فصلنامه علمی- پژوهشی تحقیقات بازاریابی نوین
سال هشتم، شماره دوم، شماره پیاپی (۲۹) تابستان ۱۳۹۷
تاریخ دریافت: ۱۳۹۶/۶/۲۲ تاریخ پذیرش: ۱۳۹۷/۷/۸
صص: ۸۶- ۶۹

Customer Behavior Analysis using Web Usage Mining

Shahrzad Jalaly[1], Neda Abdolvand[2], Saeedeh Rajaee Harandi[3]

1- Master's Degree, Information Technology Management, Faculty of Social Science and Economics, Alzahra University, Tehran, Iran
shahrzad.jalaly@gmail.com
2- Assistant Professor and Faculty Member of Management Department, Alzahra University, Tehran, Iran
n.abdolvand@alzahra.ac.ir
3- Master's Degree, Information Technology Management, Faculty of Social Science and Economics, Alzahra University, Tehran, Iran
Saeedeh.rh@gmail.com

Abstract

Recent marketing strategies consider customers as important sources of the organization. Therefore, acquiring knowledge about customers and understanding their needs is necessary for keeping customers in an e-commerce business. Online customer shopping behavior is difficult to predict because they rarely visit the stores for real shopping, which is a challenge for marketers and researchers. Therefore, online business needs to analyze customers' behavior in order to be successful. Hence, this study aims to provide a framework for increasing the accuracy of the analysis and recognition of customer groups as well as providing the model and rules for predicting customers' behavior. Therefore, CRISP-DM and K-means algorithm were used for clustering data. Then, by assigning three tags of purchase, waiting and not purchase to customers, customers were categorized by C5 decision tree. Finally, a model with a precision of 63.6% and a collection of 261 rules with a confidence of 70% was obtained.

Keywords: Customer Behavior Analysis; Data Mining; Web Usage Mining; Online Retailing

تحلیل رفتار مشتری با استفاده از کاوش کاربری وب

شهرزاد جلالی[۱]، ندا عبدالوند[۲*]، سعیده رجائی هرندی[۳]

۱- دانش‌آموختهٔ کارشناسی ارشد مدیریت فناوری اطلاعات، دانشکدهٔ علوم اجتماعی و اقتصاد، دانشگاه الزهرا (س)
shahrzad.jalaly@gmail.com
۲- استادیار گروه مدیریت دانشکدهٔ علوم اجتماعی و اقتصاد، دانشگاه الزهرا (س)
n.abdolvand@alzahra.ac.ir
۳- دانش‌آموختهٔ کارشناسی ارشد مدیریت فناوری اطلاعات، دانشکدهٔ علوم اجتماعی و اقتصاد، دانشگاه الزهرا (س)
saeedeh.rh@gmail.com

چکیده

در راهبردهای اخیر بازاریابی، مشتریان از منابع مهم سازمان قلمداد می‌شوند. بر اساس این، کسب دانش دربارهٔ مشتریان و درک نیازهای آنها برای حفظ مشتریان در تجارت الکترونیک بسیار ضروری است. پیش‌بینی رفتار خرید مشتریانِ برخط دشوار است؛ زیرا به‌ندرت بازدید آنها از فروشگاه‌ها به خرید واقعی ختم می‌شود و این موضوع برای بازاریابان و پژوهشگران نوعی چالش شده است؛ از این‌رو، برای داشتن کسب‌وکارِ برخط موفق باید رفتار مشتریان را تحلیل کرد. بنابراین، این پژوهش با دو هدفِ الف) طرح چارچوبی برای افزایش دقت تحلیل و شناخت گروه‌های مشتریان و ب) ارائهٔ مدل و قوانینی برای پیش‌بینی رفتار آنها، رفتار مشتریان را تحلیل می‌کند. در این پژوهش از روش کریسپ و الگوریتم کا-میانگین برای خوشه‌بندی مشتریان استفاده شده است؛ سپس با اختصاص سه نوع برچسب خرید، خریدنکردن و انتظار خرید به مشتریان و با استفاده از درخت تصمیم C5 مشتریان دسته‌بندی شدند. درنهایت، مدلی با دقت ۶۳.۶٪ و مجموعه‌ای از ۲۶۱ قانون مناسب با اطمینان ۷۰٪ برای کسب‌وکار به دست آمد.

کلید واژه‌ها: تحلیل رفتار مشتری، داده‌کاوی، کاوش کاربری وب، خرده‌فروشیِ برخط

* نویسنده مسؤول

۱- مقدمه و بیان مسئله

در عصر الکترونیک، شرکت‌های تجارت الکترونیک از دنیای قدیمی که در آن، محصولات استاندارد، بازارهای همگن و چرخهٔ توسعه و عمر محصول طولانی یک قانون بود، به دنیای جدیدی انتقال می‌یابند که در آن انواع محصولات استاندارد جایگزین وجود دارند. مصرف‌کنندگان می‌توانند از میان میلیون‌ها کالا در یک فروشگاهِ برخط به جای ده‌ها هزار کالا در فروشگاه‌های بزرگ، کالاهای ضروری خود را انتخاب کنند (آرورا و چپرا[1]، ۲۰۱۶). از یک طرف رشد سریع و انتشار فناوری اینترنت، محبوبیت بازدید از پایگاه‌های وب را در میان بازدیدکنندگان افزایش داده است و از طرفی دیگر، در راهبردهای اخیر بازاریابی، مشتریان جزو منابع مهم یک سازمان قلمداد می‌شود (پارک و چونگ[2]، ۲۰۰۹). بنابراین کسب دانش دربارهٔ مشتریان و درک نیازهای آنها برای حفظ مشتریان در تجارت الکترونیک بسیار ضروری است، زیرا رق؛ا به اندازهٔ یک کلیک از ما فاصله دارند (ژنگ[3] و همکاران، ۲۰۰۴). برای خرده‌فروشی برخط که در آن حفظ مشتری از عوامل کلیدی موفقیت است، توسعهٔ مؤثر حضور در وب و عملیات بخش‌های مدیریتی لازم است. یک خرده‌فروشیِ برخط موفق باید سطحی ایده‌آل از سیستم، اطلاعات و کیفیت خدمات را ارائه کند و مشتریان را برای بازدید دوباره از پایگاه وب خود جلب کند. هنگامی که بازدیدکنندگان، بازدید لذت‌بخشی داشته باشند، احتمال بازدید دوبارهٔ آنها از پایگاه وب افزایش می‌یابد (آهن[4] و همکاران، ۲۰۰۷). پیش‌بینی رفتار خرید مشتریان برخط دشوار است؛ زیرا به‌ندرت بازدید آنها از فروشگاه‌ها به خرید واقعی می‌انجامد و این برای بازاریابان و پژهشگران نوعی چالش شده است. میزان تبدیل بازدید به خرید برای کسب‌وکارهای برخط متوسط حداکثر ۳ درصد است (پارک و چونگ، ۲۰۰۹). در این صورت سازمان‌ها باید به جای هدف قراردادن تمام مشتریان به یک اندازه یا پیشنهاد مشوق‌های یکسان به همهٔ آنها، تنها مشتریانی را هدف قرار دهند که براساس نیازهای فردی یا رفتارهای خریدشان به معیارهای سودبخش مشخصی دست یافته‌اند (لیو و تزنگ[5]، ۲۰۱۰).

وب‌کاوی گام مؤثری برای رسیدن به این هدف است. وب‌کاوی، راهی تعیین‌کننده برای درک کاربران تجارت‌الکترونیک و تبدیل اطلاعات به مزیت رقابتی است و سازمان‌ها را قادر می‌سازد تا تصمیمات مبتنی بر داده بگیرند و راهبردهای تصمیم‌گیری خود را بهبود و توسعه دهند. همچنین در به دست آوردن مشتریان جدید و حفظ مشتریان موجود و بهبود رضایت مشتری نیز کمک می‌کند (ژنگ و همکاران، ۲۰۰۴؛ شانتی[6]، ۲۰۱۷). کاوش کاربری وب شاخه‌ای از وب‌کاوی است که بر کاربرد تکنیک‌های داده‌کاوی برای یافتن الگوهای مفید تمرکز دارد و می‌تواند رفتار کاربر را هنگامی تعامل با وب پیش‌بینی کند (روا و آرورا[7]، ۲۰۱۷). تجزیه و تحلیل این اطلاعات به سازمان‌ها در تعیین ارزش طول عمر مشتریان، طراحی راهبرد بازاریابی در محصولات و خدمات، ارزیابی اثربخشی کمپین‌های تبلیغاتی، بهینه‌سازی عملکرد برنامه‌های کاربردی مبتنی بر وب، به دست دادن محتوای شخصی‌سازی‌شده به بازدیدکنندگان و پیداکردن مؤثرترین ساختار منطقی برای فضای وب خود، کمک می‌کند. این نوع تجزیه و تحلیل شامل شناسایی خودکار الگوها و روابط معنی‌دار از مجموعهٔ بزرگی از

[1] Arora & Chopra
[2] Park & Chung
[3] Zhang
[4] Ahn
[5] Liou & Tzeng
[6] Shanthi
[7] Rao & Arora

داده‌های نیمه‌ساخت‌یافته است که بیشتر در پایگاه وب و برنامه‌های کاربردی لاگ‌سرور و همچنین در منابع دادهٔ عملیاتی مرتبط ذخیره می‌شوند (لیو[1] و همکاران، ۲۰۱۱). برای داشتن راه‌حل تجارت الکترونیک موفق، نیاز است که رفتارهای کلیک مشتری در پایگاه وب جمع‌آوری و بررسی شود؛ زیرا جذب مشتریان جدید و حفظ مشتریان باارزش از اهمیت بسیاری برخوردار است (ژنگ و همکاران، ۲۰۰۴). بنابراین این سؤالات مطرح می‌شود که چگونه می‌توان با ترکیب روش‌های داده‌کاوی و کاوش کاربری وب به درک و شناخت بهتری از گروه‌های مشتریان دست یافت و چگونه می‌توان با ترکیب روش‌های داده‌کاوی و کاوش کاربری وب و براساس ویژگی‌های رفتاری مشتریان پیش‌بینی کرد که آیا رفتار آن‌ها به خرید ختم می‌شود یا خیر؟ از این‌رو، این پژوهش، دو هدف اصلی را دنبال می‌کند.

نخستین هدف، طرح چارچوبی برای افزایش دقت تحلیل و شناخت گروه‌های مشتریان و دیگری، ارائهٔ مدل و قوانینی برای پیش‌بینی رفتار مشتریان بر مبنای ویژگی‌های رفتاری آنها با استفاده از ترکیب رویکردهای وب‌کاوی و داده‌کاوی است. در این پژوهش از رویکرد ترکیبی کمتر به‌کاررفتهٔ داده‌کاوی و کاوش کاربری وب استفاده شده است که به مزیت رقابتی برای شرکت می‌انجامد.

درواقع نوآوری پژوهش در اِعمال روش‌های داده‌کاوی بر ابعاد مختلف داده‌های مشتری شامل داده‌های رفتاری مرور وب و داده‌های رفتار خرید است که به بهبود دانش از مشتری خواهد انجامید. در این پژوهش، داده‌ها و رفتار کاربران درون سایت با استفاده از دو تکنیک خوشه‌بندی و دسته‌بندی تحلیل می‌شود. خوشه‌بندی مشتریان سبب می‌شود شرکت بتواند مشتریان خود و مشتریان باارزش را شناسایی کند. از این طریق شرکت برای مشتریان هر خوشه سیاست‌های متفاوتی را اعمال می‌کند و مبتنی بر دانش به‌دست‌آمده تصمیم می‌گیرد و محتوا و طراحی سایت را براساس نیازهای هر خوشه تغییر می‌دهد. در دسته‌بندی بررسی می‌شود که براساس ویژگی‌های رفتاری مشتریان، چه رفتارهایی باعث خرید شده و چه رفتارهایی به خرید ختم نمی‌شود. با ایجاد مدلی جامع و مقبول، می‌توان رفتار مشتریان جدید را با داشتن ویژگی‌های رفتاری آنان، در یکی از این دسته‌ها طبقه‌بندی و رفتار آیندهٔ آنها را پیش‌بینی کرد. این کار باعث می‌شود شرکت بتواند مشتریان باارزش خود را شناسایی کند و با سپردنِ مزایا و خدماتی به آنها به‌سوی جذب سود بیشتر حرکت کند.

با توجه به هدف، نخست مبانی نظری پژوهش و سپس مدل و روش پژوهش بررسی می‌شود و درنهایت پژوهش با بحث و نتیجه‌گیری و ارائه پیشنهادها خاتمه می‌یابد.

۲- مبانی نظری پژوهش

خرده‌فروشی برخط[2]، یک کسب‌وکار مبتنی بر اینترنت است که محصولات و خدمات را در وب ارائه می‌دهد. خرده‌فروشی برخط نه‌تنها یک سیستم اطلاعاتی است، ارائه‌ای کامل از یک فروشگاه به مشتری است (آهن و همکاران، ۲۰۰۷). با توجه به رشد روزافزون خرده‌فروشی برخط، خرده‌فروشان برخط به درک دلایل خاصی نیاز دارند که چرا مصرف‌کنندگان خرید برخط را انتخاب می‌کنند (دکا[3]، ۲۰۱۷). رفتار مشتری، حاصل تعامل پیچیده میان تعدادی از عوامل است که این عوامل شامل سطح فعالیت بازاریابی، رقابت‌پذیری محیط، درک نام تجاری، تأثیر فناوری‌های جدید و

[1] Liu

[2] Online Retailing
[3] Deka

نیازهای فردی است (لیو و تزنگ، ۲۰۱۰). به‌منظور افزایش رضایت مشتری و جلوگیری از ترک سازمان از سوی مشتری، سازمان باید بر بخش‌بندی و تأمین نیازهای فردی مشتریان متمرکز شود (تسای[1] و همکاران، ۲۰۱۵). به‌طورکلی «بخش‌بندی مشتری» فرایند تقسیم مشتریان سازمان به گروه‌های مختلف بر مبنای اطلاعات مختلف جغرافیایی، جمعیت‌شناختی، رفتار شناختی و اتخاذ راهبردهای مناسب هر گروه با توجه به مصرف کالا و خدمات و تاریخچهٔ خرید مشتریان است (تسای و همکاران، ۲۰۱۵). بخش‌بندی مشتری با تکنیک‌ها و روش‌های تحلیل داده‌ای متفاوت انجام می‌شود که در این میان استفاده از تکنیک‌های داده‌کاوی روی داده‌هایی که از تراکنش‌های برخط تولید می‌شوند، متداول‌تر از سایر روش‌هاست. داده‌کاوی قابلیت پیچیدهٔ جست‌وجوی داده است که از الگوریتم‌های آماری برای کشف الگوها و همبستگی داده‌ها استفاده می‌کند. به بیان ساده، داده‌کاوی فرایند خودکارِ یافتن اطلاعات مفید از مخازن بزرگِ داده است (بهشتیان اردکانی[2] و همکاران، ۲۰۱۸). یکی از متداول‌ترین روش‌ها برای مرتب‌سازی و تحلیل مشتریان، امتیازدهی به آنها براساس دفعات خرید و مقدار پرداخت است که از شناخته‌شده‌ترین روش‌ها برای انجام این کار مدل RFM (تازگی خرید، تکرار خرید و ارزش پولی خریدها) است که اساس بخش‌بندی برای بازاریابی مستقیم است. هنگامی‌که امتیازات RFM مشتریان تعیین می‌شود، می‌توان مشتریان را به‌صورت بخش‌هایی گروه‌بندی و متعاقباً سودآوری آنها را تحلیل کرد (مقدم[3] و همکاران، ۲۰۱۷).

وب بزرگ‌ترین پایگاهِ دادهٔ در دسترس و موضوعی چالش‌برانگیز برای داده‌کاوی است. درواقع وب‌کاوی، تحلیل رفتار الکترونیکی مشتری است و به‌نوعی استفاده از تکنیک‌های داده‌کاوی برای کشف و استخراج اطلاعات از مستندات و خدمات وب تعریف می‌شود (شیخ و مناریا[4]، ۲۰۱۷). کاربرد تکنیک‌های داده‌کاوی روی داده‌های کاربردی وب اصطلاح علمی جدیدی با نام کاوش کاربری وب ایجاد کرده است (دارمارنجان و دورایزانگاسوامی[5]، ۲۰۱۶). با کاوش کاربری وب، اطلاعات مفیدی مانند الگوهای پیمایش کاربران با استفاده از داده‌های لاگ‌وب استخراج و تحلیل می‌شود. هم طراحان و هم کاربران وب از کاوش کاربری وب سود می‌برند. از طرفی، با تحلیل الگوهای پیمایش در لاگ‌وب سرور، طراحان وب‌سایت رفتارهای بازدید کاربران وب را تعیین می‌کنند؛ بنابراین آنها درمی‌یابند معروف‌ترین صفحات روی سایت کدامند و کدام صفحات با احتمال بیشتری با هم بازدید می‌شوند. از طرفی دیگر، کاربران وب همچنین می‌توانند از این موضوع برای دسترسی مؤثرتر به وب استفاده کنند (سان و ژنگ[6]، ۲۰۰۴).

۲-۱- پیشینهٔ پژوهش

در بسیاری از پژوهش‌های انجام‌شده دربارهٔ رفتار مشتری، مقادیر دموگرافیک مشتریان برای تحلیل رفتار آنها استفاده می‌شود (لیو و تزنگ، ۲۰۱۰)؛ برای نمونه، ژنگ و همکاران (۲۰۰۴) با استفاده از داده‌های جامع که شامل لاگ‌وب و اطلاعات مشتریان در وب‌سایت‌های تجارت الکترونیک هستند، الگوهای رفتاری کاربران و درک رفتارهای خرید آنها در پایگاه وب را شناسایی کرده‌اند. در این مقاله از سه تکنیک دسته‌بندی، خوشه‌بندی و قوانین انجمنی برای بررسی روش پیشنهادی استفاده کرده‌اند. داده‌های دموگرافیک

[1] Tsai
[2] Beheshtian-Ardakani
[3] Moghadam
[4] Sheikh & Menaria
[5] Dharmarajan & Dorairangaswamy
[6] Sun & Zhang

و خرید کاربران و همچنین اطلاعات بازدید آنها از سایت شامل تعداد صفحات بازدیدشده در هر نشست و همچنین تعداد صفحات بازدیدشده در بخش‌های مختلف سایت نیز برای خوشه‌بندی مشتریان استفاده شده است.

همچنین در تحلیل رفتار مشتریان، اهمیت متغیرهای ارزشمند رفتار مشتری به طور گسترده مطالعه شده است (لیو و تزنگ، ۲۰۱۰). پژوهشگران مشاهده کرده‌اند که متغیرهای RFM نه‌تنها برای تحلیل رفتار مشتریان مفید هستند، می‌توانند به صورت مؤثر در تجسس ارزش مشتری و بازارهای گوشه[1] استفاده شوند (لیو و تزنگ، ۲۰۱۰). هسیه[2] (۲۰۰۴) روشی پیشنهاد می‌دهد که داده‌کاوی و مدل‌های امتیازدهی رفتار یعنی RFM را برای مدیریت مشتریان یک بانک یکپارچه می‌کند. او توانست مشتریان بانک را به سه گروه بزرگ سودمند دسته‌بندی کند.

به‌تازگی کاوش کاربری وب، توجه پژوهشگران و متخصصان تجارت الکترونیک را در تحلیل رفتار مشتری به خود جلب کرده است. پژوهش‌ها در زمینهٔ کاوش کاربری وب بیشتر بر توسعهٔ تکنیک‌های کشف دانش، به‌خصوص آنهایی که برای تجزیه و تحلیل داده‌های کاربری وب طراحی شده، متمرکز شده است. بیشتر این تلاش‌ها، بیشتر بر سه پارادایم اصلی قوانین وابستگی، الگوهای ترتیبی و خوشه‌بندی توجه دارند (فاکا و لنزی[3]، ۲۰۰۵)؛ برای مثال، در پژوهش ینگ و سو[4] (۲۰۱۲) از الگوریتم SVM برای خوشه‌بندی رفتار مشتریان شبکه به‌منظور ارائهٔ خدمات بهتر به آنها استفاده شده است که سازمان‌ها با استفاده از نتایج خدمات خود را بیشتر شخصی‌سازی می‌کنند. همچنین پارک و چونگ (۲۰۰۹) استفاده از داده‌های جریان‌های کلیک برای پیش‌بینی رفتارهای خرید کاربران را ارائه کردند. آنها با استفاده از تحلیل رگرسیون سلسله‌مراتبی ثابت کردند که کاربران انتقال‌داده‌شده از پایگاه وب ارجاع‌دهنده، کمتر از کاربرانی که مستقیم وارد سایت می‌شوند، خرید می‌کنند و هرچه مدت زمان ماندن این کاربران در سایت بیشتر باشد و صفحاتِ مشاهده‌کرده کمتر باشد، احتمال خریدشان بیشتر است. علاوه بر این، استفاده از تکنیک‌های داده‌کاوی شامل آمار توصیفی و قوانین انجمنی برای تحلیل رفتار هدایتی کاربران و شناسایی الگوهای گشت و گذار آنها را سیسودیا و ورما[5] (۲۰۱۲) مطرح کردند. هونگ[6] و همکاران (۲۰۱۳) در پژوهش مشابهی از روش کاوش کاربری وب روی سرویس مراقبت از خود برای افراد سالمند، به منظور بهبود درک و تحلیل رفتارهای آنها استفاده کرده‌اند. آنها این کار را با تکنیک‌های تحلیل انجمنی و مدل مارکوف همراه با الگوریتم بهبودیافته کا-میانگین انجام داده‌اند. با استفاده از نمونه‌ای از لاگ‌وب سرور مرکز فضایی ناسا و کاوش کاربری وب، پامونتا[7] و همکاران (۲۰۱۲) به اطلاعات آماری از نشست کاربر دست یافتند که می‌توان از آن برای شناسایی الگوهای دسترسی کاربران و تحلیل رفتار آنها استفاده کرد. جنامانی و همکاران (۲۰۰۳) نیز مدل فرایندی نیمه مارکوف (یک ابزار کاوش کاربری وب)، برای درک رفتار مشتریان الکترونیک را مطرح کردند که نتایج این مدل به بهبود طراحی سایت و تشخیص عملکرد آن کمک کرد. ها[8] (۲۰۰۲) سیستم شخصی‌سازی‌شدهٔ مبتنی بر وبی را پیشنهاد داد که از کاوش کاربری وب

[1] Niche Markets
[2] Hsieh
[3] Facca & Lanzi
[4] Yang & Su
[5] Sisodia & Verma
[6] Hung
[7] Pamutha
[8] Ha

برای دادن پیشنهادهایی شخصی به مشتریانِ برخط استفاده می‌کند. این سیستم برای ارائهٔ اطلاعات به مشتریان طراحی شده و به آنها در خرید کالاها کمک می‌کند. بایی[1] و همکاران (۲۰۰۳) سیستم انتخاب آگهی وبی طراحی کردند که کاربران پایگاه وب با ترجیحات مشابه را از راه کاوش کاربری وب به چند بخش تقسیم می‌کند. این سیستم با استنتاج فازی، تبلیغات مناسب را پیشنهاد می‌دهد. کیم[2] و همکاران (۲۰۰۳) روشی برای پیش‌بینی رفتار خرید مشتریان تجارت الکترونیک با ترکیب چند دسته‌بندی بر مبنای الگوریتم ژنتیک پیشنهاد داد که عملکرد بهتری نسبت به دسته‌بندهای تکی داشت. جنامانی[3] و همکاران (۲۰۰۳) با پیشنهاد مدل فرایندی نیمه‌مارکوف (یک ابزار کاوش کاربری وب)، به درک رفتار مشتریان الکترونیک پرداختند که نتایج این مدل به بهبود طراحی سایت و تشخیص عملکرد آن کمک می‌کند.

با توجه به پژوهش‌ها انجام‌شده در این زمینه، بیشتر پژوهش‌ها به پیش‌بینی رفتار مشتری با استفاده از رویکردهای وب‌کاوی یا داده‌کاوی تکیه کرده و پژوهش‌های اندکی به پیش‌بینی رفتار مشتریان با ترکیب این دو رویکرد توجه کرده‌اند و پژوهشی در این زمینه در ایران انجام نشده است. اگر اطلاعات بیشتر در دسترس باشد که با داده‌های لاگ‌وب ارتباط داشته و همچنین به مشکل هم مربوط باشند، می‌توانند به صورت چشمگیری نتایج را بهبود بخشند و به نتایج دقیق‌تر هم برسند. داده‌های مشتریان، اطلاعات بیشتری هستند که همراه با داده‌های لاگ استفاده می‌شوند. از این‌رو، در این پژوهش به‌منظور افزایش دقت تحلیل و شناخت گروه‌های مشتریان و ارائهٔ مدل و قوانینی برای پیش‌بینی رفتار مشتریان بر مبنای ویژگی‌های رفتاری آنها چارچوبی را ارائه می‌دهد. درواقع، از رویکرد ترکیبی و مغفول‌ماندهٔ داده‌کاوی و کاوش کاربری وب استفاده شده است که می‌تواند به مزیت رقابتی برای شرکت منجر شود. در این پژوهش، داده‌ها و رفتار کاربران سایت با استفاده از دو تکنیک خوشه‌بندی و دسته‌بندی تحلیل می‌شوند که به شناخت بهتر مشتریان باارزش منجر شده و براساس ویژگی‌های رفتاری مشتریان، مشخص می‌شود چه رفتارهایی باعث خرید شده و چه رفتارهایی نمی‌شود؛ شرکت با تحلیل این موارد می‌تواند با اعطای مزایا و خدماتی به مشتریان، به سوی جذب سود بیشتر حرکت کند. مزیت اصلی این روش کسب نتایج بهتر از فرایند تحلیل با استفاده از اطلاعات بیشتر مشتریان است.

۳- روش پژوهش

این پژوهش دارای رویکردی سازنده است و با دو هدفِ الف) ارائهٔ چارچوبی برای افزایش دقت تحلیل و شناخت گروه‌های مشتریان و ب) ارائهٔ مدل و قوانینی برای پیش‌بینی رفتار آنها رفتار مشتریان را تحلیل می‌کند. داده‌ها و رفتار کاربران درون سایت با استفاده از دو تکنیک خوشه‌بندی و دسته‌بندی تحلیل می‌شود. این پژوهش برمبنای متدولوژی کریسپ[4] انجام شده که یکی از قوی‌ترین روش‌های تحلیلی برای اجرای پروژه‌های داده‌کاوی است و شامل شش مرحلهٔ فهم کسب‌وکار، فهم داده، آماده‌سازی، مدل‌سازی، ارزیابی و به‌کارگیری است (بهشتیان اردکانی و همکاران، ۲۰۱۸). ترتیب و توالی این شش مرحله انعطاف‌پذیر است. روش کریسپ بسیار کامل و مستند است. تمامی مراحل آن، به‌موقع سازماندهی، ساخته و تعریف می‌شود که اجازه می‌دهد یک پروژه بتواند به‌راحتی

[1] Bae
[2] Kim
[3] Jenamani
[4] CRISP-DM

درک یـا تجدیـدنظر شـود (آزودو[1]، ۲۰۰۸). بـرای بخش‌بندی و ارزش‌گذاری و همچنین الگویابی رفتار خرید مشتری از الگوریتم‌هـا و تکنیـک‌هـای مختلفـی استفاده می‌شود که در دو گروه کلی خوشـه‌بنـدی و کشف قواعد انجمنی دسته‌بندی مـی‌شـوند. بـر همـین اساس، در این پژوهش سعی بر آن است تا از تکنیک خوشـه‌بنـدی و الگـوریتم کـا-میـانگین اسـتفاده شـود. الگوریتم خوشه‌بندی کا-میانگین یکـی از سـاده‌تـرین و البته مشهورترین الگوریتم‌های یـادگیری بـدون نظـارت است. این الگوریتم، کاربردی‌تـرین روش خوشـه‌بنـدی داده‌هاست که از مزایایی همچون سرعت و سـاده‌بـودن در اجـرا برخـوردار اسـت و در مسـائل بـزرگ بسـیار کارایی دارد. این الگوریتم، روشی بسیار مناسـب بـرای یافتن گروه‌های مشتریان با رفتارهای مشابه است و و در تلاش است تا عدم تشابه میان گروه‌هـای مشـتریان را بـه حـداقل برسـاند (تاسـینگر و هـوبر[2]، ۲۰۰۰)؛ سـپس به‌منظور سنجش صحت[3] نتایج، از معیار ارزیابی دیویس بولدین بـا اسـتفاده از نـرم‌افـزار رپیـدماینر روی ۷۴.۳۴۴ رکورد نرمال‌شده انجام می‌شود (اطلاعـات مربـوط بـه سازمان و داده‌ها در بخش بعد آورده شده است). بـرای بررسـی تمـایز خوشـه‌هـا آزمـون آنـووا بـا اسـتفاده از نرم‌افزار اس‌پی‌اس‌اس ۱۹ روی خوشه‌های به‌دست‌آمده اجـرا مـی‌شـود. در مرحلـهٔ بعـد بـا توجـه بـه ضـرورت استخراج قوانین، به دلیل اینکه شبکهٔ عصبی ماننـد یـک جعبهٔ سیاه عمل می کند، با وجود دقت بیشتر دسته‌بندی شـبکهٔ عصـبی، از الگـوریتم درخـت تصـمیم C5، بـا اختصاص سه برچسب به نام‌های خرید، انتظـار خریـد و خریدنکردن با استفاده از نرم‌افزار کلمنتاین برای بررسی رفتارهـا بـا توجـه بـه ویژگـی‌هـای رفتـاری کـاربران و دسته‌بندی مشتریان براساس ویژگی‌های رفتـاری آنهـا و ایجاد مدلی برای پیش‌بینـی نـوع رفتـار مشـتریان آینـده استفاده می‌شود. درخت تصـمیم در بـین الگـوریتم‌هـای طبقه‌بندی، روش قدرتمندی است کـه محبوبیـت آن بـا رشد داده‌کاوی به‌طور فزاینده‌ای در حال افزایش است. درخت تصمیم روشی بـرای نمـایش دسـته‌ای از قـوانین است که به یک رده یا مقدار منتهی می‌شـود. الگـوریتم درخـت تصـمیم C5 روشـی افزایشـی از هـرس‌کـردن درخت را به کار می‌گیرد تا خطای طبقه‌بندی ناشـی از نویز یـا جزئیـات خیلـی زیـاد را در داده‌هـای آموزشـی کاهش دهد. این درخـت مـی‌توانـد دسـته‌بنـدهایی هـم برای درخت تصمیم و هم برای مجموعـه قـوانین تولیـد کنـد. در پایـان بـا اسـتفاده از معیارهـای ارزیـابی الگوریتم‌های خوشـه‌بنـدی و دسـته‌بنـدی نتـایج ایـن دو تکنیک ارزیابی می‌شود.

٤- پیاده‌سازی

این پژوهش برمبنای فرایند کریسپ انجام شده و شـامل مراحل زیر است:

فاز ۱. درک کسب و کار

این فاز بر درک اهداف و نیازمنـدی‌هـای پـروژه از دیدگاه کسب و کار تمرکز مـی‌کنـد. ایـن پـژوهش در شرکت XYZ انجام شده است که مرجع تخصصی نقد و بررسی و فـروش اینترنتـی کـالا در ایـران اسـت. ایـن شـرکت در زمینـهٔ فـروش بـرخط محصـولات مختلـف فعالیت مـی‌کنـد و در حـال حاضـر در حـدود ۶۷ هـزار کالای مختلف را در سایت خود بـه فـروش مـی‌رسـاند. سایت این شرکت به طور میـانگین روزانـه بـیش از ۴۰۰ هزار بازدیدکننـده دارد و از پربازدیـدترین سـایت‌هـای کشور است. کاربران و مشتریان XYZ می‌توانند با حـق انتخاب متنوع و با دریافت اطلاعاتی کامل برای انتخاب درست کالای مدنظر خـود، وب‌سـایت ایـن شـرکت را

[1] Azevedo
[2] Theusinger & Huber
[3] Goodness

بررسی و با حداکثر اطمینان کالای خود را انتخاب و خرید کنند.XYZ با ارائهٔ طیف گسترده‌ای از معتبرترین برندها در گروه‌های مختلف و با همکاری نزدیک با واردکنندگان و توزیع‌کنندگان اصلی این کالاها در ایران، تلاش می‌کند نیازهای متفاوت مشتریان با کاربری‌های متفاوت آنان را برآورده سازد. کیفیت و سهولت استفاده از پایگاه وب و خدمات ارائه‌شده در آن، همواره یکی از مهم‌ترین و بااولویت‌ترین موضوعات در XYZ بوده است و این شرکت تلاش می‌کند در پایگاه وب XYZ، خدماتی شایسته و تجربه‌ای خوشایند را برای مخاطبان خود به ارمغان بیاورند.

فاز ۲ و ۳. شناخت و آماده‌سازی داده

این فاز با جمع‌آوری داده‌های اولیه و تصمیم اینکه کدام داده‌ها و در چه فرمت و اندازه‌ای لازم خواهند بود، آغاز می‌شود و فعالیت‌ها را به‌منظور آشنایی با داده‌ها، شناسایی مشکلات کیفیت داده‌ها، دریافت بینش‌های مقدماتی دربارهٔ داده‌ها یا کشف زیرمجموعه‌های جالب برای شکل‌دادن فرضیه‌ها برای اطلاعات پنهان ادامه می‌دهد. براساس رویکرد پیشنهادی، دو سری داده (پایگاه دادهٔ لاگ و پایگاه دادهٔ مشتریان) ضروری است. بنابراین، داده‌های ۱۰۰.۰۰۰ کاربر تصادفی و یکتا در بازهٔ شش‌ماه دوم سال ۱۳۹۳ از پایگاه دادهٔ لاگ جمع‌آوری شده و داده‌های مربوط به سبد این کاربران هم از پایگاه داده مشتریان گرفته شده است. داده‌های دموگرافیک این کاربران به دلیل وجود برخی مسائل امنیتی و فنی در دسترس نبودند؛ از این‌رو، از استفاده از آنها صرف‌نظر می‌شود. تعداد نشست‌های این کاربران در حدود ۱.۲ میلیون و داده‌های لاگ در حدود ۷.۵ میلیون رکورد و داده‌های مربوط به سبد خرید این کاربران در حدود ۱۴۲ هزار رکورد بود. در پایگاه دادهٔ مشتریان کلیهٔ اطلاعات مربوط به پروفایل مشتریان ذخیره می‌شوند؛ اما در این پژوهش، داده‌های سبد خرید مشتریان شامل شناسهٔ کاربر، شمارهٔ سبد، تاریخ و زمان، مبلغ، شناسهٔ کالای موجود در سبد خرید، نام کالای موجود در سبد خرید و دستهٔ هر کالای موجود در سبد خرید و داده‌های لاگ‌سرور شامل تعداد نشست‌ها، زمان کل، صفحات مختلف، صفحات کالا، صفحات جست‌وجو، تأخر[1]، مدت مراجعه، مدت زمان بین نشست‌ها، کالاهای سبد، ارزش مالی، دسته‌های کالاهای سبد و برچسب دسته است. برای ایجاد مسیرهای طی‌شدهٔ کاربر در هر نشست و همچنین تکمیل مسیرهای آنها برنامه‌ای به زبان برنامه‌نویسی پایتون[2] نوشته شد و برای هر کاربر مسیرهای هر نشست و زمان صرف‌شده در هر نشست آن به دست آمد. پس از به دست آوردن مسیر دسترسی هر کاربر در هر نشست، با ترکیب ویژگی‌های رفتاری استخراج‌شده از سبد کالا و لاگ‌سرور بر اساس مدل RFM، ویژگی‌های مشترک آنها شامل ۱۱ ویژگی تعداد نشست‌ها، زمان کل، صفحات مختلف، صفحات کالا، صفحات جست‌وجو، تأخر، مدت زمان بین نشست‌ها، کالاهای سبد، ارزش مالی، دسته‌های کالاهای سبد، برچسب دسته از پایگاه دادهٔ لاگ و مشتریان استخراج شد.

وظیفهٔ اصلی در کاوش کاربری وب، پیش‌پردازش داده است که شامل پاک‌سازی داده، شناسایی کاربر، شناسایی نشست، کامل‌کردن مسیر و شکل‌دهی داده است (پامونتا و همکاران، ۲۰۱۲). سه مرحلهٔ نخست از مراحل پیش‌پردازش داده، در هنگام ذخیره‌سازی لاگ‌ها در جداول پایگاه داده انجام شد. تنها یک مرحلهٔ پاک‌سازی دیگر روی داده‌های لاگ انجام شد. به این صورت که درخواست‌های تکراری پشت سر هم در هر نشست و نشست‌هایی که تنها یک درخواست

[1] . Recency
[2] . Python

برای آنها ثبت شده بود، از داده‌ها حذف شد. همچنین تمامی داده‌ها از لحاظ داشتن مقادیر ازدست‌رفته و داده‌های غیرطبیعی بررسی شدند. پس از پاک‌سازی، تعداد کاربران به ۷۴.۳۴۴ رسید. پس از استخراج ویژگی‌ها و حذف داده‌های پرت نوبت به نرمال‌سازی آنها می‌رسد. به دلیل وجود محدوده‌ها و مقیاس‌های مختلف اندازه‌گیری برای هر ویژگی، لازم است برای یکسان‌کردن این محدوده‌ها، تمامی ویژگی‌ها نرمال شوند. در این پژوهش، برای نرمال‌سازی داده‌ها از روش نرمال‌سازی مینیمم-ماکزیمم استفاده شده است. در این روش نرمال‌سازی، تبدیل خطی روی داده‌های اصلی انجام می‌دهد. فرض کنید که $\min_A$ و $\max_A$، کمترین و بیشترین مقدار ویژگی A باشند؛ سپس نرمال‌سازی مینیمم-ماکزیمم، مقداری مانند v از ویژگی A را به v' در محدوده $[newmin_A, newmax_A]$ می‌نگارد. رابطهٔ (۱) چگونگی این نگاشت را نشان می‌دهد:

رابطهٔ (۱)

$$v' = \frac{v - min_A}{max_A - min_A}(newmax_A - newmin_A) + newmin_A$$

فاز ٤. مدلسازی

در فاز مدل‌سازی، برای شناخت گروه‌های مختلف مشتریان از تکنیک خوشه‌بندی و به‌منظور بررسی رفتارها با توجه به ویژگی‌های رفتاری کاربران و ایجاد مدلی برای پیش‌بینی نوع رفتار مشتریان آینده از تکنیک درخت تصمیم استفاده می‌شود.

- **شناخت گروه‌های مشتریان**

اولین هدف این پژوهش، شناخت الگوها و گروه‌های مختلف مشتریان بر مبنای رفتار آنهاست. بدین منظور از الگوریتم خوشه‌بندی کا-میانگین استفاده می‌شود که باید تعداد خوشه‌های آن از پیش تعیین شوند و در ادامه بر مبنای تعداد خوشه‌های اولیه داده‌ها در خوشه‌های مختلف قرار می‌گیرند (مقدم و همکاران، ۲۰۱۷). یکی از معایب الگوریتم کا-میانگین این است که تعداد خوشه‌ها را همچون پارامتر ورودی الگوریتم می‌گیرد و نمی‌تواند تعداد خوشه‌های بهینه را بیابد. بنابراین، با استفاده از شاخص دیویس بولدین الگوریتم برای تعداد خوشه‌های مختلف امتحان شده و بهترین نتیجه برای تعداد خوشه‌ها انتخاب می‌شود. براساس سه معیار شباهت مختلف (فاصلهٔ اقلیدسی، فاصلهٔ چبیشف و فاصلهٔ منهتن) و برای تعداد خوشه‌های مختلف، الگوریتم کا-میانگین اجرا می‌شود تا زمانی که معیار دیویس بولدین آن به بالای عدد یک برسد. نتایج اعتبارسنجی الگوریتم کا-میانگین با استفاده از شاخص دیویس بولدین در جدول ۱ نشان داده شده است. با توجه به نتایج به‌دست‌آمده بهترین جواب برای شاخص دیویس بولدین مربوط به معیار فاصلهٔ اقلیدسی و با چهار خوشه است. درواقع هرچه مقدار این شاخص کمتر باشد، خوشه‌ها در بیشترین فاصله از هم قرار دارند.

جدول ۱. نتایج اعتبارسنجی الگوریتم کا-میانگین با استفاده از شاخص دیویس بولدین

تعداد خوشه	فاصلهٔ چبیشف	فاصلهٔ اقلیدسی	فاصلهٔ منهتن
۳	۰.۸۰۸	۰.۸۱۹	۰.۸۲۹
٤	۰.۷۲۵	**۰.۷۱٦**	۰.۷۱۸

سپس خوشه‌بندی با پارامترهای چهار خوشه و معیار شباهت برابر با فاصلهٔ اقلیدسی انجام می‌شود (جدول ۲). با توجه به جدول ۲، بیشترین تعداد مشتریان در خوشهٔ اول و کمترین تعداد مشتریان در خوشهٔ دوم قرار گرفته‌اند. نقاط مرکزی هر چهار خوشه که درواقع میانگین فواصل نقاط موجود در آن خوشه به ازای تمامی ویژگی‌های به‌کاررفته، در جدول ۲ آورده شده است.

جدول ۲. نتایج خوشه‌بندی با چهار خوشه و نقاط مرکزی

خوشه	تعداد	نقاط مرکزی									
		تعداد نشست	زمان کل	صفحات مختلف	صفحات کالا	صفحات جستجو	تأخر	مدت مراجعه	زمان بین نشست‌ها	کالاهای سبد	ارزش مالی	دستهٔ کالاهای سبد
۱	۲۶۴۴۱	۰	۰.۰۵	۰.۰۲	۰.۰۲	۰.۰۲	۰.۴۶	۰	۱	۰.۰۱	۰.۰۱	۰.۰۳
۲	۱۲۲۸۰	۰.۰۴	۰.۰۵	۰.۰۲	۰.۰۲	۰.۰۲۰	۰.۱۳	۰.۶۴	۰.۰۸	۰.۰۰	۰.۰۱	۰.۰۲
۳	۱۷۷۳۶	۰.۰۱	۰.۰۵	۰.۰۲	۰.۰۲	۰.۰۲	۰.۶۸	۰.۱۱	۰.۰۳	۰.۰۱	۰.۰۱	۰.۰۳
۴	۱۷۸۸۷	۰.۰۱	۰.۰۵	۰.۰۲	۰.۰۲	۰.۰۲	۰.۱۵	۰.۱۲	۰.۰۳	۰.۰۱	۰.۰۱	۰.۰۳
کل	۷۴۳۴۴	–	–	–	–	–	–	–	–	–	–	–

• دسته‌بندی مشتریان براساس ویژگی‌های رفتاری

برای دسته‌بندی رفتار مشتریان شرکت XYZ مشکل توزیع نامتوازن کلاس‌ها وجود دارد. به بیان دیگر، دستهٔ خرید در میان دو دستهٔ دیگر، کلاس اقلیت است و نمونه‌های بسیار کمتری در مقایسه با دو دستهٔ دیگر دارد؛ پس باید یک دسته‌بند ایجاد شود که هزینهٔ کلی دسته‌بندی نادرست را کمینه کند. برای مجموعهٔ داده‌های بزرگ با بیش از ۱۰.۰۰۰ نمونه، الگوریتم یادگیری حساس به هزینه بهترین نتایج را نسبت به روش‌های نمونه‌برداری ایجاد می‌کند (ویس و همکاران، ۲۰۰۷). بر این اساس، در این پژوهش نیز با داشتن مجموعهٔ داده بزرگ از روش یادگیری حساس به هزینه استفاده می‌شود. بهترین نسبت هزینه در این پژوهش هم به‌صورت تجربی و براساس مجموعه اعتبارسنجی تعیین شده است. هدف اصلی این پژوهش از دسته‌بندی، علاوه بر مدل‌سازی برای پیش‌بینی رفتار مشتریان، استخراج قوانین به صورت واضح و دقیق به‌منظور استفاده در کسب‌وکار است. انتخاب الگوریتم درخت تصمیم C5.0 هم بر این مبنا صورت گرفته است؛ زیرا قوانین درخت تصمیم بسیار ساده و قابل تفسیر است. در این الگوریتم فیلدهای ورودی ویژگی‌های به‌دست‌آمده و فیلد خروجی برچسب‌های کاربران است. سپس داده‌ها با استفاده از گره پارتیشن به سه گروه آموزشی، آزمایشی و اعتبارسنجی تقسیم می‌شوند. در گره C5، برای افزایش قابلیت اطمینان اعتبارسنجی مدل، گزینهٔ اعتبارسنجی متقابل با ده تکرار انتخاب شده است. این روش بر پایهٔ تقسیم مجموعه‌داده به ۱۰ قسمت مساوی است که در آن ۹ قسمت از مجموعه‌داده آموزش مدل و بقیهٔ آزمایش مدل را انجام می‌دهند. همچنین ساخت درخت تصمیم با استفاده از حالت ساده و مطلوب تعمیم انجام شده است.

این حالت باعث می‌شود بیشتر پارامترهای درخت به صورت خودکار تنظیم شود. به دلیل کم‌بودن تعداد نمونه‌های خرید مشتریان و ارزشمندی این دسته در پژوهش، در ماتریس هزینهٔ درخت، هزینهٔ تشخیص نادرست مشتریان واقعی با عنوان دسته‌های دیگر افزایش یافته است. پس از اجرای الگوریتم، پایین‌ترین سطح شکست‌ها دوباره بررسی می‌شوند و آنهایی که کمک شایانی به ارزش مدل نمی‌کنند، از درخت حذف یا به عبارتی دیگر هرس می‌شوند. پس از اجرای این روند، درختی با میانگین دقت ۶۳.۶٪ و خطای استاندارد ۰.۳ به دست آمد. در نهایت تعداد ۲۶۱ قانون با اطمینان ۷۰٪ حاصل شد.

فاز ۵. ارزیابی

• ارزیابی خوشه‌ها

پس از مشخص‌شدن نقاط مرکزی خوشه‌ها، به منظور بررسی این‌که به ازای تمامی ویژگی‌ها آیا

میانگین‌های به‌دست‌آمده، اختلاف معناداری با هم دارند یا خیر، آزمون آنووا در سطح معناداری کمتر از ۰.۰۵ اجرا می‌شود. برای این کار، چهار میانگین به‌دست‌آمده به ازای هر ویژگی با اجرای این آزمون با هم مقایسه می‌شوند. همان‌طور که از جدول ۳ و نتایج به‌دست‌آمده مشخص است، با بررسی ستون سطح معناداری مشاهده می‌شود که این مقدار برای تمامی متغیرها برابر با ۰.۰۰۰ و کمتر از ۰.۰۵ است؛ ازاین‌رو فرض همگن‌بودن میانگین‌های جامعه رد می‌شود و نشان می‌دهد گروه‌ها میانگین‌های مختلفی دارند.

جدول ۳. نتایج آزمون آنوا برای الگوریتم‌های کا-میانگین

متغیر	منبع تغییر	مجموع مجذورات	درجهٔ آزادی	میانگین مجذورات	آزمون F	معناداری
تعداد نشست‌ها	بین گروهی	۱۲.۸۸۹	۳	۴.۲۹۶	۶۰۰۵.۲۵۳	۰.۰۰۰
	درون گروهی	۵۳.۱۸۴	۷۴۳۴۰	۰.۰۰۱		
	مجموع	۶۶.۰۷۳	۷۴۳۴۳	–		
زمان کل	بین گروهی	۰.۴۲۰	۳	۰.۱۴۰	۶۵.۳۶۹	۰.۰۰۰
	درون گروهی	۱۵۹۰۲۸۱	۷۴۳۴۰	۰.۰۰۲		
	مجموع	۱۵۹.۷۰۱	۷۴۳۴۳	–		
صفحات مختلف	بین گروهی	۰.۳۱۲	۳	۰.۱۰۴	۱۴۹.۵۹۴	۰.۰۰۰
	درون گروهی	۵۱.۶۳۰	۷۴۳۴۰	۰.۰۰۱		
	مجموع	۵۱.۹۴۲	۷۴۳۴۳	–		
صفحات کالا	بین گروهی	۰.۱۹۷	۳	۰.۰۶۶	۱۳۲.۴۱۵	۰.۰۰۰
	درون گروهی	۳۶.۹۲۸	۷۴۳۴۰	۰.۰۰۰		
	مجموع	۳۷.۱۲۶	۷۴۳۴۳	–		
صفحات جستجو	بین گروهی	۰.۰۲۲	۳	۰.۰۰۷	۱۳.۴۷۷	۰.۰۰۰
	درون گروهی	۴۱.۰۰۵	۷۴۳۴۰	۰.۰۰۱		
	مجموع	۴۱.۰۲۸	۷۴۳۴۳	–		
تأخر	بین گروهی	۳۴۱۹.۹۶۰	۳	۱۱۳۹.۹۸۷	۲۴۵۱۹.۳۲۴	۰.۰۰۰
	درون گروهی	۳۴۵۶.۳۱۹	۷۴۳۴۰	۰.۰۴۶		
	مجموع	۶۸۷۶.۲۷۸	۷۴۳۴۳	–		
مدت مراجعه	بین گروهی	۳۶۳۹.۱۸۱	۳	۱۲۱۳.۰۶۰	۱۰۰۳۰۲.۵۷۶	۰.۰۰۰
	درون گروهی	۸۹۹.۰۶۹	۷۴۳۴۰	۰.۰۱۲		
	مجموع	۴۵۳۸.۲۵۰	۷۴۳۴۳	–		
زمان بین	بین گروهی	۱۵۵۶۶.۱۹۳	۳	۵۱۸۸.۷۳۱	۱۵۳۳۶۶۷.۵۱۸	۰.۰۰۰
	درون	۲۵۱.۵۰۸	۷۴۳۴۰	۰.۰۰۳		

متغیر	منبع تغییر	مجموع مجذورات	درجهٔ آزادی	میانگین مجذورات	آزمون F	معناداری
نشست‌ها	گروهی					
	مجموع	۱۵۸۱۷.۷۰۲	۷۴۳۴۳	–		
کالاهای سبد	بین گروهی	۰.۱۱۲	۳	۰.۰۳۷	۱۳۳.۵۳۹	۰.۰۰۰
	درون گروهی	۲۰.۸۴۸	۷۴۳۴۰	۰.۰۰۰		
	مجموع	۲۰.۹۶۱	۷۴۳۴۳	–		
ارزش مالی	بین گروهی	۰.۴۳۴	۳	۰.۱۴۵	۱۰۱.۵۴۹	۰.۰۰۰
	درون گروهی	۱۰۵.۹۱۰	۷۴۳۴۰	۰.۰۰۱		
	مجموع	۱۰۶.۳۴۴	۷۴۳۴۳	–		
دسته‌های کالای سبدها	بین گروهی	۱.۵۹۴	۳	۰.۵۳۱	۳۳۰.۴۱۶	۰.۰۰۰
	درون گروهی	۱۱۹.۵۳۸	۷۴۳۴۰	۰.۰۰۲		
	مجموع	۱۲۱.۱۳۲	۷۴۳۴۳	–		

• تحلیل خوشه‌ها

برای تحلیل بهتر خوشه‌ها، نخست بررسی می‌شود در هر خوشه هر مشتری چه برچسبی دارد، هر کدام از این گروه‌ها چه درصدی از کل خوشه را به خود اختصاص داده‌اند و همچنین چند درصد از کل هر گروه در هر خوشه است. نتایج بررسی تعداد هر برچسب در هر خوشه در جدول ۴ آورده شده است.

جدول ٤. نتایج حاصل از بررسی تعداد برچسب در هر خوشه

خوشه	برچسب	تعداد نمونه	درصد در خوشه	درصد در برچسب	تعداد کل
خوشهٔ اول	خرید	۴	۰.۰۱	۱.۳۴	۲۶۴۴۱
	انتظار	۱۰۸۷۸	۴۱.۱۴	۳۶	
	خریدنکردن	۱۵۵۵۹	۵۸.۸۵	۳۵.۵	
خوشهٔ دوم	خرید	۱۷۱	۱.۴	۵۸.۳۳	۱۲۲۸۰
	انتظار	۳۲۰۵	۲۶.۱	۱۰.۶	
	نخریدن	۸۹۰۴	۷۲.۵	۲۰.۳	
خوشهٔ سوم	خرید	۷۳	۰.۴۱	۲۵	۱۷۷۳۶
	انتظار	۸۸۰۴	۴۹.۵۹	۲۹.۱	
	خریدنکردن	۸۸۵۹	۵۰	۲۰.۲	
خوشه چهارم	خرید	۴۵	۰.۲۵	۱۵.۳۳	۱۷۸۸۷
	انتظار	۷۳۶۳	۴۱.۱۵	۲۴.۳	
	خریدنکردن	۱۰۴۷۹	۵۸.۶	۲۴	

باتوجه به نتایج جدول ۴ خوشه‌ها تحلیل می‌شود:

خوشهٔ اول: در خوشهٔ نخست که بیشترین تعداد

کاربران را داشت، کاربرانی‌اند که در بازهٔ مطالعه‌شده، تنها یک بار از سایت بازدید داشته و نسبت به خوشه‌های دیگر از صفحات مختلف بیشتری بازدید کرده‌اند. حدود نیمی از کاربران، برخی از این کالاها را به سبد خرید خود اضافه کرده‌اند؛ اما خرید خود را نهایی نکرده و سبد خرید آنها در حالت باز قرار دارد. نیمی دیگر از کاربران هم تمایلی به تشکیل سبد نداشته و تنها به بازدید از صفحات اکتفا کرده‌اند. در این خوشه، تعداد کاربرانی که بازدیدشان به خرید ختم شده بسیار محدود است.

خوشهٔ دوم: در خوشهٔ دوم که کمترین تعداد کاربران را داشت، کاربرانی هستند که در بازهٔ مطالعه‌شده، بیشترین دفعات بازدید از سایت را داشته‌اند. نسبت به دیگر خوشه‌ها، کمترین مدت زمان بازدید از سایت و کمترین تعداد صفحات مختلف بازدیدشده در هر مراجعه را دارند و بیشتر به جست‌وجوی کالاها می‌پردازند. فاصلهٔ زمانی بین بازدیدهای این گروه از خوشه‌های دیگر بیشتر است. بیشترین تعداد سبدهایی که به خرید ختم شده است، در این گروه قرار دارد که می‌توان کاربران این خوشه را دارای احتمال خرید دانست.

خوشهٔ سوم: کاربرانی که در خوشهٔ سوم‌اند، در بازهٔ زمانی کوتاهی از سایت بازدید داشته‌اند. همچنین کمترین فاصلهٔ زمانی بین بازدیدها هم متعلق به این گروه است. بعد از خوشهٔ اول، بیشترین صفحات بازدیدشده متعلق به این خوشه است که می‌توان نتیجه گرفت بیشتر به جست‌وجوی کالاها در سایت و بازدید از آنها پرداخته‌اند. بعد از خوشهٔ اول، بیشترین تعداد سبدهای باز در این خوشه قرار دارد. سبدهای این گروه از کاربران دارای بیشترین تعداد کالاست که از متنوع‌ترین سبدها هم محسوب می‌شود. همچنین مبالغ سبدهای تشکیل‌شده از سایر خوشه‌ها بیشتر است.

خوشهٔ چهارم: کاربرانی که در خوشهٔ چهارم‌اند، همانند خوشهٔ سوم تنها در بازهٔ زمانی کوتاهی از سایت بازدید کرده‌اند؛ با این تفاوت که این بازدیدها اغلب در اواخر بازهٔ مطالعه‌شده انجام گرفته است. همچنین فاصلهٔ زمانی بین بازدیدها در این گروه کم است. بعد از خوشهٔ اول، بیشترین تعداد کاربرانی که تمایلی به تشکیل سبد نداشته‌اند، در این خوشه است؛ با این حال، باز هم سبدهای این گروه از کاربران دارای بیشترین تعداد و متنوع‌ترین کالاها بعد از خوشهٔ سوم است.

- **ارزیابی دسته‌بندی**

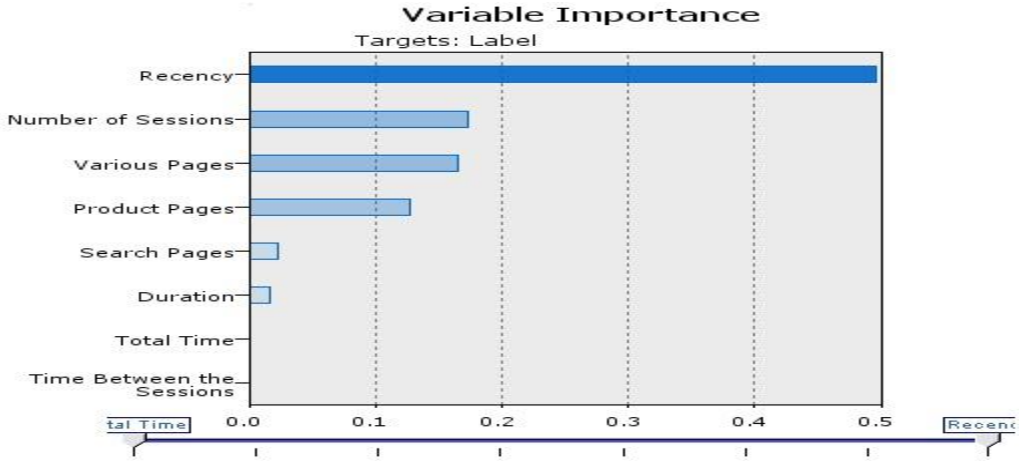
برای ارزیابی درخت تصمیم از ماتریس درهم‌ریختگی استفاده می‌شود. این ماتریس ابزار مفیدی برای تحلیل چگونگی عملکرد روش دسته‌بندی در تشخیص داده‌ها یا مشاهدات دسته‌های مختلف است. مهم‌ترین معیار برای تعیین کارایی تکنیک دسته‌بندی، معیار دقت است. این پژوهش به دقت دسته‌بندی ۶۳.۶٪ دست یافت که با توجه به نوع دسته‌ها و تعداد نمونه‌های موجود در هر دسته، دقت قابل قبولی است. معیارهایی که به صورت جداگانه عملکرد یک دسته‌بند[1] را روی دسته‌های مختلف برآورد می‌کنند حساسیت، شفافیت و صحت‌اند. این شاخص‌ها برای هر سه دسته با استفاده از ماتریس درهم‌ریختگی محاسبه شده‌اند. نتایج ارزیابی درخت تصمیم در جدول ۵ ارائه شده است.

[1] . Classifier

جدول ۵. نتایج ارزیابی درخت تصمیم

برچسب	Recall+	Recall-	Precison+	Precision-
خرید	٪۸۰	٪۷۸	٪۱.۶	٪۹۹
انتظار خرید	٪۶۶	٪۸۳	٪۷۶	٪۷۴
خریدنکردن	٪۶۵	٪۷۷	٪۸۲	٪۵۶

با توجه به نامتعادل‌بودن دسته‌ها و دقت به‌دست‌آمده از اعتبارسنجی متقابل با ده تکرار یعنی ٪۶۳.۶، ثابت می‌شود که مدل دارای کارایی قابل قبولی است. براساس مدل (شکل ۱)، تأخر مهم‌ترین و مدت زمان و زمان بین نشست‌ها کم‌اهمیت‌ترین ویژگی‌ها در دسته‌بندی بوده‌اند.



شکل ۱. ویژگی‌های مهم در دسته‌بندی

فاز ٦. به‌کارگیری

هدف از دسته‌بندی داده‌ها در این پژوهش، ارائهٔ مدلی برای پیش‌بینی رفتار مشتریان و کشف دانشی مفید از پایگاه داده‌هاست. این مرحله شامل برنامه‌ریزی برای دانش کشف‌شده است که کجا و چگونه به کار رود. برنامه‌ای هم برای نظارت بر پیاده‌سازی دانش کشف‌شده باید ایجاد و تمام پروژه مستندسازی شود. کشف دانش از راه بررسی مدل و قوانین ایجاد شده انجام می‌گیرد. در این پژوهش ۲۶۱ قانون با اطمینان ٪۷۰ حاصل شد که تعدادی از آنها به شرح زیر است:

۱- اگر کاربر هیچ صفحهٔ کالایی را مشاهده نکند، اصلاً خرید نخواهد کرد.

۲ - کاربرانی با مدت مراجعه یک‌روزه و داشتن تعداد نشست‌های کمتر از ۱۲ و تعداد صفحات مختلف بزرگ‌تر از ۶، از سایت خرید نمی‌کنند.

۳- کاربرانی که تعداد نشست‌هایشان کمتر از ۱۲ و تعداد صفحات مختلف کمتر از ۶ و تعداد صفحات جست‌وجوی آنها بیشتر از ۳ است، خرید نمی‌کنند.

۴- کاربرانی که ۹ صفحه کالا را دیده‌اند و تعداد نشست‌هایشان بیشتر از ۱۲ است، زمانی را که درون سایت می‌گذرانند، بیشتر از ۲۸ دقیقه است و همچنین دارای مدت مراجعه بزرگ‌تر از ۶۷ هستند، از سایت خرید می‌کنند.

۵- کاربرانی که تأخرشان بین ۸۲ تا ۸۵ روز است،

دارای تعداد نشست بالای ۱۲ و تعداد صفحات مختلف بیشتر از ۶ هستند، فقط تمایل به تشکیل سبد دارند.

۵- نتیجه‌گیری

شرکت‌های تجارت الکترونیک از دنیای قدیمی که در آن، محصولات استاندارد، بازارهای همگن و چرخهٔ توسعه و عمر محصول طولانی یک قانون بود، به دنیای جدیدی انتقال می‌یابند که در آن انواع محصولات استانداردِ جایگزین وجود دارند. مصرف‌کنندگان می‌توانند از بین میلیون‌ها کالا در یک فروشگاه برخط به جای ده‌ها هزار کالا در فروشگاه‌های بزرگ، کالاهای خود را انتخاب کنند. بنابراین، پیش‌بینی رفتار مشتریان دشوار است؛ چون به‌ندرت بازدید آنها از سایت‌ها به خرید ختم می‌شود؛ از این‌رو، برای داشتن کسب‌وکار برخط موفق باید به تحلیل رفتار مشتریان پرداخت. استفاده از داده‌کاوی و وب‌کاوی به ما در تحلیل رفتار مشتریان و کشف دانش در این زمینه کمک می‌کند. این دانش، سازمان‌ها را قادر می‌سازد تا تصمیمات مبتنی بر داده گرفته و راهبردهای تصمیم‌گیری خود را بهبود و توسعه دهند. همچنین برای به دست آوردن مشتریان جدید و حفظ مشتریان موجود و بهبود رضایت مشتری نیز کمک می‌کند. از آنجا که به نظر می‌رسد در ایران پژوهشی در این زمینه از تجارت الکترونیک وجود ندارد، این پژوهش با هدف طرح چارچوبی برای افزایش دقت تحلیل و شناخت گروه‌های مشتریان و همچنین ارائهٔ مدل و قوانینی برای پیش‌بینی رفتار مشتریان بر مبنای ویژگی‌های رفتاری آنها انجام شده است. در این پژوهش، با استفاده از رویکرد ترکیبی داده‌کاوی و کاوش کاربری وب، رفتار مشتریان تحلیل شده است؛ این روش مزیت رقابتی برای شرکت را باعث می‌شود. درواقع نوآوری این پژوهش اِعمال روش‌های داده‌کاوی بر ابعاد مختلف داده‌های مشتری شامل داده‌های رفتاری مرور وب و داده‌های رفتار خرید است که به بهبود دانش از مشتری می‌انجامد. این پژوهش مبتنی بر داده‌های واقعی است و داده‌های واقعیِ مشتریان شرکت XYZ را بررسی می‌کند که محدودهٔ فعالیت آن، تجارت الکترونیک و از نوع خرده‌فروشی اینترنتی است. در این پژوهش اطلاعات حساب‌های کاربری مشتریان به همراه اطلاعات موجود در لاگ‌های وب سرور که در پایگاه دادهٔ شرکت ذخیره می‌شوند، استفاده شده است. شرکت XYZ به‌منظور درک سازمانی بهتر نیازها، خواسته‌ها و تمایلات مشتریان و همچنین وجود انسجام و دیدی مشخص در سازمان برای پیام‌های بازاریابی و تعریف مشتریان بالقوه، می‌تواند از خوشه‌های به‌دست‌آمده در این پژوهش، برای این منظور بهره ببرد و به شناخت دقیقی از مشتریان خود دست یابد و می‌تواند راهبردهای مقتضی را در هر گروه پیاده‌سازی کند و احتمال موفقیت خود را افزایش دهد. در کل خرده‌فروشان برخط باید فعالیت‌های بازاریابی خود در رسانه‌های اجتماعی را گسترش داده تا بتوانند متناسب با نیازها و فعالیت‌های مشتریان خود (کاربران این رسانه‌ها) عمل کنند و فعالیت‌های جذب مشتری جدید و حفظ مشتریان موجود را ادامه دهند.

محدودیت‌های زیادی در زمینهٔ استفاده از داده‌های دنیای واقعی برای تحلیل رفتار مشتریان وجود دارد. به دلیل وجود بازار رقابتی، شرکت‌ها به‌سادگی داده‌های مشتریان خود را در اختیار قرار نمی‌دهند یا در صورت انجام این کار، داده‌های دست‌کاری‌شده و ناقص در اختیار آنها قرار می‌دهند. در واقع مدیران شرکت برای سپردن اطلاعات به افراد خارج از سازمان تمایلی ندارند و چون نتایج داده‌کاوی کاملاً به صحت داده‌های اولیه بستگی دارد، این امر ممکن است سبب به دست آمدن

نتایج نامعتبر شود. همچنین در برخی از شرکت‌ها، ممکن است کارمندان به دلیل ترسِ از دست دادن موقعیت شغلی خود، حاضر به همکاری با پژوهشگران نباشند و در شناخت کسب‌وکار و داده‌ها کمک چندانی نکنند.

در این پژوهش به دلیل زیاد بودن حجم داده‌ها و کم‌بودن مقدار حافظه، امکان اجرای تکنیک‌های دیگر داده‌کاوی مانند الگوریتم SOM وجود نداشت؛ ازاین‌رو پیشنهاد می‌شود پژوهش‌های آینده از سایر الگوریتم‌های داده‌کاوی که به مشخص‌کردن تعداد خوشه‌های بهینه نیاز ندارند و خودشان این تعداد را به دست می‌آورند، برای خوشه‌بندی استفاده کنند.

در این پژوهش داده‌های دموگرافیک مشتریان شامل جنس، سن، مکان سکونت، تحصیلات و غیره به‌دلیل برخی مسائل فنی و امنیتی دردسترس نبودند؛ از این‌رو پیشنهاد می‌شود این ویژگی تحلیل رفتار مشتریان بررسی شود. علاوه بر آن، پیشنهاد می‌شود ویژگی‌های دیگری مانند نوع خرید، روش خرید، نوع پرداخت و غیره استخراج و تأثیر آنها در این رویکرد ترکیبی بررسی شود. هم‌چنین بررسی تأثیر وب‌کاوی بر ارزش مشتری برای پاسخ به این سؤال که آیا وب‌کاوی می‌تواند ارزش مشتری را پیش‌بینی کند یا خیر، نیز می‌تواند در پژوهش‌های آینده بررسی شود.

منابع

1- Ahn, T., Ryu, S., & Han, I. (2007). The impact of Web quality and playfulness on user acceptance of online retailing. *Information & management*, *44*(3), 263-275.

2- Arora, M., & Chopra, A. B. (2016). Impact of online selling on physical retail in India. *International Journal of Research in IT and Management*, *6*(10), 57-68.

3- Azevedo, A. I. R. L., & Santos, M. F. (2008). KDD, SEMMA and CRISP-DM: a parallel overview. *IADS-DM*.

4- Bae, S. M., Park, S. C., & Ha, S. H. (2003). Fuzzy web ad selector based on web usage mining. *IEEE intelligent Systems*, *18*(6), 62-69.

5- Beheshtian-Ardakani, A., Fathian, M., & Gholamian, M. (2018). A novel model for product bundling and direct marketing in e-commerce based on market segmentation. *Decision Science Letters*, *7*(1), 39-54.

6- Deka, P. K. (2017). A Conceptual Model for Determining Factors Influencing Online Purchasing Behavior. *Journal of Management in Practice*, *2*(1).

7- Dharmarajan, K., & Dorairangaswamy, D. M. (2016). Web Usage Mining: Improve The User Navigation Pattern Using Fp-Growth Algorithm. *Elysium journal of engineering research and management (EJERM)*, *3*(4).

8- Facca, F. M., & Lanzi, P. L. (2005). Mining interesting knowledge from weblogs: a survey. *Data & Knowledge Engineering*, *53*(3), 225-241.

9- Ha, S. H. (2002). Helping online customers decide through web personalization. *IEEE Intelligent systems*, *17*(6), 34-43.

10- Hsieh, N. C. (2004). An integrated data mining and behavioral scoring model for analyzing bank customers. *Expert systems with applications*, *27*(4), 623-633.

11- Hung, Y. S., Chen, K. L. B., Yang, C. T., & Deng, G. F. (2013). Web usage mining for analysing elder self-care behavior patterns. *Expert Systems with Applications*, *40*(2), 775-783.

12- Jenamani, M., Mohapatra, P. K., & Ghose, S. (2003). A stochastic model of e-customer behavior. *Electronic Commerce Research and Applications*, *2*(1), 81-94.

13- Kim, E., Kim, W., & Lee, Y. (2003). Combination of multiple classifiers for the customer's purchase behavior prediction. *Decision Support*

Systems, *34*(2), 167-175.

14- Liou, J. J., & Tzeng, G. H. (2010). A dominance-based rough set approach to customer behavior in the airline market. *Information Sciences*, *180*(11), 2230-2238.

15- Liu, B., Mobasher, B., & Nasraoui, O. (2011). Web usage mining. In *Web Data Mining* (pp. 527-603). Springer Berlin Heidelberg.

16- Moghaddam, Q., S. Abdolvand, N., & Harandi, R. S. (2017). A RFMV Model and Customer Segmentation Based on Variety of Products. *Information Systems & Telecommunication*, 5(3), 155- 161

17- Pamutha, T., Chimphlee, S., Kimpan, C., & Sanguansat, P. (2012). Data preprocessing on web server log files for mining users access patterns. *International Journal of Research and Reviews in Wireless Communications (IJRRWC) Vol*, *2*.

18- Park, J., & Chung, H. (2009). Consumers' travel website transferring behaviour: analysis using clickstream data-time, frequency, and spending. *The Service Industries Journal*, *29*(10), 1451-1463.

19- Shanthi, S. (2017). Survey on Web Usage Mining using Association Rule Mining. *International Journal of Innovative Computer Science & Engineering*, (4) 3; 65-67

20- Sheikh, A. M., & Menaria, S. (2017). An Approach of Security in E-Commerce with Web Mining Framework. *International Education and Research Journal*, *3*(5).

21- Sisodia, D. S., & Verma, S. (2012, May). Web usage pattern analysis through web logs: A review. In *Computer Science and Software Engineering (JCSSE), 2012 International Joint Conference on* (pp. 49-53). IEEE.

22- Sun, L., & Zhang, X. (2004, April). Efficient frequent pattern mining on web logs. In *Asia-Pacific Web Conference* (pp. 533-542). Springer, Berlin, Heidelberg.

23- Theusinger, C., & Huber, K. P. (2000, August). Analyzing the footsteps of your customers. In *Proc. of the Sixth ACM SIGKDD Internat. Conf. on Web KDD 2000* (pp. 44-52).

24- Tsai, C. F., Hu, Y. H., & Lu, Y. H. (2015). Customer segmentation issues and strategies for an automobile dealership with two clustering techniques. *Expert Systems*, *32*(1), 65-76.

25- Weiss, G. M., McCarthy, K., & Zabar, B. (2007). Cost-sensitive learning vs. sampling: Which is best for handling unbalanced classes with unequal error costs?. *DMIN*, *7*, 35-41.

26- Yang, Z., & Su, X. (2012). Customer behavior clustering using SVM. *Physics Procedia*, *33*, 1489-1496.

27- Zhang, X., Gong, W., & Kawamura, Y. (2004, January). Customer behavior pattern discovering with web mining. In *APWeb* (pp. 844-853).