جینی یک گره به صورت ساده: مجموع مربع نسبت‌های دسته‌هاست. گره خالص ضریب جینی یک دارد و گره متوازن امتیاز جینی (۰/۵) خواهد داشت که در این شاخص هر چقدر مقدار این امتیاز به یک نزدیکتر شود خوشه ایجاد شده کیفیت بهتری داشته است(شهرابی، ۱۳۹۴). اگر یک برگ کاملاً خالص باشد آنگاه دسته‌های این برگ را به راحتی این‌گونه می‌توان توصیف کرد که همگی آن‌ها در یک دسته‌بندی جای می‌گیرند. مقدار خالص بودن بین یک (خوشه‌بندی بهینه) و صفر (خوشه‌بندی بد) است که در این پژوهش از فرمول۱برای محاسبه میزان خلوص دسته‌های ایجاد شده به روش خوشه‌بندی استفاده شده است (ویسی،۱۳۹۶).

1. $Purity(Cluster\ , Class) = \frac{1}{N\sum_{k=1}^{K} Max_j[\text{Clusterk}\cap\text{Class j}]}$

در فرمول۱،K نشان‌دهنده تعداد خوشه‌ها وj نشان‌دهنده تعداد دسته‌هاست وN نشاندهنده تعدادکل داده‌هاست.

آنتروپی میزان بی نظمی یک سیستم است. آنتروپی یک گره خاص در درخت تصمیم عبارت است از جمع نسبت‌های داده‌های متعلق به یک دسته خاص برای تمام دسته‌هایی که در گره نشان داده شده‌اند که در لگاریتم پایه دو آن نسبت ضرب شده است (معمولاً در (۱-) ضرب می‌کنند تا عددی مثبت بدست آید) (ویسی،۱۳۹۶، ۲۴). در این پژوهش از فرمول ۲ برای بدست آوردن مقدار آنتروپی هر روش خوشه-بندی استفاده شده است. در مقایسه با ضریب جینی، شاخص آنتروپی گره‌هایی را ترجیح می‌دهد که خالصترند حتی اگر کوچکتر شوند.

2. $H(cluster) = -\sum_{K=1}^{k} p(Cluster_k)\log(Cluster_k) = \sum_{k=1}^{k} \frac{|(Cluster_k)|}{N}\log\frac{|(Cluster_k)|}{N}$

در فرمول۲، P نشان‌دهنده مقدار فراوانی نسبی (احتمال) هر خوشه است و K نشان‌دهنده تعداد خوشه‌ها بوده و N نشاندهده تعداد کل داده‌هاست.

روش دیگر محاسبه کیفیت اطلاعات بدست آمده از خوشه‌بندی استفاده از معیار خارجی[۲۹] اطلاعات متقابل نرمال شده[۳۰]است. این شاخص نشان می‌دهد آیا اطلاعات مفید و غیر تصادفی از خوشه‌بندی بدست آمده است یا خیر (عدد این شاخص بین صفر و یک است).مقدار یک، نشاندهنده آن است که هرخوشه دقیقاً بیان‌کننده یک دسته است؛ مقدار صفر (خوشه‌بندی تصادفی) یعنی دانستن خوشه کمکی

به افزایش اطلاعات ما از دسته نمی کند؛ شاخص اطلاعات متقابل نرمال شده به تعداد خوشه‌ها حساس است که نحوه محاسبه آن به شرح فرمول‌های ۳ و ۴ است (ویسی، ۱۳۹۶، ۲۴- ۲۵).

3. $NMI(Cluster, Class) = I(Cluster, Class)/([H(Cluster) + H(Class)\ ]/2)$

4. $I(Cluster, Class) = \sum_{k=1}^{k} \sum_{j=1}^{j} P(Cluster_k \cap class_j) * \log \frac{P(Cluster_k \cap class_j)}{P(cluster_k) * p(Classj)}$

در فرمول ۳ و۴، NMI شاخص اطلاعات متقابل نرمال شده‌بوده و P_{kj} احتمال(فراوانی نسبی)دسته j در خوشهK است، K نشان‌دهنده تعداد خوشه‌هاست و I مقدار اطلاعات متقابل قبل از نرمال‌کردن است.

یکی از تکنیک‌های ساده و کاربردی برای بخش‌بندی[۳۱] مشتریان، تحلیل ارزش مشتریان با مدل تازگی، تکرار و ارزش پولی خرید[۳۲]است. این تکنیک کمک می‌کند مشتریانی که برای شرکت‌ها سودآورترند را شناسایی کرده و آن‌ها را به طبقه‌های مختلف پلاتینی، طلایی، نقره‌ای، برنزی و... تقسیم نمود. با این کار می‌توان برای هر طبقه، سیاست قیمت‌گذاری و پرداخت خاصی در نظر گرفت. تازگی، تکرار و ارزش پولی خرید، حروف اول سه متغیرند. تازگی خرید: فاصله زمانی بین آخرین خرید صورت گرفته تا امروز؛ تکرار خرید: تعداد خریدهایی که مشتری در دوره زمانی خاص انجام داده است؛ ارزش پولی خرید: مقدار پولی که مشتری در دوره زمانی خاص برای خرید اختصاص داده است. در این پژوهش با استفاده از فرمول ۵ و بر اساس ضرایبی که از نظر مشورتی خبرگان صنعت برای تازگی، تکرار و ارزش پولی خرید مشتری با روش مقایسه‌های زوجی تعیین شده، در نهایت مقدار WRFM بدست آمده است تا خوشه‌بندی مشتریان بر این اساس صورت بگیرد. در این پژوهش برای متغیرهای تازگی، تکرار و ارزش پولی خرید به ترتیب مقدارهای (۰/۵۷۱)، (۰/۱۴۵) و (۰/۲۸۴) در نظرگرفته شده است.

5. $RFMScore = RMonetaryScore + RRecencyScore + RFrequency\ Score.$

در فرمول ۵، Monetary Score نمره یا مقدار ارزش پولی خرید است که در این پژوهش بر اساس متوسط خرید فاکتور مشتری تعیین می‌شود. Recency Score مقدار تازگی (تأخر) خرید مشتری است که در این پژوهش بر اساس مدت زمانی سپری شده از آخرین خرید سال۱۳۹۶بدست می آید و در نهایت Frequency Score نمره تعداد خرید مشتری است که در این پژوهش بر اساس تعداد کل خط فاکتور خرید انجام شده توسط مشتری بدست می آید.

پیشینه پژوهش

پیشینه پژوهش‌های انجام شده داخلی در زمینه داده‌کاوی

رییسی وانانی و همکاران(۱۳۹۹)، در پژوهشی با عنوان مدلی برای بخش‌بندی یادگیرندگان و بهبود عملکرد آموزشی با استفاده از الگوریتم‌های داده‌کاوی؛ داده‌های مربوط به دانش‌پذیران بین المللی، را بر

اساس روش تحقیق علم طراحی و با استفاده از روش‌های داده‌کاوی مورد بررسی قرار داده‌اند. در این راستا داده‌های تحصیلی و غیر تحصیلی یادگیرندگان درسه دسته خانوادگی، حمایتی و رفتار تحصیلی با استفاده از داده‌کاوی، خوشه‌بندی کرده‌اند و پس از اعتبارسنجی خروجی الگوریتم‌ها توسط شاخص‌های مرتبط و تعیین تعداد خوشه بهینه در هر بخش، خوشه‌ها را نام‌گذاری و تحلیل کرده اند. تحلیل خوشه‌های شناسایی شده، نشان‌دهنده تجربه موفقیت یا شکست تحصیلی دانش پذیران و ریشه‌های عملکرد مؤثر در هر بخش داشته است و روش نام‌گذاری ارایه شده، یک روش نوین و قابل استفاده در اغلب مراکز آموزشی جهت تفکیک و تبیین عملکرد آموزشی بوده است (رییسی وانانی، وانانی و تقوی فرد، ۱۳۹۹). رنگریز و بایرامی(۱۳۹۸)، در پژوهش خود با عنوان تأثیر مدیریت ارتباط با مشتری الکترونیکی بر وفاداری مشتریان با استفاده از تکنیک‌های داده‌کاوی، به دنبال بررسی اثر E-CRM بر وفاداری مشتریان بانک ملت بوده‌اند؛ که در پژوهش خود از روش‌های Kمیانگین، و شبکه عصبی با الگوریتم پس انتشار و مدل LRFM به کمک نرم افزارهای اکسل و متلب استفاده کرده‌اند که نتایج پژوهش حکایت از وجود رابطه غیرخطی بین وفاداری مشتریان، E-CRM و مؤلفه‌های LRFM داشته است (رنگریز و بایرامی شهریور، ۱۳۹۸). مشایخی و همکاران (۱۳۹۷)، از الگوریتم‌های داده‌کاوی برای تحلیل سودمندی دوربین‌های ثبت تخلف راهنمایی و رانندگی بزرگراه تهران-کرج استفاده کرده‌اند و برای مدلسازی پژوهش خود از سری زمانی و رگرسیون نیز استفاده کرده‌اند؛ نتایج پژوهش آن‌ها حکایت از آن داشته که مدل‌های برازش شده سری زمانی از خطای پیش‌بینی کمتری نسبت به مدل‌های رگرسیونی برخوردار بودند و از الگوریتم‌های داده‌کاوی می‌توان برای بهبود مدیریت شهری با هدف کاهش تخلف‌ها و تصادف‌ها استفاده کرد (مشایخی، نوراله و دقیق نژاد، ۱۳۹۷). محمودی و همکاران (۱۳۹۶)، با استفاده از رویکرد داده‌کاوی به کمک الگوریتم‌های Apriori وID3 به بررسی و تعیین عوامل مؤثر بر بروز سرطان معده بر روی۴۹۰مراجعه کننده به بیمارستان امام رضای تبریز پرداخته‌اند. هدف این پژوهشگران استخراج قوانین و الگوهای پنهان از داده‌های سرطان معده بوده که به کمک آن بتوانند بدون نیاز به روش‌های تشخیصی، احتمال ابتلا به بیماری را تشخیص داده و نیز عوامل مؤثر در این بیماری را شناسایی کنند. دقت الگوریتم مذکور(۸۵/۵۶) برآورد شده که نشان از قدرت خوب در پیش‌بینی سرطان معده داشته است (محمودی، میرزایی و محمودی، ۱۳۹۶). کتابی و همکاران (۱۳۹۵)، برای بخش‌بندی بازار و طبقه‌بندی مشتریان بازارهای الکترونیکی به کمک روش‌های نگاشت‌های خودسازمانده[۳۳]، Kمیانگین و خوشه‌بندی دو مرحله‌ای در قالب رویکردی تجمیعی الگوریتمی یکپارچه ارایه کرده‌اند و از معیار سیلوئت[۳۴](ضریب نیمرخ) برای بررسی صحت و اعتبار نتایج و همچنین از شاخص‌های تازگی، تکرار و ارزش پولی برای خوشه‌بندی استفاده

کردند. از روش دلفی فازی برای جبران کمبود داده‌های پژوهش بهره برده‌اند سپس بر پایه نتایج حاصل از خوشه‌بندی و داده‌های گردآوری شده از طریق مصاحبه تلفنی، مشتریان با استفاده از روش‌های آماری تحلیل ممیزی و رگرسیون لجستیک و همچنین روش‌های یادگیری ماشینی، شبکه‌های عصبی مصنوعی[۳۵] و ماشین‌بردارپشتیبان[۳۶] طبقه‌بندی شده‌اند. نتایج پژوهش مذکور نشان می‌دهد، روش طبقه‌بندی برنامه‌ریزی خطی درحین سادگی و شفافیت نتایج دقیق‌تری به ویژه برای بازاریابان در بر داشته است (کتابی و همکاران، ۱۳۹۵). عاشوری و همکاران (۱۳۹۵)، به منظور شناسایی الگویی برای تشخیص وضعیت اهداکنندگان خون، از روش خوشه‌بندی بهره برده‌اند؛ جامعه آماری پژوهش آن‌ها داده‌های سازمان انتقال خون بیرجند در ماه‌های خرداد تا شهریور۱۳۹۲بوده است. این پژوهشگران برای تحلیل داده‌ها از نرم افزار کلمنتاین[۳۷]استفاده کرده‌اند؛ در پژوهش مذکور ابتدا از خوشه‌بندی دومرحله‌ای و سپس الگوریتم‌های (C5.0)، (C&RT)[۳۸]، (CHAID)[۳۹] و(QUEST)[۴۰] استفاده شده که نتایج مقدار صحت بدست آمده از اجرای الگوریتم‌های (C5.0) درخت (C&RT)، (CHAID) وQUEST به ترتیب(۰/۹۹۸)، (۰/۹۹۶)، (۰/۹۹۳)، (۰/۸۹۱) بوده است. مقادیر به‌دست آمده برای شاخص‌های حساسیت[۴۱]، شفافیت، صحت[۴۲]، دقت[۴۳]، شاخصF، میانگین هندسی، نرخ مثبت غلط، نرخ منفی غلط و نرخ خطا برای مدل (C5.0) نشان‌دهنده عملکرد بهتر این الگوریتم نسبت به سایر الگوریتم‌ها بوده است و تأثیرگذارترین شاخص‌ها در تولید مدل، دسته فشارخون، وضعیت اهدای خون و دمای بدن بوده است (عاشوری، محمدی و حسینی، ۱۳۹۵). فیروزی جهان تیغ و عامری (۱۳۹۴)، با انجام پژوهشی توصیفی تحلیلی به کمک روش خوشه‌بندی K میانگین، به بررسی ویژگی‌های بیماران مبتلا به سل پرداخته‌اند. پژوهش مذکور با هدف دسته‌بندی و پیداکردن ارتباط بین ویژگی‌های بالینی ودموگرافیک بیماران مختلف انجام شده‌است. پژوهش مذکور روی۶۰۰ بیمار مرکز تحقیقات سل بیمارستان مسیح دانشوری انجام شده است و پژوهشگران برای دسته‌بندی و تعیین شاخص‌های مشترک بین بیماران از الگوریتم‌های خوشه‌بندی kمیانگین و قوانین باهم‌آیی (انجمنی)[۴۴] به کمک نرم افزارکلمنتاین نسخه ۱۴استفاده کرده اند همچنین به کمک شاخص دان[۴۵]، تعداد سه خوشه به عنوان خوشه بهینه انتخاب شده است. نتایج حاصل از این مطالعه، مهمترین عوامل شناسایی شده با استفاده از خوشه‌بندی را: هموگلوبین، سن، جنسیت، مصرف سیگار، مصرف الکل وکراتینین دانسته است همچنین با توجه به قوانین باهم‌آیی (انجمنی)، بیشترین ارتباط بین سرفه،کاهش وزن و سرعت رسوب گلوبول‌های قرمز دیده شده است(فیروزی و عامری، ۱۳۹۴). غفاری و سلماسی (۱۳۸۸)، به کمک روش خوشه‌بندی به شناسایی مشترکان تلفن همراه که تمایل به ترک شرکت دارند پرداخته‌اند تا بتوانند الگوهایی را برای شناخت

مشترکان تلفن همراه پیداکنند؛ برای این منظور فرایند داده‌کاوی را بر روی مجموعه داده‌هایی شامل حدود یکصد هزار سابقه و ۳۱ مشخصه بر اساس روش استاندارد داده‌کاوی [۴۶]با استفاده از نرم افزارکلمنتاین اجرا نمودند که خوشه‌های ایجادشده بیانگر الگوهای موجود درداده‌ها بودندکه نتایج پژوهششان نشان داد مشترکانی که از طرح‌های رنگی استفاده می‌کنند یا آن‌هایی‌که از پیام کوتاه بیشتر استفاده می‌کنند کمتر تمایل به ترک شرکت دارند (غفاری و سلماسی، ۱۳۸۸).

پیشینه پژوهش‌های انجام شده خارجی در زمینه داده کاوی

ویشنوواردهان و لکشیمپادماجا[۴۷](۲۰۱۸)، درپژوهشی با عنوان بهبود عملکرد طبقه‌بندی با استفاده از الگوریتم انتخاب تصادفی ویژگی‌های زیرمجموعه (RSFS)[۴۸]به منظورداده‌کاوی، الگوریتمی اصلاح شده را با روش K نزدیکترین همسایگی[۴۹]ترکیب کرده که نتایج پژوهش نشان از بهبود پایداری مدل و دقت طبقه‌بندی داده‌کاوی داشته است و از پایگاه داده‌های ماهیتاً علمی که برای تشخیص سرطان به‌کاررفته استفاده کرده‌اند (ویشنوواردهان و لکشیمپادماجا، ۲۰۱۸). انزوا و لاورنس[۵۰] (۲۰۱۷) درپژوهشی به منظور پیش‌بینی تولید چای درکنیا به استفاده گام‌به‌گام از خوشه‌بندی و قوانین انجمنی پرداخته‌اند. در این پژوهش از قوانین انجمنی و بر اساس ارقام ارایه شده مشخص شده است که میزان تولید آینده به طور عمده توسط سه متغیر اصلی توصیف خواهد شد. این پژوهشگران داده‌کاوی را در رشته کشاورزی به عنوان یک تعهد جدید برای رفاه عمومی مردم معرفی کرده‌اند (انزوا و لاورنس، ۲۰۱۷). فمینا بهری و الایدوم[۵۱](۲۰۱۵)، در پژوهشی که در زمینه داده‌کاوی انجام داده‌اند به دنبال ارایه چارچوبی کارآمد برای داده‌کاوی مدیریت ارتباط با مشتری با هدف پیش‌بینی رفتار مشتریان بوده‌اند؛ در مدل مذکور از دو الگوریتم طبقه بندی (Naïve Byes) و شبکه عصبی به منظور بهبود و ارتقای فرآیند تصمیم‌گیری برای حفظ مشتریان با ارزش بهره برده‌اند ونتایج حکایت ازدقت وکارآمدی بیشتر شبکه عصبی داشته است (فمینا بهاری، الائیدوم، ۲۰۱۵). هانگ[۵۲](۲۰۱۱)، در مقاله‌ای با استفاده از روش استاندارد داده‌کاوی برای کمک به شناخت بهتر وام‌دهی بانک‌ها، مشتریان را بر حسب داده‌های دموگرافیک (جمعیت شناختی) و بازپرداخت اقساطشان با استفاده از شبکه عصبی مصنوعی خوشه‌بندی نموده و سپس جهت توسعه مدل از روش درخت تصمیم به پیش‌بینی مدلی برای اعطای اعتبار به مشتریان پرداخته است (هانگ، ۲۰۱۱). یوهانگ گیو[۵۳](۲۰۱۰)، در مقاله خود به شناسایی الگوهای رفتاری خرید دارندگان کارت اعتباری به کمک خوشه‌بندی با استفاده از داده‌های تلفیقی[۵۴]پرداخته است. در این مقاله به جای استفاده از داده‌کاوی سنتی که از مقطعی از داده‌ها استفاده می‌کند، داده‌کاوی را بر روی مجموعه‌ای از متغیرها در طول زمان اجرانموده است (یوهانگ گیو، ۲۰۱۰). لیو و همکاران[۵۵](۲۰۱۰)، در پژوهش خود از برخی روش‌های

داده‌کاوی برای تحلیل رفتار مشتریان جهت ساخت الگوی رفتاری مشتری و شناخت مشتریان وفادار برای تصمیم‌گیری‌های استراتژیک در بازاریابی آژانس هواپیمایی استفاده کرده است (لیو و تی ژنگ، ۲۰۱۰). ناگای و همکاران[۵۶](۲۰۰۹)، در مقاله مروری خود به ادبیات موضوعی استفاده از داده‌کاوی در مدیریت ارتباط با مشتری پرداخته است؛ در پژوهش مذکور تاریخچه و پیشینه پژوهش در بازه زمانی سال‌های ۲۰۰۰- ۲۰۰۶ میلادی با پوشش۲۴ مجله و ارایه شمای کلی از طبقه‌بندی بالغ بر۹۰۰ مقاله که به طورمستقیم به مدیریت ارتباط با مشتری مربوط بوده بررسی شده است. در نهایت مقاله‌ها به ۹ طبقه فرعی از عناصر مدیریت ارتباط با مشتری تحت روش‌های مختلف داده‌کاوی بر اساس تمرکز اصلی هر مقاله طبقه بندی شدند. یافته‌ها نشان از آن داشته که زمینه پژوهش حفظ مشتری بیشترین توجه پژوهشگران را به خود جلب نموده است که از این تعداد بیشتر به بازاریابی رو در رو و برنامه‌های وفاداری مربوط می‌شود. همچنین مدل‌های طبقه بندی و قوانین انجمنی دو مدل معمول برای داده‌کاوی در مدیریت ارتباط با مشتری بوده‌اند (ناگای، خیو و چائو، ۲۰۰۹).

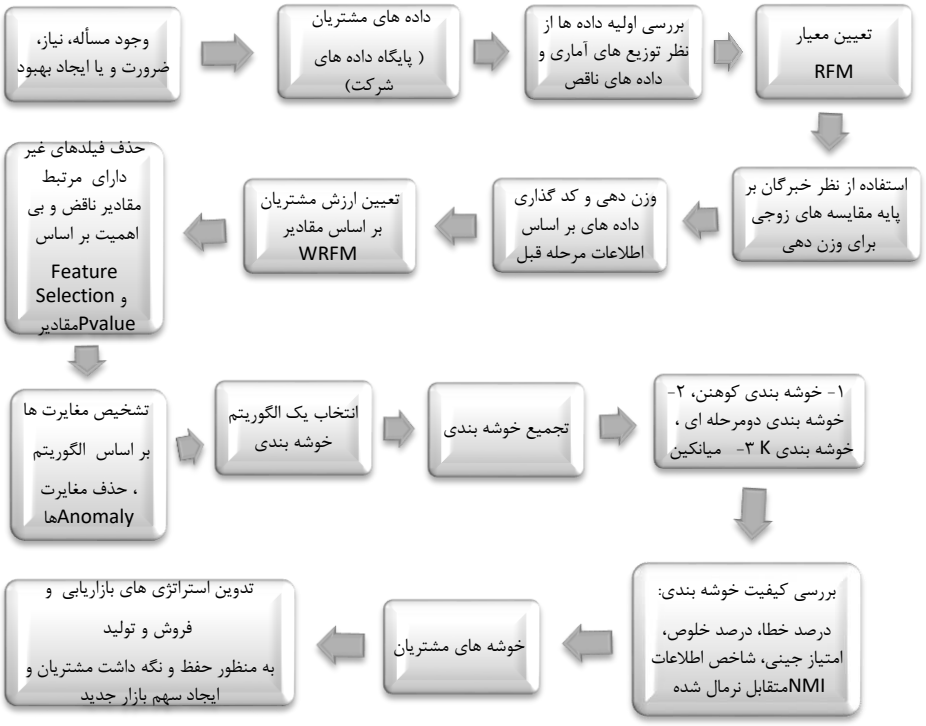
با توجه به بررسی پیشینه پژوهش‌های انجام شده در زمینه داده‌کاوی آنچه مشخص است پژوهشی که روش‌های مختلف خوشه‌بندی را به منظور بخش‌بندی مشتریان بازار محصولات بهداشتی و آرایشی با توجه معیارهای گوناگون میزان صحت و کیفیت دسته‌بندی، بررسی کرده و سپس به ارایه روش خوشه‌بندی بهینه بر این اساس پرداخته، مشاهده نشده است؛ همچنین با توجه به روند رو به رشد مصرف محصول‌های بهداشتی و آرایشی در سطح کشور و انباشت اطلاعات مشتریان در این صنعت، انجام پژوهشی که بتواند الگوهای نهان و آشکار خرید مشتریان را از دل این داده‌های انباشته‌شده استخراج کند تا به منظور بخش‌بندی بهتر مشتریان مورد استفاده قرارگیرد بیش از پیش ضرورت می‌یابد.

روش پژوهش

این پژوهش کاربردی- توسعه‌ای و از نظر ماهیت جنبه اکتشافی- تحلیلی دارد که هدف آن توسعه کاربرد داده‌کاوی بر بستر بازاریابی به منظور دستیابی به الگوهای رفتاری مشتریان است تا بتواند به ارایه الگویی برای خوشه‌بندی مشتریان دست یابد. با توجه به ماهیت اکتشافی پژوهش حاضر در ابتدا فرضیه‌های مشخصی برای پژوهش نمی‌توان متصور شد بلکه در حین اجرای پژوهش در ذهن پژوهشگر ایجاد می‌شود. جامعه آماری این پژوهش را مشتریان استان تهران شرکتی فعال در عرصه تولید محصول‌های بهداشتی و آرایشی از گروه صنعتی گلرنگ، که درقالب یکی از گروه‌های مشتریان شامل: خرده فروشی، تعاونی مصرف و محلی از محصول‌های بهداشتی و آرایشی شرکت در بازه سال‌های ۱۳۹۶-۱۳۹۷استفاده کرده‌اند تشکیل می‌دهد. نمونه آماری این پژوهش که شامل ۶۵۵۳۴ نمونه

(مشتری) بوده با روش نمونه‌گیری هدفمند در دسترس از پایگاه داده‌های مشتریان شرکت که به صورت اکسل است از شعبه جنوب شرق استان تهران شامل: مناطق ۸، ۱۲، ۱۳، ۱۴، ۱۵ (شهرداری تهران) و شهرهای دماوند، گرمسار، قرچک، ورامین، پاکدشت، ایوانکی می‌شوند انتخاب شده‌اند. برای گردآوری اطلاعات پیشینه پژوهش از روش مطالعه‌های کتابخانه‌ای و با بهره‌گیری از موتورهای جستجوگر مقاله‌ها، کتاب‌ها و مستندها، پایان‌نامه‌ها و نظر کارشناسان مرتبط با موضوع پژوهش استفاده شده است سپس در مرحله آماده‌سازی داده‌ها برای انجام داده‌کاوی و همچنین برای تعیین مقدار وزن متغیرهای تازگی، تکرار و ارزش پولی خرید از نظر خبرگان صنعت و مدیران فروش شرکت به روش مقایسه زوجی (تحلیل سلسله مراتبی) استفاده شده؛ به این شکل که ابتدا از خبرگان و کارشناسان خواسته شده است تا هریک جداگانه نسبت به اهمیت هریک از معیارهای تازگی، تکرار و ارزش پولی خرید نظرهای خود را در قالب مقایسه دو به دو اعلام نمایند سپس بعد از بدست آوردن این وزن‌ها مقدار هریک از این مؤلفه‌ها را محاسبه و تعیین کرده که این سه مؤلفه به همراه متغیر جدیدی که از حاصل جمع موزون این سه متغیر بدست می‌آید به داده‌های پژوهش اضافه شده است. مشتریان (رکوردها) بر اساس بازه‌هایی که برای حاصل جمع WRFM آنها ایجاد می‌شود به چهارگروه مشتری تقسیم شده‌اند. اگر نمره این معیار بالاتر از ۴/۵ باشد مشتریان طلایی محسوب شده و اگر در بازه ۳ تا ۴/۵ مشتری نقره‌ای، در بازه ۱/۵تا۳ مشتری معمولی و کمتر از ۱/۵ مشتری پر ریسک محسوب شده‌اند. بعد از این مرحله نوبت به غربالگری[۵۷]منطقی و اولیه داده‌ها می‌رسد. مؤلفه‌هایی چون کدشناسایی مشتری، طول و عرض جغرافیایی و...که از نظر منطقی تأثیری در پیش‌بینی ندارند حذف می‌شوند تا بتوان دقت داده‌کاوی را بالابرده و هم از هدررفت زمان جلوگیری نمود. سپس با استفاده از نرم افزار کلمنتاین نسخه۱۲و به کمک گره گزینش مؤلفه‌های مهم [۵۸](غربالگری مرحله دوم) به منظور حذف متغیرهایی که اثری درپیش بینی متغیر نوع مشتریان ندارد اقدام می‌شود که در نهایت۲۰ مؤلفه با توجه به مقدار پی [۵۹]انتخاب شده‌اند (با اهمیت بوده‌اند). بعد از این مرحله به کمک گره تشخیص مغایرت[۶۰] از بین نمونه‌های اولیه، رکوردهای (نمونه‌های) مغایر تشخیص داده شده را حذف کرده و سپس از۷۰درصد آنها برای آموزش داده‌کاوی و از مابقی آن برای تست صحت و اعتبارسنجی مدل این پژوهش استفاده شده است. برای تجزیه و تحلیل اطلاعات و داده‌های مشتریان از روش‌های مختلف و ترکیبی داده‌کاوی مثل خوشه‌بندی دومرحله‌ای، شبکه‌های عصبی خودسازمانده، kمیانگین استفاده شده است و در جای جای این پژوهش خصوصاً در زمان تجزیه و تحلیل اطلاعات و همچنین به هنگام استفاده از نوع روش مناسب برای داده‌کاوی از نظرخبرگان این صنعت و نخبگان علمی نیز استفاده شده و روایی و پایایی این پژوهش مورد تأیید قرارگرفته است؛ همچنین برای

بررسی میزان دقت، خلوص وکیفیت خوشه‌بندی از معیارهای میزان خطا، امتیازجینی، درصد خلوص و شاخص اطلاعات متقابل نرمال شده استفاده شده است. با توجه به مطالب بیان شده الگویی که برای دسته‌بندی مشتریان استفاده شده است و مدل مفهومی که این پژوهش بر اساس آن طراحی شده به شرح شکل ۱ است.



شکل ۱- مدل مفهومی پژوهش حاصل از متدولوژی پژوهشگر

یافته‌های پژوهش

یافته‌های حاصل از تعیین مؤلفه‌های مهم

از بین فیلدهای اولیه پایگاه داده‌های مشتریان، کدشناسایی مشتری، طول و عرض جغرافیایی، فیلد سال(۱۳۹۶)، پوشش مشتری و... و فلیدهای گروه مشتریان سطح یک، گروه مشتریان سطح دو، گروه مشتریان سطح سه به خاطر اینکه این فیلدها دارای رکوردهای زیادی در یک دسته‌اند که برای متغیر هدف تعیین شده این پژوهش اطلاعات ارزشمندی تولید نکرده‌اند حذف شده‌اند (بر اساس تشخیص الگوریتم مؤلفه‌های مهم). همچنین فیلد روز خرید مشتریان با مقدار اهمیت (۰/۹۰۳) در مرز بین با

اهمیت و بی اهمیت بودن قرارگرفته است. ۲۰ فیلد با اهمیت این پژوهش (به تشخیص الگوریتم مؤلفه‌های مهم و بر اساس مقدار پی) عبارتند از: متغیرهای شهر، وضعیت مشتری (مشتری فعال، موقت، غیرفعال و لیست سیاه)، منطقه شهری، مدت زمان سپری شده از آخرین خرید سال ۱۳۹۶(هرچقدر این مقدار کمتر است یعنی مشتری به تازگی خرید کرده است)، ماه خرید، گروه بازار محصول‌ها، مقدار فاکتور فروش، تعداد خط فاکتور فروش، مقدار خرید قطعی (برحسب ریال)، متوسط فروش هر فاکتور، کل مبلغ تخفیفات ریالی، مدت زمان سپری شده از آخرین خرید، خالص تعداد فروش، مقدار برگشتی از فروش، خالص تخفیف‌های کالایی (برحسب ریال)، میانگین کل تخفیفات (برحسب ریال)، مقدار تازگی، تکرار و مقدار ارزش پولی خرید موزون، مقدار مجموع تازگی، تکرار و ارزش پولی خرید وزن‌دهی شده (مقدار اهمیت برابر یک منهای مقدار پی هست که این مقدار احتمال، از آزمون تی- استیودنت یا مربع‌کای بدست می‌آید. در مورد هر دو نوع فیلد بازه‌ای و گسسته هرچه مقدار اهمیت بیشتر باشد امکان تصادفی بودن تغییرهای مقداریک فلید در بین خوشه‌ها کمتر و امکان وجود یک تفاوت اساسی بین خوشه‌ها بیشتر خواهد بود. مقدار اهمیت حاصل به صورت یک آزمون آماری بدست می‌آید که برای متغیرهای طبقه‌ای آزمون مربع کای است. فرض صفر این است که توزیع طبقه‌های داخل خوشه‌ها برای تمامی خوشه‌ها یکسان است. اگر این متغیر طبقه‌ای واقعاً در خوشه‌بندی مؤثر است فرض صفر رد و سطح اهمیت به یک نزدیک می‌شود و برای متغیرهای پیوسته از آزمون تی- استیودنت استفاده می‌شود. فرض صفر به صورت برابری میانگین متغیر درخوشه‌های مختلف بیان شده است. اگر این متغیر پیوسته بوده و واقعاً درخوشه‌بندی مؤثر باشد فرض صفر رد و سطح اهمیت به یک نزدیک می شود).

یافته‌های روش تشخیص مغایرت

بر اساس تنظیم‌های اولیه مبنی بر مقدار برش بر پایه یک درصد بیشترین رکوردهای مغایر در داده‌های آموزشی و همچنین انتخاب دامنه حداقل یک و حداکثر پانزده گروه غالب بر این اساس خود سیستم به طور اتوماتیک مقدار شاخص مغایرت را (۲/۱۴۵۲۵) تخمین زده است (مقدار شاخص برابر نسبت انحراف گروه از میانگین خوشه‌ای است که به آن تعلق دارد. هر چقدر مقدار این شاخص بزرگتر باشد انحراف از میانگین نیز بیشتر است.

مواردی با مقدار شاخص بزرگتر از۲ می‌تواند کاندیدای خوبی برای مغایرت شود زیرا انحراف از میانگین حداقل دو برابر خواهد بود) (علیزاده و ملک محدی، ۱۳۹۳، ۱۲۶). دو گروه غالب در این پژوهش تشخیص داده شده که در گروه غالب شماره یک تعداد ۱۲۴۷۸رکورد که از این تعداد۱۵۰رکورد آن رفتار مغایر

داشته و در گروه غالب شماره دو، ۵۳۰۵۶ رکورد موجود بوده که ۵۰۵ رکورد آن رفتاری مغایر داشته است؛ در مجموع از۶۵۵۳۴رکورد مشتری ۶۵۵ رکورد (مشتری) رفتار مغایر داشته‌اندکه با حذف مشتریان مغایر تعداد ۶۴۸۷۹ رکورد (مشتری) باقی خواهد ماند. در گروه‌بندی اول و دوم تنها به توصیف آماری بعضی از اطلاعات جمعیت شناختی مشتریان اشاره می‌شود:

۱. درگروه غالب یک: بیشترین خرید توسط مشتریانی از منطقه ۱۵ با وضعیت غیر فعال بوده که بیشترین خرید خود را در تیر ماه انجام داده‌اند و بیشترین محصولی که خریداری کرده اند از بین ۲۶ قلم کالایی مایع‌دستشویی با فراوانی ۱۴/۱۴درصد بوده است، همچنین مقدار تازگی- تکرار- ارزش پولی موزون این گروه ۲/۱۸۷ با انحراف استاندارد ۰/۷۴۹ بوده است.

۲. در گروه غالب دو: بیشترین خرید توسط مشتریانی از منطقه ۸ با فراوانی۳۲/۸۴ درصد صورت گرفته است که خرید این مشتریان بیشتر در ماه بهمن با فراوانی ۱۰/۳۹درصد بوده است و اکثریت مشتریان در دسته مشتریان فعال شرکت بوده‌اند و همچنین بیشترین محصول مورد استفاده این گروه از بین۲۶ قلم کالایی مایع دستشویی با فراوانی ۱۳/۰۱درصد بوده است؛ مقدار تازگی- تکرار- ارزش پولی موزون این گروه ۳/۲۴۳ با انحراف استاندارد ۰/۷۷۲ بوده است.

نتایج این دو گروه غالب نشان می‌دهد که درگروه غالب یک با توجه به مقدارWRFM مشتریان غالب در این گروه، مشتریان معمولی با وضعیت غیر فعال بوده‌اند و در گروه غالب دو مشتریان نقره‌ای با احتمال۹۹/۵ درصد جزو مشتریان فعال بوده‌اندکه بیشتر این مشتریان در شهر تهران قرار داشته‌اند و در هر دوگروه مقدار تغییرهای انحراف استاندارد نشان ازپراکندگی و تنوع رفتاری مشتریان دارد.

با توجه به جدول۲ ذیل مشخص است که از بین ۶۵۵ مغایرت تشخیص داده شده ۳۷۴ مورد مربوط به مشتریان تهران، ۱۰۹ مورد مربوط به مشتریان پاکدشت، ۸۱ مورد مربوط به مشتریان ورامین،۶۰ مورد مربوط به مشتریان قرچک و۳۱ مورد مربوط به مشتریان پیشواست و شهرهای دماوند، گرمسار و ایوانکی بدون مغایرت بوده‌اند؛ همچنین بیشترین مغایرت در دسته مشتریان از آن مشتریان نقره‌ای است با ۲۹۴مورد، سپس مشتریان طلایی با ۲۶۸ مورد و بعد از آن مشتریان معمولی با ۶۵ مورد و مشتریانی که در دسته خاصی قرار نگرفته‌اند (دارای مقادیر خالی یا صفر در بعضی ازفیلدهای داده‌های اولیه خود بوده‌اند) با ۲۸مورد و در نهایت مشتریان پرریسک که هیچگونه مغایرتی در آن‌ها دیده نشده است.

جدول۲- مقایسه تغییرهای ایجاد شده با استفاده از تشخیص مغایرت بر روی نوع و شهر مشتری

	داده های اولیه (دسته های مشتری)						داده ها بعد از حذف مغایرت ها (دسته های مشتری)					
شهر	طلایی	نقره ای	معمولی	پرریسک	دسته ناقص	جمع کل	طلایی	نقره ای	معمولی	پرریسک	دسته ناقص	جمع کل
تهران	۲۲۴۰	۲۱۸۳۹	۲۱۳۴۲	۲۳۳۹	۵۵۸۹	۵۳۳۴۹	۲۰۹۶	۲۱۶۸۱	۲۱۲۹۰	۲۳۳۹	۵۵۷۳	۵۲۹۷۵
پیشوا	۷۱	۱۹۴	۱۲۷	۲۳	۸۳	۴۹۸	۵۶	۱۸۹	۱۲۷	۲۳	۸۰	۴۶۷
پاکدشت	۴۶۵	۲۷۵۳	۲۱۲۶	۹۶	۶۱۱	۶۰۵۱	۴۱۱	۲۶۹۲	۲۱۲۱	۹۶	۶۰۷	۵۹۴۲
قرچک	۱۰۷	۱۱۹۶	۱۰۲۴	۵۵	۲۷۸	۲۶۶۰	۹۰	۱۱۶۶	۱۰۱۶	۵۵	۲۷۶	۲۶۰۰
ورامین	۲۲۷	۱۲۱۹	۱۰۶۷	۵۷	۳۱۷	۲۸۸۷	۱۸۹	۱۱۷۹	۱۰۶۷	۵۷	۳۱۴	۲۸۰۶
دماوند	۰	۲۰	۲۵	۰	۳	۴۸	۰	۲۰	۲۵	۰	۳	۴۸
گرمسار	۰	۰	۲۵	۱	۱	۲۷	۰	۰	۲۵	۱	۱	۲۷
ایوانکی	۰	۰	۱۲	۲	۰	۱۴	۰	۰	۱۲	۲	۰	۱۴
جمع کل	۳۱۱۰	۲۷۲۲۱	۲۵۷۴۸	۲۵۷۳	۶۸۸۲	۶۵۵۳۴	۲۸۴۲	۲۶۹۲۷	۲۵۶۸۳	۲۵۷۳	۶۸۵۴	۶۴۸۷۹

منبع: یافته‌های پژوهشگر

یافته‌های روش خوشه‌بندی سلسله مراتبی (دو مرحله‌ای)

با استفاده از تنظیم‌های اولیه خود سیستم (تعداد خوشه‌ها بین۲ تا ۱۵) در این حالت دو خوشه تخمین زده می‌شود ولی وقتی الگوریتم خوشه‌بندی دو مرحله‌ای داده‌های پرت و دورافتاده را نادیده می‌گیرد چهار خوشه ایجاد می‌شود که در جدول۳ ذیل تعداد مشتریان در هر خوشه آمده است (علت تفاوت در جمع کل داده‌های آموزشی با تعداد داده‌های این قسمت به دو دلیل است: ۱- گردشدن اعداد، ۲- وجود داده‌هایی با مقادیر خالی برای نوع مشتریان بر اساس مقدار تازگی- تکرار- ارزش پولی است که مشتریان با این مشخصه‌ها از پایگاه داده‌ای موجود حذف شده‌اند. کل داده‌های آموزشی بر اساس تقریب ۷۰/۰۵۵درصد از ۶۴۸۷۹رکورد باقیمانده از بخش تشخیص مغایرت بدست آمده است که بر این اساس باید حدود۴۵۴۵۱رکورد برای آموزش موجود باشد درحالی‌که تنها۴۰۶۸۰رکورد دارای مقادیر بر اساس مقدار WRFM بوده‌اند و حدود۴۷۷۱ رکورد با این تفاسیر دارای مقادیر خالی بوده‌اند که حذف شده‌اند).

جدول ۳- مقدار تازگی- تکرار- ارزش پولی هر خوشه در خوشه بندی دو مرحله‌ای

وضعیت خوشه‌بندی	تعداد خوشه	شماره خوشه	میانگین تازگی	میانگین تکرار	میانگین ارزش پولی	میانگین WRFM	تعداد رکورد هر خوشه	درصد هر شاخه
حالت اولیه	۲	۱	۱/۴۵۸	۰/۷۲۴	۱/۲۳۵	۳/۴۱۷	۲۷۰۷	۶/۶۵
		۲	۱/۷۳۵	۰/۴۴۸	۰/۸۱۷	۳	۳۷۹۷۳	۹۳/۳۵
بدون فرض داده های پرت[۶۱]	۴	۱	۱/۸۷۵	۰/۴۴۴	۰/۸۳۶	۳/۱۵۵	۱۸۲۶۰	۴۴/۸۹
		۲	۱/۴۶۴	۰/۷۲۵	۱/۲۳۷	۳/۴۲۶	۲۷۲۴	۶/۷۰
		۳	۱/۸۸	۰/۴۶۸	۰/۸۰۸	۳/۱۵۶	۶۸۸۹	۱۶/۹۳
		۴	۱/۴۵۵	۰/۴۴۱	۰/۷۹۴	۲/۶۹	۱۲۸۰۷	۳۱/۴۸
جمع کل (مشتریان)							۴۰۶۸۰	۱۰۰

منبع: یافته‌های پژوهشگر

با توجه به جدول۳ در حالت اولیه خوشه دوم حدود ۹۳/۳۵ درصد مشتریان را شامل شده و این مشتریان در مرز بین مشتریان معمولی و نقره‌ای قرارگرفته‌اند (بر اساس مقادیر WRFM). در حالت دوم چهارخوشه‌ای، خوشه شماره یک با ۴۴/۸۹درصد مشتریان، بیشترین فراوانی را دارد و مشتریان آن بیشتر از نوع مشتریان نقره‌ای بوده‌اند. از نتایج بدست آمده مشهود است که در هر دو روش خوشه‌بندی دو مرحله‌ای بیشترین فراوانی از آن مشتریان نقره‌ای بوده است.

یافته‌های خوشه‌بندی حاصل از شبکه‌های عصبی خودسازمانده

با استفاده از شبکه‌عصبی خودسازمانده کوهنن یک ماتریس۳×۵ ایجاد شده است (در ماتریس کوهنن هر چه تعداد داده‌ها (مشتریان) دریک خوشه بیشتر باشد، رنگ خوشه قرمز تندتر خواهد بود بنابراین با توجه به شکل ۴ می‌توان وجود ۳ تا ۴ خوشه را برای این خوشه‌بندی تخمین زد (ماتریس کوهنن ذیل بر اساس تنظیم زمانی توقف ۵ دقیقه‌ای ایجاد شده است)). برای اینکه بررسی کنیم کدام تعداد خوشه در این روش کیفیت بهتری دارد از معیارهای امتیاز جینی، درصدخلوص، شاخص اطلاعات متقابل نرمال شده استفاده می شود.

شکل ۴- خروجی نقشه‌های خود سازمانده با ابعاد۳×۵ بر اساس نوع مشتریان

منبع: یافته‌های پژوهشگر

در ماتریس ایجاد شده مذکور، ۵ سلول آن نسبت به مابقی سلول‌ها دارای بیشترین تعداد مشتری بوده که در این پژوهش تنها به این ۵ سلول اشاره می‌شود (مابقی سلول‌ها یا خالی از داده‌هاست یا فراوانی داخل سلول‌های آنها بسیار ناچیز بوده است) که شرح آن‌ها در جدول ۵ و در ادامه پژوهش آمده است.

جدول ۵- خوشه‌بندی کوهنن و مقدار تازگی- تکرار- ارزش پولی هر خوشه

روش	ابعاد ماتریس	نرون‌های[۶۲] لایه ورودی	نرون‌های لایه خروجی	سلول	تعداد مشتریان هر سلول	مقدار تازگی	مقدار تکرار	مقدار ارزش پولی	مقدار WRFM	درصد فراوانی هر سلول
خوشه بندی کوهنن	۵×۳	۴۶	۱۵	(X_0,Y_0)	۷۵۲۹	۲/۱۴۲	۰/۴۵۷	۰/۷۳۹	۳/۳۳۹	۱۸/۵۱
				(X_0,Y_2)	۶۸۱۹	۱/۹۶۶	۰/۴۸۹	۰/۸۵۴	۳/۳۰۹	۱۶/۷۶
				(X_4,Y_2)	۵۹۹۷	۱/۵۴۶	۰/۴۲۸	۰/۸۷۸	۲/۸۵۲	۱۴/۷۴
				(X_4,Y_0)	۵۱۳۴	۰/۵۷۷	۰/۴۵۹	۰/۸۲۶	۱/۸۶۱	۱۲/۶۲
				(X_2,Y_2)	۴۲۵۰	۲/۲۸۹	۰/۴۹۱	۰/۸۹۲	۳/۶۷۲	۱۰/۴۵

منبع: یافته‌های پژوهشگر

در اینجا تنها به اطلاعات جمعیت شناختی و خرید موجود در ۵ سلول ماتریس کوهنن ایجاد شده با بالاترین فراوانی اشاره می‌شود:

۱) در سلول (X_0,Y_0) مشتریان با احتمال۱۰۰درصد از منطقه ۸ تهران بوده و در وضعیت فعال قرار داشته‌اند؛ همچنین این مشتریان بیشتر در آذرماه با فراوانی ۱۲/۷۲درصد خرید کرده‌اند. این خوشه با فراوانی در حدود ۱۸/۵۱درصد با اصل پارتو یا اصل۲۰-۸۰ (بیست درصد مشتریان ۸۰ درصد درآمد شرکت ایجاد می‌کنند) تا حدی تطابق داشته و به علت مقداری که برای معیار تازگی- تکرار- ارزش پولی ایجاد شده است خوشه ارزشمندی است.

۲) در سلول (X_0,Y_2) مشتریان با احتمال۵۰/۱۲ درصد از پاکدشت بوده و با ۹۷/۸۳درصد در وضعیت فعال قرار داشته‌اند؛ همچنین این مشتریان بیشتر در بهمن ماه با فراوانی ۱۰/۰۵ درصد خرید کرده‌اند. مشتریان این خوشه با فراوانی نزدیک ۹۶/۶۴درصد از مناطقی به جز تهران بوده‌اند. با توجه به درصد فراوانی و مقدار تازگی- تکرار- ارزش پولی ایجاد شده این خوشه ارزشمند است.

۳) در سلول (X_4,Y_2) مشتریان با احتمال ۶۷/۹۸ درصد از منطقه ۱۵بوده و با احتمال ۹۸/۳۳درصد بیشتر مشتریان این خوشه از شهر تهران بوده‌اند؛ همچنین این مشتریان بیشتر در تیرماه

با فراوانی۱۳/۱۴درصد خریدکرده‌اند. این مشتریان کاملاً در وضعیت فعال قرارداشته‌اند و با توجه به درصد فراوانی و مقدار تازگی- تکرار- ارزش پولی ایجاد شده این خوشه نسبتاً ارزشمند است.

۴) در سلول (X_4,Y_0) مشتریان با احتمال ۲۹/۲۸ درصد از منطقه ۱۵بوده و با احتمال۸۷/۸۷ درصد بیشتر مشتریان این خوشه از شهر تهران بوده‌اند؛ همچنین این مشتریان بیشتر در تیرماه با فراوانی ۱۴/۸۸درصد خرید کرده‌اند، این مشتریان با احتمال ۹۹/۷۵ درصد در وضعیت غیر فعال قرارداشته‌اند.

۵) در سلول (X_2,Y_2) مشتریان با احتمال۴۳/۹۳درصد از منطقه ۱۳بوده و این مشتریان کاملاً از شهر تهران بوده‌اند همچنین در وضعیت فعال قرارداشته اند؛ این مشتریان بیشتر در اسفند ماه با فراوانی۱۹/۶۷درصد خرید کرده‌اند. با توجه به مقدار تازگی- تکرار- ارزش پولی ایجاد شده این خوشه نسبت به چهار خوشه (سلول) اشاره شده در بندهای قبل دارای ارزش بیشتری برای شرکت است چراکه با مقدار۳/۶۷۲ برای معیار تازگی- تکرار- ارزش پولی بالاترین ارزش را برای شرکت ایجاد کرده است.

یافته‌های خوشه‌بندی K میانگین

در این قسمت خوشه‌بندی‌های K میانگین با فرض۳، ۴، ۵ خوشه اجرا شده‌اند که نتایج آن به شرح جدول ۶ ذیل بوده است.

جدول ۶- خوشه بندی K میانگین بر اساس داده های آموزشی به تفکیک نوع مشتری و WRFM

شماره خوشه	مشتری طلایی (تعداد)		مشتری نقره ای (تعداد)		مشتری معمولی (تعداد)		مشتری پرریسک (تعداد)		تعداد	درصد آموزش	مقدار WRFM
	درصد آموزش	درصد آزمون	درصد آموزش	درصد آزمون	درصد آموزش	درصد آزمون	درصد آموزش	درصد آزمون			
۱	۱۸۰۳		۷۳۸۲		۲۴۷۲		۲۵		۱۱۶۸۲	۲۸/۷۲	۳/۶۹۵
	۹۱/۰۶	۹۲/۴۳	۳۹/۰۸	۳۷/۹۴	۱۳/۷۴	۱۴/۵۶	۱/۳۷	۱/۷۷			
۲	۰		۶۸		۴۷۲۵		۱۳۶۵		۶۱۵۸	۱۵/۱۴	۱/۹۰۶
	۰	۰	۰/۳۶	۰/۳۵	۲۶/۲۷	۲۶/۹۵	۷۴/۷۹	۷۴/۸			
۳	۱۷۷		۱۱۴۳۸		۱۰۷۹۰		۴۳۵		۲۲۸۴۰	۵۶/۱۴	۲/۹۸۸
	۸/۹۴	۷/۵۷	۶۰/۵۶	۶۱/۷۱	۵۹/۹۹	۵۸/۴۹	۲۳/۸۴	۲۳/۴۳			
شماره خوشه	مشتری طلایی (تعداد)		مشتری نقره ای (تعداد)		مشتری معمولی (تعداد)		مشتری پرریسک (تعداد)		تعداد مشتریان آموزش	درصد مشتریان آموزش	مقدار WRFM
۱	۱۴۵۱		۱۲۷۶۷		۱۰۸۸		۰		۱۵۳۰۶	۳۷/۶۳	۳/۷۹۳
	۷۳/۲۸	۷۴/۸۷	۶۷/۵۹	۶۷/۸۷	۶/۰۵	۵/۹	۰	۰			
۲	۵۲۹		۳۷۵۶		۲۴۱۴		۳۶		۶۷۳۵	۱۶/۵۶	۳/۳۰۵
	۲۶/۷۲	۲۵/۱۳	۱۹/۸۹	۱۸/۷۸	۱۳/۴۲	۱۴/۰۹	۱/۹۷	۲/۳۶			
۳	۰		۶۸		۴۶۶۶		۱۳۵۶		۶۰۹۰	۱۴/۹۷	۱/۹۰۶

			۷۴/۲۱	۷۴/۳	۲۶/۷۴	۲۵/۹۴	۰/۳۵	۰/۳۶	۰	۰
۲/۴۸۸	۳۰/۸۵	۱۲۵۴۹	۴۳۳		۹۸۱۹		۲۲۹۷		۰	
			۲۳/۴۳	۲۳/۷۳	۵۳/۲۷	۵۴/۵۹	۱۳/۰۰	۱۲/۱۶	۰	۰
مقدار WRFM	درصد مشتریان آموزش	تعداد مشتریان آموزش	مشتری پرریسک (تعداد)		مشتری معمولی (تعداد)		مشتری نقره ای (تعداد)		مشتری طلایی (تعداد)		شماره خوشه
۳/۲۶۹	۲۷/۳۰	۱۱۱۰۶	۱۱۱		۴۰۰۸		۶۴۴۳		۵۴۴		۱
			۵/۳۱	۶/۰۸	۲۲/۱۵	۲۲/۲۸	۳۴/۳۶	۳۴/۱۱	۲۸/۴	۲۷/۴۷
۳/۳۱	۱۶/۴۱	۶۶۷۶	۳۶		۲۳۷۲		۳۷۳۹		۵۲۹		۲
			۲/۳۶	۱/۹۷	۱۳/۶۹	۱۳/۱۹	۱۸/۶۱	۱۹/۸	۲۵/۱۳	۲۶/۷۲
۱/۹۰۵	۱۴/۹۵	۶۰۷۹	۱۳۵۴		۴۶۵۸		۶۷		۰		۳
			۷۴/۰۲	۷۴/۱۹	۲۶/۶۸	۲۵/۹	۰/۳۳	۰/۳۵	۰	۰
۳/۷۵۲	۲۲/۷۸	۹۲۶۸	۰		۸۰۶		۷۵۵۵		۹۰۷		۴
			۰	۰	۴/۵۵	۴/۴۸	۴۰/۴۱	۴۰	۴۶/۴۷	۴۵/۸۱
۲/۴۳۶	۱۸/۵۶	۷۵۵۱	۳۲۴		۶۱۴۳		۱۰۸۴		۰		۵
			۱۸/۳۱	۱۷/۷۵	۳۲/۹۳	۳۴/۱۵	۶/۲۹	۵/۷۴	۰	۰
	۱۰۰	۴۰۶۸۰	۱۸۲۵		۱۷۹۸۷		۱۸۸۸۸		۱۹۸۰		جمع کل

منبع: یافته‌های پژوهشگر

در خوشه‌بندی با kهای مختلف مشتریان معمولی و نقره‌ای در تمامی خوشه‌ها حضور داشته‌اند. در خوشه اول تمامی این خوشه‌بندی‌ها، مشتریان نقره‌ای بیشترین فراوانی را داشته‌اند. با توجه به مقدار درصد داده‌های مربوط به نوع مشتریان در بخش آموزش و آزمون مدل می‌توان گفت خوشه‌بندی برای مشتریان نوع طلایی بالاترین میزان تغییرها را داشته است ولی برای مشتریان پر ریسک تقریباً داده‌های آموزشی و آزمون یک چیز را پیش‌بینی کرده‌اند. در جدول۷ ذیل مقایسه بین روش‌های مختلف خوشه-بندی مورد استفاده در این پژوهش با توجه به فرمول‌های امتیاز جینی، درصدخلوص و میزان خطای خوشه‌بندی و شاخص اطلاعات متقابل نرمال شده آمده است.

جدول ۷- مقایسه روش‌های خوشه‌بندی بر اساس معیارهای مختلف سنجش کیفیت خوشه‌بندی

روش خوشه‌بندی	تعدا خوشه یا گروه ها	تعداد تکرار تنظیم اولیه	میزان خطا	درصد خلوص	شاخص اطلاعات متقابل نرمال شده	امتیاز جینی
تشخیص مغایرت بار دوم	۳ گروه	--	--	۰/۵۳	۰/۰۳	۰/۴۲۹
دو مرحله ای	۴	بدون داده های پرت		۰/۵۳۳	۰/۰۲۶	۰/۴۳۱
دومرحله ای	۲	با تنظیم های اولیه		۰/۴۶۴	۰/۰۱	۰/۴۱۷
Kمیانگین	۳	۲۰	۰/۰۷۹	۰/۵۷۹	۰/۱۹۶	۰/۴۹۷

	۴	۲۰	۰/۰۰۶	۰/۷۶۱	۰/۳۱۱	۰/۶۳۶
	۵	۲۰	۰/۰۱۶	۰/۷۰۱	۰/۲۱۵	۰/۵۷۸

منبع: یافته‌های پژوهشگر

با مقایسه یافته‌های حاصل از خوشه‌بندی‌های مختلف بر پایه میزان کیفیت، خطای خوشه‌بندی و دیگر معیارهای مورد بررسی این پژوهش از بین روش‌های مختلف خوشه‌بندی، خوشه‌بندی K میانگین کیفیت بهتری داشته است و در این روش تعداد ۴خوشه نسبت به دیگر روش‌ها دارای درصد خلوص، امتیازجینی و شاخص اطلاعات متقابل نرمال شده بیشتری بوده است (شاخص‌های اطلاعات متقابل نرمال شده هر چقدر به یک نزدیکتر شود بهتر است و امتیاز جینی هر چقدر به یک نزدیکتر شود بهتر است).

بحث و نتیجه‌گیری

با توجه به پیشینه پژوهش حاضر این نکته قابل مشاهده است که زمینه‌های به‌کارگیری داده‌کاوی بسیارمتنوع بوده ولی تمرکز بیشتر پژوهش‌های صورت گرفته درحیطه مسایل پزشکی برای پیش بینی دسته‌های مختلف بیماران و علایم مؤثر بر بیماری بوده و یا در زمینه بانکداری به منظور تعیین مشتریان کم‌ریسک برای پرداخت بهتر وام‌دهی و یا پیش‌بینی احتمال کلاهبرداری صورت گرفته است و یا در بازار بورس برای دسته‌بندی شرکت‌های فعال در این عرصه بوده است و کمتر تمرکز آنها برروی میزان دقت و کیفیت دسته بندی‌های ایجادشده بوسیله روش‌های مختلف خوشه بندی بوده است درحالی‌که در این پژوهش سعی شده از شیوه‌های مختلف دقت و کیفیت خروجی روش‌های مختلف خوشه بندی بهبود یابد.

پژوهش حاضر با پژوهش‌های رییسی وانانی و همکاران (۱۳۹۹) و رنگریز و بایرامی (۱۳۹۸)، از این منظر که از داده‌کاوی برای دسته‌بندی و بخش‌بندی بازار مشتریان استفاده کرده‌اند وجه اشتراک دارد ولی از منظر اینکه بازارهای متفاوتی را مورد بررسی قرارداده‌اند تفاوت آشکار دارد و همچنین در این پژوهش‌ها شاخصی برای بررسی کیفیت خوشه‌بندی معرفی نشده است که این خود وجه تمایز دیگری است؛ پژوهش حاضر پژوهش‌های کتابی و همکاران (۱۳۹۵) از این نظر که سه روش خوشه‌بندی K میانگین، دو مرحله‌ای، شبکه‌های عصبی خودسازمانده در کنار معیار تازگی، تکرار و ارزش پولی خرید برای بخش‌بندی مشتریان استفاده کرده‌اند بیشترین اشتراک را داشته ولی از این نظرکه درپژوهش مذکور از معیار سیلوئت برای بررسی صحت و اعتبار نتایج خوشه‌بندی استفاده شده است و از نظر حیطه بازار مورد بررسی تفاوت‌هایی دارد. این پژوهش با پژوهش‌های عاشوری و همکاران (۱۳۹۵)، فیروزی جهان تیغ و عامری (۱۳۹۴) و غفاری و سلماسی (۱۳۸۸) از حیث این‌که از نرم افزار کلمنتاین برای خوشه‌بندی

استفاده کرده‌اند شباهت دارد ولی از حیث بازارهای مورد بررسی متفاوت است. بیشتر پژوهش‌ها برای تعیین تعداد خوشه‌های بهینه وکیفیت خوشه‌بندی تنها به یک شاخص مانند شاخص سیلوئت و یا شاخص دان اکتفا نموده‌اند مانند پژوهش‌های کتابی و همکاران (۱۳۹۵) و جهان تیغ و عامری (۱۳۹۴) که با این پژوهش که از روش‌های مختلفی مانند مقدارامتیاز جینی، آنتروپی، درصد خلوص و شاخص اطلاعات متقابل نرمال شده کیفیت خوشه‌بندی را سنجیده است تفاوت دارد. این پژوهش با پژوهش فمینا بهری و الایدوم (۲۰۱۵) از این نظرکه از داده‌کاوی به منظور ایجاد چارچوبی کارآمد برای مدیریت ارتباط با مشتری با هدف پیش‌بینی رفتار مشتریان استفاده کرده‌اند وجه اشتراک داشته ولی از این نظرکه این پژوهشگران الگوریتم‌های متفاوتی را برای بهبود و ارتقای دقت پیش‌بینی خود به‌کارگرفته‌اند تفاوت دارد. می‌توان گفت پژوهش هانگ (۲۰۱۱) که به دنبال استفاده از روش داده‌کاوی برای خوشه‌بندی مشتریان بر پایه داده‌های دموگرافیک و تراکنش‌های مالی مشتریان است با این پژوهش که به دنبال دسته‌بندی مشتریان براساس متغیرهای جمعیت شناختی و معیار تازگی- تکرار - ارزش پولی خرید است از نظر نوع متدولوژی پژوهش مورد استفاده شباهت داشته ولی از نظر بازار مورد بررسی متفاوت است. این پژوهش با پژوهش‌های ناگای و همکاران(۲۰۰۹) از این منظرکه از روشk میانگین برای خوشه بندی استفاده کرده‌اند تشابه دارد ولی از منظر ماهیت بازار مورد بررسی متفاوتند. کتابی و همکاران (۱۳۹۵) که درباره بخش‌بندی بازار شرکت‌های الکترونیکی پژوهشی انجام داده‌اند با استفاده از نظر کارشناسان و با محاسبه-های معادله‌های سلسله مراتبی فازی برای تازگی، تکرار و ارزش پولی وزن‌هایی تعیین کرده‌اند درحالی‌که در پژوهش حاضر با توجه به نظر مشورتی خبرگان صنعت و با محاسبه‌های معادله‌های سلسله مراتبی (مقایسه‌های زوجی) برای این معیارها وزن‌های متفاوتی تعیین شده است که این نشان‌دهنده آن خواهد بود که این مقادیر وزنی می‌تواند متناسب با نظر کارشناسان و حتی صنعت مورد بررسی دستخوش تغییرهایی شود.

محدودیت‌ها و پیشنهادهای پژوهش

محدودیت‌های این پژوهش آن است که نتایج مربوط به یک شرکت مورد مطالعه است. حتی این نتایج با افزایش داده‌ها در پایگاه داده مشتریان همین شرکت نیز قطعاً دستخوش تغییرهایی خواهند شد. زیرا اساساً استخراج دانش از داده‌ها امری پویاست و از این رو ضروری است که شرکت به تناوب محاسبه‌ها را تکرارکند تا نتایج به‌روز رسانی شوند.

از محدودیت‌های دیگر این پژوهش می‌توان به به نوع جامعه آماری آن اشاره کرد چراکه تنها اطلاعات مشتریان استان تهران بررسی شده و مقایسه‌ای با دیگر الگوهای رفتاری و خرید مشتریان در دیگر

استان‌ها صورت نگرفته است (آب و هوا، اقلیم، منطقه‌بندی و نوع سبک زندگی مردم در ایجاد الگوهای خرید متفاوت تأثیرگذار خواهد بود و نتایج متفاوتی را ایجاد خواهد کرد).

پیشنهاد می‌شود پژوهشگران درکنار معیارهای این پژوهش ازمعیارهای شاخص سیلوئت و شاخص دان نیزکمک بگیرند و نتایج آنها را مقایسه نمایند.

با توجه به ماهیت پویای داده‌کاوی و تغییرهای سلیقه و رفتار مشتریان از آنجایی‌که این پژوهش مربوط به اطلاعات بخشی از مشتریان استان تهران شرکت مورد بررسی در بازه سال‌های ۱۳۹۶-۱۳۹۷بوده است؛ به مدیران فروش، بازاریابی و تحقیقات بازار شرکت پیشنهاد می‌شود برای سال‌های ۱۳۹۸-۱۳۹۹ نیز این پژوهش مجدد بررسی شود تا در صورت معتبر بودن و یکسانی نتایج بتوان آن را به کل مشتریان استان تهران شرکت نسبت داد.

با توجه به کیفیت و تعداد خوشه‌بندی‌های ایجاد شده پیشنهاد می‌شود مدیران فروش و بازاریابی شرکت اولاً با شناسایی مشتریان پر ریسک و بررسی ویژگی‌های جمعیت شناختی این نوع مشتریان وضعیت ریسک درآمدی خود را کاهش داده و میزان بهره‌وری خود را افزایش دهند و سپس به مشتریان معمولی که در بعضی از ویژگی‌های جمعیت شناختی و کمی خرید در خوشه‌های مشتریان پرریسک قرارگرفته‌اند نیز توجه ویژه نمایند چراکه در آینده این احتمال وجود دارد اگر برنامه بازاریابی و فروش خاصی برای این دسته از مشتریان در نظرگرفته نشود به مشتریان پرریسک فردا تبدیل شوند؛

همچنین پیشنهاد می‌شود مدیران فروش و بازاریابی شرکت بر اساس نمره WRFM، مشتریانی که در این شاخص نمره بالاتری داشته‌اند و از نظر خوش حسابی در وضعیت خوبی‌اند را شناسایی کنند تا بتوانند برای آنها برنامه‌های ویژه‌تری از فروش و بازاریابی در نظر بگیرند.

توان تحلیل به موقع و دقیق داده‌های انباشته شده از مشتریان می‌تواند منجر به یک مزیت رقابتی شود. هدف این پژوهش نشان دادن این امر بوده است که تحلیل داده‌های حاصل از مشتریان می تواند دانش با ارزشی در اختیار تصیم‌سازان بخش بازاریابی، فروش و تحقیقات بازار و مدیران ارشد سازمان قرار دهد. با بررسی اجمالی نتایج این پژوهش و میزان صحت و دقت روش‌های داده‌کاوی آن، می‌توان این اطمینان را به مدیران فروش و بازاریابی شرکت‌های فعال در زمینه محصول‌های بهداشتی و آرایشی داد که به راحتی با استفاده از روش‌های مختلف داده‌کاوی می‌توانند تحول بزرگی درزمینه شناسایی الگوهای رفتاری خرید مشتریان خود و همچنین عملیات میدانی تحقیقات بازار-که مقوله‌ای هزینه‌بر است- ایجادکرده و بتوانند برنامه‌ریزی‌های دقیقتر و باکیفیت‌تری را انجام داده و در نهایت تصمیم‌های مفیدتری را با هزینه‌های کمتر اتخاذ نمایند.

منابع

۱) داگلاس، ایوان جی (۱۳۷۵) "اقتصاد مدیریت"، سیدجواد پورمقیم، تهران، نشر نی.

۲) رنگریز، حسن؛ بایرامی‌شهریور، زهرا (۱۳۹۸) تأثیرمدیریت ارتباط با مشتری الکترونیکی بر وفاداری مشتریان با استفاده از تکنیک‌های داده‌کاوی، مطالعات مدیریت کسب و کار هوشمند، (۷)۲۷، صص ۱۷۵-۲۰۵.

۳) رییسی وانانی، سینا، رییسی وانانی، ایمان، تقوی فرد، محمدتقی (۱۳۹۹) مدلی برای بخش‌بندی یادگیرندگان و بهبود عملکرد آموزشی با استفاده از الگوریتم‌های داده کاوی، مطالعات مدیریت کسب و کار هوشمند، ()، doi: 10.22054/ims.2020.49579.1668

۴) شهرابی، جمال (۱۳۹۴) داده‌کاوی، تهران، سروش گیتا.

۵) صادقی مال امیری، منصور (۱۳۹۲) طراحی مدل تحلیل رفتار مصرف‌کننده بر اساس آموزه‌های اسلام، فصلنامه علمی و پژوهشی اندیشه مدیریت راهبردی، سال هفتم، شماره اول، صص ۱۲۳- ۱۵۶.

۶) عاشوری، مریم، محمدی، شهریار، حسینی ایوری، هدی سادات (۱۳۹۵) کشف الگوی وضعیت اهداکنندگان خون از طریق خوشه‌بندی، مجله دانش و تندرستی در علوم پزشکی، دوره ۱۱، شماره ۴، صص۷۳- ۸۲.

۷) غفاری، باهره، سلماسی، ناصر (۱۳۸۸) شناسایی مشترکان تلفن همراه که تمایل به ترک شرکت دارند به روش خوشه‌بندی، سومین کنفرانس داده کاوی، تهران.

۸) فیروزی جهان تیغ، فرزاد. عامری، حکیمه (۱۳۹۴) بررسی ویژگی‌های بیماران مبتلا به سل با استفاده از روش خوشه‌بندی K-Means، مجله انفورماتیک سلامت و زیست پزشکی، دوره دوم، شماره ۳، صص ۱۴۹- ۱۵۹.

۹) کتابی، سعیده. ایزدی، بهرام، رنجبریان بهرام، نصیری مفخم، فریا (۱۳۹۵) یک رویکرد جامع برای بخش‌بندی بازار و طبقه‌بندی مشتریان با استفاده از روش‌های داده‌کاوی و برنامه‌ریزی خطی، فصلنامه علمی و پژوهشی مدیریت تولید و عملیات، دوره هفتم، شماره (۱) پیاپی (۱۲)، بهار و تابستان، صص ۱-۲۲.

۱۰) محمدی، مرتضی و سهرابی، طهمورث (۱۳۹۶) تأثیر مدیریت ارتباط با مشتری الکترونیک بر رضایت مشتریان، فصلنامه مطالعات مدیریت کسب وکار هوشمند، سال ششم، شماره ۲۲، صص ۶۰۴-۶۲۳.

۱۱) محمودی، سید عباس، میرزایی، کمال، محمودی، مصطفی (۱۳۹۶) تعیین عوامل مؤثر بر سرطان معده با استفاده از رویکرد داده‌کاوی، مجله دانشکده پیراپزشکی دانشگاه علوم پزشکی تهران (پیاورد سلامت)، دوره ۱۱، شماره ۳، صص ۳۳۲-۳۴۱.

۱۲) مشایخی، هدی، نوراله، زهرا، دقیق نژاد، بی بی هانیه (۱۳۹۷) استفاده از الگوریتم‌های داده‌کاوی برای تحلیل سودمندی دوربین‌های راهنمایی و رانندگی، فصلنامه علمیو پژوهشی مهندسی حمل و نقل، دوره ۱۰، شماره ۱ ، شماره پیاپی ۳۸، صص ۱۱۷- ۱۳۵.

۱۳) ملک محمدی، سمیرا، علیزاده، سمیه (۱۳۹۳) "داده‌کاوی و کشف دانش گام‌به‌گام با نرم افزار Clementine

"، چاپ سوم، تهران، انتشارات دانشگاه صنعتی خواجه نصیر الدین طوسی.

۱۴) ویسی، هادی (۱۳۹۶) جزوه روش‌های آماری در پردازش زبان طبیعی(خوشه بندی)، دانشکده علوم و فنون نوین دانشگاه تهران.

15) Blattberg, R. C. , Kim, B. & Neslin, S. A. (2008). "Database Marketing: analyzing and managing customers" , Springer, New York.

16) Chao, A. & Schor, J. B. (1998)."Empirical Tests of Status Consumption: Evidence from Women's Cosmetics", Journal of Economic Psychology, 19(1), pp. 107-131.

17) Chen, S. C. , & Huang, M. Y. (2011). " Constructing credit auditing and control & management model with data mining technique". Expert Systems with Applications, 38(5), 5359–5365

18) Clark , R. A. ,Zbojab , J. J.& Goldsmith, R.E. (2007). "Status Consumption and Role-relaxed Consumption: A Tale of Two Retail Consumers", Journal of Retailing and Consumer Services, 14(1), pp. 45-59.

19) 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, pp 4558–4565.

20) Chiu, T., Fang, D., Chen, j., Wang, y., & Jeris, C. (2001). "A Robust and Scalable Clustering Algorithm for Mixed Type Attributes in Large Database Environment". Paper presented at the international conference on knowledge discovery and data mining. California, San Francisco. August 2001 pp 263–268 .https://doi.org/10.1145/502512.502549

21) Daniel C. MC. , Lamb, C. & Hair, J. (2003). Marketing, London: South-Western.

22) Engel, J. F. , Kollart, D.J. & Blackwell, R.D. (1968). Consumer Behavior, New York, NY: Holt, Rinehart & Winston

23) Femina Bahari T., Sudheep Elayidom M. (2015). "An Efficient CRM-Data Mining Framework for the Prediction of Customer Behavior". Procedia Computer Science, 46, pp 725–731.https://doi.org/10.1016/j.procs .2015.02.136

24) Goldsmith, R. E. & Clark, R. A. (2008). "An Analysis of Factors Affecting Fashion Opinion Leadership and Fashion Opinion Seeking" , Journal of Fashion Marketing and Management, 12(3), pp. 308-322.

25) Guangli N., Yibing C., Lingling Z., Yuhong G. (2010). "Credit Card Customer Analysis Based on Panel Data Clustering". Procedia Computer Science. 1(1):2489–2497.

26) Hanafizadeh, P., & Mirzazadeh, M. (2011). "Visualizing market segmentation using self-organizing maps and fuzzy delphi method – ADSL market of a telecommunication company" ,Expert Systems with Applications, 38(1), pp198-205. https:// doi.org/10.1016/j.eswa .2010.06.045

27) Howard, J. A., & Sheth, J. N. (1969). The theory of buyer behavior. New York: Wiley

28) James J.H. Liou, Gwo-Hshiung T.Z. (2010). "A Dominance-based Rough Set Approach to customer behavior in the airline market" , Information Sciences. 180(11), pp 2230-2238 https://doi.org/ 10.1016/ j.ins.2010.01.025

29) Kahle, L.R. (1995). "Role-relaxed Consumers: A Trend of the Nineties", Journal of Advertising Research, 35(2) pp. 66-71

30) Kotler . P., Armstrong, G., Saunders,J. & Wong, V. (2001*).* Principle of marketing, London: Prentice-Hall
31) Lakshmipadmaja,D. ,Vishnuvardhan, B. (2018). "Classification and Performance Improvement Using Random Subset Feature Selection Algorithm for Data Mining",Big data research,*Elsevier*.12 ,pp 1-12. https://doi.org/10.1016/j.bdr.2018.02.007
32) Low, P. & Freeman, I. (2007) . "Fashion Marketing to Women in Kazakhstan" , Journal of Fashion Marketing and Management, 11(1), pp. 41-55. https://doi.org/10.1108/13612020710734391
33) Ngai, E.W.T., Xiu L.& Chau, D.C.K. (2009). "Application of data mining techniques in customer relationship management: A literature review and classification" .Expert Systems with Applications 36(2), pp .2592–2602, https://doi.org/10.1016/j.eswa.2008.02.021
34) Nzuva, M.S. , Lawrence N. (2017). "Prediction of Tea Production in Kenya Using Clustering and Association Rule Mining Technique " .American Journal of Computer Science and Information Technology .5(2). https://doi.org/10.21767/2349-3917.100006.
35) O'Cass, A. (2001). " Consumer Self-monitoring, Materialism and Involvement in Fashion Clothing", Australasian Marketing Journal *(AMJ)*, 9(1), pp. 46-60. https://doi.org/10.1016/S1441-3582 (01)70166-8
36) Wilke, L. W. (2000). "Consumer Behaviors". New York: John Wiley and Sons Inc.

یادداشت‌ها

1 Daniel C. MC. , Lamb, C. & Hair, J
2 Kotler. P., Armstrong, G., Saunders, J. & Wong, V.
3 Engel, J.F. , Kollart, D. J. & Blackwell, R.D.
4 Howard, J. A. , & Sheth, J. N.
5 Wilke, L. W.
6 Chao, A. & Schor, J. B
7 O' Cas , A.
8 Low , P. & Freeman, I.
9 Clark, R.A. ,Zbojab, J. J.& Goldsmith, R.E
10 Kahle, L.R.
11 Goldsmith, R.E. & Clark, R.A.
12 Data Mining
13 Customer Relationship Management (CRM)
14 UnSupervised Data Mining
15 Classification
16 Prediction
17 Clustering
18 Data Description and Visualization
19 Blattberg, R. C. , Kim, B. & Neslin, S. A.
20 Self-Organizing Neural Net

21 Kohonen Net
22 Chiu, C. Y., Chen, Y. F., Kuo, I. T., & Ku, H. C.
23 TwoStep Clustering
24 Chiu, T., Fang, D., Chen, j., Wang, y., & Jeris, C.
25 Purity
26 Gini Score
27 Entropy
28 Gain Information Ratio
29 External Criterion
30 Normalized Mutual Information (NMI)
31 Segmentation
32 Recency, Frequency, Monetary Value (RFM) Index
33 Self-Organizing Maps (SOM)
34 Silhouette Coefficient
35 Artificial Neural Net (ANN)
36 Support Vector Machine (SVM)
37 Clementine Software
38 Classification and Regression Tree(CART)/(C&RT)
39 Chi Square Automatic Interaction Detection (CHAID)
40 Quick, Unbiased, Efficient, Statistical Tree (QUEST)
41 Sensitivity
42 Validation
43 Accuracy
44 Association Rules
45 Dunn's Index
46 CRISP-DM (Cross Industrial Standard Process –Data Mining)
47 Lakshmipadmaja, D., Vishnuvardhan, B.
48 Random Subset Feature Selection (R.S.F.S.)
49 K- Nearest Neighbor (KNN)
50 Nzuva, M.S. ,Lawrence N.
51 Femina Bahari T., Sudheep Elayidom M
52 Chen, S. C. , & Huang, M.
53 Guangli N., Yibing C., Lingling Z., Yuhong G
54 Panel Data
55 James J.H. Liou, Gwo-Hshiung T.Z.
56 Ngai.E.W.T, Xiu L. & Chau D.C.K.
57 Screening
58 Feature Selection
59 P-Value
60 Anomaly Detection Node
61 Exclude Outliers
62 Input Layer Neurons