مجله محاسبات نرم، جلد 13، شماره 2 (پاییز و زمستان 1403)، صفحه 78_93
https://doi.org/10.22052/scj.2024.253114.1156

مجله محاسبات نرم
SOFT COMPUTING JOURNAL
تارنمای مجله: scj.kashanu.ac.ir

University of Kashan

تشخیص اولین لحظه حقیقت در خرید برخط با استفاده از روش‌های پیش‌پردازش داده‌ها و طبقه‌بندهای تلفیقی ✧

محسن امیرافضلی[1*]، دانشجوی دکتری، حسین غفاریان[1]، استادیار

[1] دانشکده فنی و مهندسی، دانشگاه اراک، اراک، ایران.

اطلاعات مقاله

تاریخچه مقاله:
دریافت 25 خرداد ماه 1402
پذیرش 30 آبان ماه 1402

کلمات کلیدی:
خرید برخط
اولین لحظه حقیقت
داده‌کاوی
پیش‌پردازش
طبقه‌بند تلفیقی

چکیده

این مقاله اقدام به ارائه یک استراتژی، با هدف افزایش دقت تشخیص زودهنگام خریداران از مشتریان در حال گشت و گذار در یک فروشگاه برخط، نموده است. این روزها مردم تمایل به کاوش برخط برای پیدا کردن اقلام مورد نیاز خود و خرید از طریق تراکنش‌های برخط دارند. با این حال، تعداد خریداران واقعی هنوز در مقایسه با تعداد کل بازدیدکنندگان از این وبگاه‌ها بسیار کم است. تحلیل رفتاری، پیش‌بینی و شناسایی زودهنگام بازدیدکنندگانی که قصد خرید از فروشگاه برخط را دارند، زمینه ارائه محتوای سفارشی مناسب‌تر برای آنها را فراهم می‌آورد. از دیدگاه مدیریتی به این زمان به اصطلاح اولین لحظه حقیقت گفته می‌شود. مزیت اصلی این پیش‌بینی کاهش خطر از دست دادن کاربران با احتمال خرید بالا و افزایش نرخ تبدیل می‌باشد. به دلیل ثابت بودن چارچوب پیش‌بینی و تشخیص در داده‌کاوی، تمرکز این مقاله بر استفاده بهینه از روش‌های پیش‌پردازش، با هدف بهبود کیفیت داده‌های ورودی به الگوریتم‌های طبقه‌بندی می‌باشد. به همین دلیل، در استراتژی پیشنهادی، مجموعه‌ای از الگوریتم‌های تبدیل محتوای اسمی به عددی، نرمال‌سازی، تشخیص داده‌های پرت، انتخاب ویژگی و متوازن‌سازی بکار گرفته شده است. سپس داده‌های اصلاح شده به مجموعه‌ای از الگوریتم‌های طبقه‌بندهای مختلف، شامل درخت تصمیم C4.5 و پرسپترون چند لایه و الگوریتم‌های طبقه‌بندی تلفیقی جنگل تصادفی، Bagging و Gradient Boosting داده شده است. ارزیابی نتایج نشان می‌دهد که بیشترین مقدار دقت به دست آمده در این پژوهش با استفاده از طبقه‌بندهای تلفیقی به 94/42 % رسیده است که در مقام مقایسه با بهترین نتایج کارهای پیشین، دقت تشخیص افزایش داشته است.

1. مقدمه

تجارت الکترونیکی که شامل خرید و فروش کالا از طریق اینترنت می‌باشد، تحولات شگرفی را در زندگی روزمره انسان‌ها به همراه داشته است. یکی از ساده‌ترین و کارآمدترین نقش‌های تجارت الکترونیک در زندگی روزمره خرید برخط می‌باشد. این اقدام یا نگرش به دلیل ویژگی‌هایی از قبیل راحتی، دسترسی آسان، عدم وجود جمعیت و صرفه‌جویی در زمان که وبگاه‌های تجارت الکترونیک در اختیار کاربران قرار می‌دهند، زندگی

✧ نوع مقاله: پژوهشی

* نویسنده مسئول
پست(های) الکترونیک: m.a.afzali1991@gmail.com (امیرافضلی)
h-ghaffarian@araku.ac.ir (غفاریان)

نحوه ارجاع به مقاله: محسن امیرافضلی، حسین غفاریان، «تشخیص اولین لحظه حقیقت در خرید برخط با استفاده از روش‌های پیش‌پردازش داده‌ها و طبقه‌بندهای تلفیقی»، مجله محاسبات نرم، جلد 13، شماره 2، ص78_93، پاییز و زمستان 1403.

انسان‌ها را راحت‌تر و آسان‌تر کرده است. به گونه‌ای که در سال‌های 2015 تا 2017 درصد مشتریانی که هنوز خرید از فروشگاه‌های فیزیکی را ترجیح می‌دهند، از 85 درصد به 70 درصد کاهش یافته است [1]. در سال 2013 حدود یک میلیارد نفر در سراسر جهان کالاهای خود را به صورت برخط خریداری می‌کردند. این عدد در سال 2018 به 1/8 میلیارد نفر افزایش یافت که بیش از یک پنجم جمعیت جهان است. علاوه بر این، در سال 2017 فروش تجارت الکترونیکی در سراسر جهان بالغ بر 2/3 تریلیون دلار بوده است [2].

پس از اینکه جان کارلزون[1] در اوایل دهه 1980 عبارت لحظه حقیقت (MOT [۲]) را معرفی نمود، تاثیر عوامل اجتماعی و رفتار فروشندگان در قبال خرید مشتریان به موضوع داغی بدل شد. عبارت لحظه حقیقت به معنی زمانی است که یک مشتری وارد یک کسب و کار شده و در حال تصمیم گیری برای خرید کالا و یا خدمات است [3]. اولین زمان اخذ تصمیم مشتری برای خرید به اولین لحظه حقیقت (FMoT [۳]) معروف است. با رشد اینترنت کم کم مفاهیم جدید دیگری نظیر تصمیم‌گیری چندکاناله[4] در مباحث فروش‌های برخط و غیربرخط مورد توجه قرار گرفت. جیم لسینسکی[5] از گوگل مفهوم لحظه صفر حقیقت (ZMOT[6]) را در ابتدای قرن جدید مطرح نمود [4]. با استمرار این روند، مفاهیمی نظیر ZMOT، Showrooming و Webrooming در زمره موارد رفتاری تاثیرگذار بر فرآیند خرید مشتریان قرار گرفتند. اما در تمامی این مسائل، یک نکته بسیار مهم وجود دارد و آن این است که چگونه در یک سیستم فروش برخط باید متوجه بشویم که یک کاربر آیا به واقع با قصد و نیت خرید مراجعه کرده است و یا صرفا در شرایط حاضر برای کسب اطلاعات وارد سیستم شده است. رسیدن به پاسخ دقیق این سوال نیازمند استفاده از سیستم‌های ردیابی و تحلیل حجم انبوهی از داده‌های مرتبط با هر کاربر است که شاید تنها توسط برخی شرکت‌های بزرگ دنیا نظیر گوگل قابل انجام باشد. حال آنکه در بسیاری از کسب و کارهای برخط، چنین اطلاعاتی در دسترس نمی‌باشد. بنابراین باید به دنبال معیارهای دیگری برای رسیدن به نتیجه واضح در خصوص نیت و تصمیم کاربر بود.

در خرید فیزیکی، یک فروشنده بر اساس تجربیاتی که در طی سال‌ها کسب کرده است، به خریدار پیشنهاداتی را در مورد خرید کالا می‌دهد که این کار به شدت بر روی نرخ تبدیل[7] موثر است. نرخ تبدیل تجارت الکترونیکی نشان دهنده درصد بازدیدها از وبگاه است که به خرید منجر شده‌اند [5]. هر چه نرخ تبدیل رشد داشته باشد به معنای موفقیت وبگاه می‌باشد. با این حال، تعداد خریداران واقعی هنوز در مقایسه با تعداد کل بازدیدکنندگان از این وبگاه‌ها بسیار کم است. از آنجا که در خرید برخط هیچ تعامل فردی بین خریدار و فروشنده صورت نمی‌گیرد، با بررسی تاریخچه رفتار کاربران می‌توان رفتار بعدی آن‌ها را پیش‌بینی کرد. در یک وبگاه، دسترسی به داده‌های گذشته کاربران در هر جلسه که شامل، صفحاتی که دسترسی داشته‌اند، محصولاتی که مشاهده کرده و/یا خریداری کرده‌اند، کلیک‌هایی که انجام داده‌اند، زمانی که صرف کرده‌اند و بسیاری موارد دیگر آسان است. ویژگی‌های جلسه و اطلاعات رفتاری کاربر به عنوان دنباله‌ای از درخواست‌های HTTP ارسال شده توسط مشتری در فایل‌های log سرور وب ذخیره می‌شوند [6]. بنابراین می‌توان با تجزیه و تحلیل داده‌های قبلی خریداران، قصد خرید آن‌ها را پیش‌بینی کرد، که به یک حوزه نوظهور از تحقیقات در حوزه محاسبه و داده‌کاوی تبدیل شده است.

در زمینه تشخیص رفتار خریداران برخط کارهای متفاوتی انجام شده است. در [7] ساکار[8] و همکارانش با استفاده از مجموعه داده Online Shoppers Intention [8]، یک سیستم تجزیه و تحلیل رفتار خریداران برخط در زمان بلادرنگ طراحی کردند. سیستم پیشنهادی متشکل از دو پیمانه است که قصد خرید بازدید کننده و احتمال ترک وبگاه را پیش‌بینی می‌کند. در پیمانه اول، قصد خرید بازدید کننده با استفاده از اطلاعات صفحه

[1] Jan Carlzon
[2] Moment of Truth
[3] First Moment of Truth
[4] multi-channel decision making
[5] Jim Lecinski
[6] Zero Moment of Truth

[7] Conversion rate
[8] Sakar

نمایش جمع آوری شده در طول بازدید همراه با برخی از اطلاعات جلسه و کاربر پیش‌بینی می‌شود. در پیمانه دوم، نویسندگان از یک شبکه عصبی مکرر مبتنی بر حافظه طولانی کوتاه مدت بر اساس اطلاعات جریان کلیک متوالی استفاده کردند تا احتمال ترک وبگاه توسط بازدیدکننده، بدون انجام معامله را پیش‌بینی کنند. همچنین در این مقاله از روش‌های نمونه‌برداری بیش از حد[1] و کاهش ویژگی برای بهبود عملکرد و مقایس‌پذیری مجموعه‌ای از الگوریتم‌های یادگیری ماشین نظارت شده، استفاده شده است.

در [9] باتی[2] و همکارش یک سیستم پیش‌بینی رفتار خریداران برخط پیشنهاد کردند تا به محض بازدید از وبگاه، کاربران با احتمال بالای خرید شناسایی شده و محتوای مناسب به آنها ارائه شود. برای این منظور نویسندگان بخشی از ویژگی‌های اطلاعات جلسه و کاربر موجود در مجموعه داده [8] را استفاده کردند. آنها برای ارزیابی روش پیشنهادی خود از طبقه‌بندهای درخت تصمیم C4.5 و Naïve Bayes و جنگل تصادفی استفاده کردند. علاوه بر این آنها از روش متعادل‌سازی SMOTE برای بهبود عملکرد و مقایس‌پذیری طبقه‌بندها استفاده کرده‌اند. در [10] کبیر[3] و همکارش عملکرد الگوریتم‌های نظارت شده و روش‌های تلفیقی[4] را بر روی مجموعه داده [8] بررسی نموده‌اند. هدف از اینکار شناسایی مدلی مناسب است که بتواند با دقت بیشتری قصد خرید یک خریدار که از صفحات وب یک فروشگاه برخط بازدید می‌کند را پیش‌بینی کند. برای این منظور آنها ابتدا داده‌ها را پیش پردازش کردند، که شامل تبدیل ویژگی‌های اسمی به ویژگی‌های عددی بود و سپس عملکرد الگوریتم‌های مختلف را مورد بررسی و آزمایش قرار داده‌اند.

در [11] اوبیدات[5]، عملکرد پنج روش نمونه‌برداری بیش از حد مختلف را بر روی مجموعه داده [8] استفاده شده، مورد بررسی و آزمایش قرار داد. او در ابتدا داده‌ها را پیش پردازش کرد و سپس سه الگوریتم طبقه‌بندی پرسپترون چند لایه، درخت تصمیم و جنگل تصادفی را با یکدیگر مقایسه کرد، که طبقه‌بند جنگل تصادفی عملکرد بهتری در مقایسه با دو طبقه‌بند دیگر از خود نشان داد. به همین دلیل روش‌های نمونه‌برداری بیش از حد مختلف برای بهبود نتایج فقط بر روی این طبقه‌بند اعمال شدند، که روش متعادل سازی SVMSMOTE نسب به سایر روش‌های متعادل سازی عملکرد بهتری از خود نشان داد.

با تکیه به این حقیقت که رفتار خریداران برخط با افرادی که فقط به عنوان بازدیدکننده در وبگاه به گشت و گذار می‌پردازند، متفاوت است، این مقاله، با هدف افزایش دقت در کشف اولین لحظه حقیقت، رفتار خریداران برخط را برای پیش‌بینی اینکه آیا آنها یک محصول را در بازدید از وبگاه خریداری می‌کنند یا خیر، مورد مطالعه قرار داده است. به این ترتیب می‌توان کاربران با قصد خرید بالا را زودتر و بهتر شناسایی نموده و تبلیغات سفارشی متناسب‌تری را به آنها ارائه داد. مزیت اصلی این شناسایی، کاهش ریسک از دست دادن کاربران با احتمال خرید بالا و افزایش نرخ تبدیل می‌باشد. در این مقاله ما یک استراتژی کشف به موقع اولین لحظه حقیقت را بر پایه روش‌های داده-کاوی پیشنهاد داده‌ایم. تکنیک‌های داده‌کاوی و یادگیری ماشین در دیگر زمینه‌ها نظیر [12]- [14] استفاده گسترده‌ای دارند. با توجه به اینکه فرایند کشف و پیش‌بینی در حوزه داده کاوی و یادگیری ماشین، دارای چارچوب و الگوریتم‌های مشخص می-باشد، تمرکز این مقاله بر استفاده از روش‌های پیش‌پردازش، با هدف بهبود کیفیت داده‌ها، قبل از ارسال آنها به الگوریتم‌های طبقه‌بندی می‌باشد. استراتژی پیشنهادی این مقاله شامل تبدیل داده‌های اسمی به عددی، نرمال‌سازی داده‌های عددی، تشخیص و حذف داده‌های پرت، انتخاب ویژگی، متوازن سازی داده‌ها و سرانجام پیشبینی اولین لحظه حقیقت به کمک الگوریتم‌های طبقه‌بندی می‌باشد. در بین طبقه‌بندهای مورد استفاده، تمرکز این مقاله برروی استفاده از ایده‌های تلفیقی است که عموما نتایج بهتری نسبت به سایر روش‌ها ارائه می‌دهند. نتایج پیاده‌سازی ارزیابی استراتژی پیشنهادی به زبان پایتون، نشانگر بهبود دقت کشف خریداران برخط به نسبت کارهای انجام شده قبلی می‌باشد.

[1] Oversampling
[2] Baati
[3] Kabir
[4] Ensemble
[5] Obiedat

2. روش پیشنهادی

در این بخش روش پیشنهادی مورد بررسی قرار گرفته است. اما قبل از آن ابتدا به معرفی مجموعه داده مورد استفاده در این مقاله پرداخته‌ایم و سپس روش پیشنهادی بیان می‌گردد.

1.2. توصیف مجموعه داده

همانند کارهای مشابه بررسی شده در بخش مقدمه، در این پژوهش نیز از مجموعه داده Online Shoppers Intention [8] استفاده شده است. این مجموعه داده دارای 12330 رکورد می‌باشد که هر رکورد متعلق به یک کاربر متفاوت است. این مجموعه داده در یک دوره زمانی یک ساله جمع‌آوری شده است، تا از تاثیر روزهای خاص جلوگیری شود. دو دسته افراد در این مجموعه داده وجود دارند: کسانی که خرید کرده‌اند و کسانی که خریدی انجام نداده‌اند. از میان 12330 رکورد 84/5 درصد (تعداد 10442 رکورد) متعلق به کلاس منفی می‌باشند که خریدی را انجام نداده‌اند و 15/5 درصد باقیمانده (تعداد 1908 رکورد) متعلق به کلاس مثبت می‌باشند که با خرید خاتمه یافته‌اند. مجموعه داده از 10 ویژگی عددی و 8 ویژگی اسمی تشکیل شده است. توضیحات هر یک از این ویژگی‌های اسمی و عددی به ترتیب در جداول (1) و (2) ارائه شده است.

2.2. روش پیشنهادی

همان‌گونه که پیش از این ذکر شد می‌توان با بررسی تاریخچه رفتاری بازدیدکنندگان قصد خرید آن‌ها را پیش‌بینی کرد. برای بررسی تاریخچه رفتاری بازدید کنندگان می‌توان از دانش نوین داده‌کاوی استفاده کرد. داده‌کاوی از جمله دانش‌های در حال توسعه است که در سال‌های اخیر در تمامی عرصه‌ها جایگاه خود را تثبیت کرده است. در واقع به فرآیند استخراج دانش پنهان از داده‌ها و استفاده از آن اطلاعات برای ساخت مدل‌های پیش‌بینی کننده داده‌کاوی گفته می‌شود [15]. در این پژوهش با کمک روش‌های داده‌کاوی تاریخچه رفتاری بازدید کنندگان مورد بررسی قرار گرفته است تا بازدیدکنندگان با قصد خرید بالا شناسایی شوند. استراتژی که در این پژوهش برای پیش‌بینی قصد خرید بازیدکنندگان ارائه شده است شامل چندین گام اساسی می‌باشد که در شکل (1) نشان داده شده‌اند. در ادامه جزئیات هر یک از این گام‌ها بیان شده است.

3.2. پیش پردازش داده‌ها

کیفیت نتایج خروجی یک الگوریتم رابطه مستقیمی با کیفیت داده‌های ورودی آن دارد. هر چه داده‌های با کیفیت‌تری به الگوریتم‌ها داده شود، نتایج مطلوب‌تر و دقیق‌تری از آن‌ها استخراج می‌شود. ولی داده‌های امروزی به دلایل متعددی مستعد داده‌های نادرست و ناسازگار هستند که این داده‌های نادرست و ناسازگار می‌توانند در نتایج عملیاتی مثل داده‌کاوی خلل ایجاد کنند. در نتیجه داده‌ها باید قبل از هر گونه پردازشی، آماده‌سازی شوند.

جدول (1): ویژگی‌های اسمی استفاده شده در مجموعه داده ONLINE SHOPPERS INTENTION

نام ویژگی	توضیحات
Operating system	سیستم عامل بازدید کننده
Browser	مرورگر بازدید کننده
Region	منطقه جغرافیایی که بازدید کننده از آنجا نشست را شروع کرده است.
Traffic Type	نوع تبلیغاتی که بازدیدکننده را به سمت سایت کشانده است (از قبیل پیامک، آگهی تبلیغاتی)
Visitor Type	نوع بازدید کننده (کاربر جدید، کاربر قدیم، دیگران)
Weekend	مقدار دودویی که نشان می‌دهد تاریخ بازدید آخر هفته بوده است یا خیر.
Month	ماهی که بازدید در آن صورت گرفته است
Revenue	بر چسب کلاس که نشان می‌دهد که مشاهده سایت توسط بازدید کننده به تراکنش منجر شده است یا خیر.

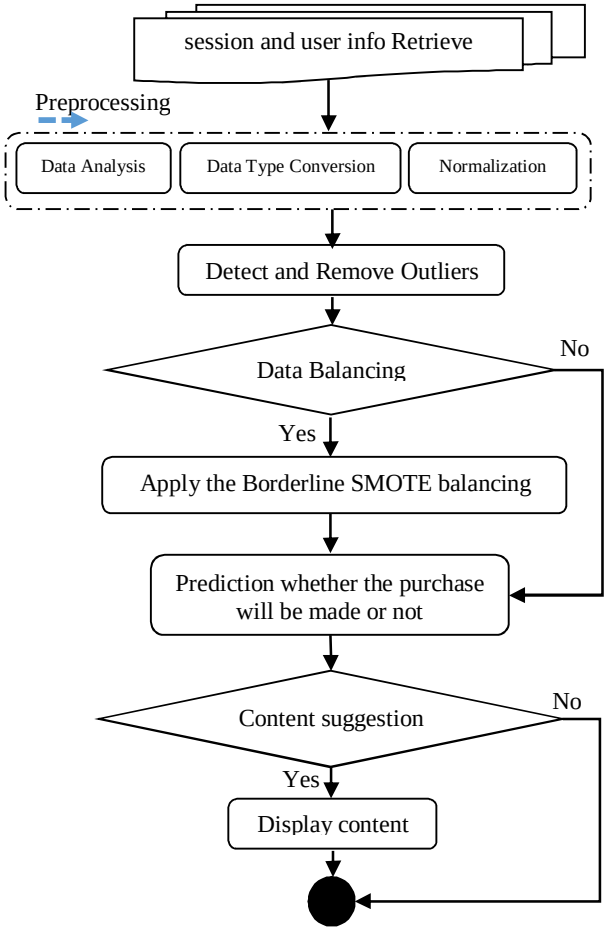
جدول (2): ویژگی‌های عددی استفاده شده در مجموعه داده Online Shoppers Intention

نام ویژگی	توضیحات	حداقل مقدار	حداکثر مقدار	مقدار میانگین	انحراف استاندارد
Administrative	تعداد صفحاتی که بازدید کننده در مورد مدیریت حساب بازدید کرده است.	0	27	2/315	3/322
Administrative duration	کل زمانی (بر حسب ثانیه) که توسط بازدیدکننده در صفحات مربوط به مدیریت حساب صرف شده است.	0	3398/75	80/819	176/779
Informational	تعداد صفحاتی که بازدیدکننده در مورد وب سایت، ارتباطات و اطلاعات آدرس سایت خرید، ملاقات کرده است.	0	24	0/504	1/27
Informational duration	کل زمانی (بر حسب ثانیه) که توسط بازدیدکننده در صفحات اطلاعاتی صرف شده است.	0	2549/375	34/472	140/749
Product related	تعداد صفحاتی که بازدیدکننده درباره صفحات مرتبط با محصول ملاقات کرده است.	0	705	31/731	44/476
Product related duration	کل زمانی (بر حسب ثانیه) که توسط بازدیدکننده در صفحات مرتبط با محصول صرف شده است.	0	63973/522	1194/746	1913/669
Bounce rate	مقدار میانگین نرخ پرش صفحاتی که توسط بازدیدکننده، بازدید شده است.	0	0/2	0/022	0/048
Exit rate	مقدار میانگین نرخ خروج از صفحاتی که توسط بازدیدکننده بازدید شده است.	0	0/2	0/043	0/049
Page value	میانگین مقدار صفحه از صفحاتی که توسط بازدیدکننده بازدید شده است.	0	361/764	5/889	18/568
Special day	نزدیک بودن زمان بازدید از سایت به یک روز خاص	0	1	0/061	0/199

آماده‌سازی و انتقال داده‌ها به یک شکل مناسب برای فرآیند استخراج دانش از داده‌ها پیش‌پردازش گفته می‌شود که یکی از مهم‌ترین وظایف داده‌کاوی است [16]. برای این منظور، در این مقاله گام‌های زیر به عنوان مراحل پیش‌پردازش داده‌ها در نظر گرفته شده‌اند:

- **آنالیز داده‌ها:** آنالیز داده‌ها برای به دست آوردن اطلاعات مفیدی در مورد مجموعه داده مورد ارزیابی انجام می‌شود. شناسایی مقادیر مفقود یا تهی موجود در مجموعه داده، شناسایی و حذف رکوردهای تکراری و شناسایی ویژگی‌های اسمی و عددی از جمله مواردی است که در این آنالیز، مورد توجه قرار می‌گیرد. انجام آنالیز ابتدایی داده‌ها، کمک زیادی به انجام مراحل بعدی پردازش می‌کند.
- **تبدیل ویژگی‌های اسمی به عددی:** در مقام مقایسه در کار با داده‌های عددی، متاسفانه عمده الگوریتم‌ها و روش‌های هوش مصنوعی، عملکرد مطلوبی در مواجه با داده‌های غیر عددی ندارند. به همین دلیل، نیازمند تبدیل ویژگی‌های اسمی به ویژگی‌های عددی هستیم. برای این منظور، در این مقاله از روش one hot encoding استفاده شده است. در این روش به ازای n مقادیر ممکن برای یک ویژگی اسمی، n ستون به مجموعه داده اضافه می‌شود که مقدار آنها به صورت دودویی است. اگر مقدار ویژگی اسمی x باشد، در ستون x ایجاد شده، مقدار یک ثبت می‌شود و $n-1$ ستون دیگر مقدار صفر می‌پذیرند.
- **نرمال‌سازی داده‌ها:** در هنگام پردازش داده‌ها واحد اندازه‌گیری استفاده شده برای ویژگی می‌تواند روی تحلیل داده‌ها اثرگذار باشد. برای مثال تغییر واحد اندازه‌گیری از متر به اینچ برای ویژگی قد ممکن است نتایج بسیار

متفاوتی را در فرآیند کاوش به همراه داشته باشد. به طور کلی بیان یک ویژگی در واحدهای کوچک‌تر باعث ایجاد بازه بزرگ‌تری برای آن می‌شود. در نتیجه این ویژگی دارای وزن بیشتری در تحلیل داده‌ها خواهد بود. برای اینکه واحدهای اندازه‌گیری متفاوت لطمه‌ای به تحلیل داده‌ها نزنند، داده‌ها باید نرمال‌سازی شوند.



شکل (1) : مراحل روش پیشنهادی

در نرمال‌سازی داده‌ها سعی می‌شود تا وزن یکسانی به کلیه‌ی ویژگی‌ها داده شود. انجام این کار باعث می‌شود تا مقادیر هر یک از ویژگی‌ها در محدوده یکسانی قرار گیرند، تا تاثیر ویژگی‌های با دامنه بزرگتر خنثی شود. برای این منظور از روش نرمال‌سازی Z-Score [17] استفاده می‌شود. روابط مورد استفاده در این روش نرمال‌سازی به شکل زیر می‌باشد:

$$x_i = \frac{v_i - \mu}{\sigma} \quad (1)$$

$$\mu = \frac{1}{n}\sum_{i=1}^{n} v_i \quad (2)$$

$$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(v_i - \mu)^2} \quad (3)$$

که در آن v_i مقدار i-ام ویژگی، μ میانگین و σ انحراف معیار مقادیر ویژگی می‌باشد.

پس از نرمال‌سازی داده‌ها به کمک روش Z-Score، اعداد مجموعه قدیمی به اعدادی با میانگین صفر و انحراف معیار یک تبدیل می‌گردند. در حالت کلی، انتظار می‌رود تا یک مجموعه داده در آزمون‌های آماری از یک الگوی مشخص پیروی کنند. مزیت بزرگ استفاده از روش Z-Score این است که به کمک آن داده‌های خارج از محدوده که از الگو پیروی نکنند، به عنوان داده پرت شناسایی می‌شوند.

4.2. تشخیص و حذف داده‌های پرت

در مجموعه داده‌ها ممکن است نمونه‌هایی وجود داشته باشد که با دیگر نمونه‌ها تفاوت چشمگیری دارند. این نمونه‌ها می‌توانند تحلیل‌ها و پیش‌بینی‌های انجام شده را تحت تاثیر قرار دهند و سبب افزایش خطا در نتایج آماری و کاهش دقت و کارایی الگوریتم‌های داده‌کاوی و مدل‌های پیش‌بینی شوند. بنابراین باید چنین داده‌های را قبل از هر گونه پردازشی شناسایی و حذف کرد.

به زیرمجموعه‌ای از داده‌ها که با دیگر داده‌ها در تناقض باشند، داده پرت می‌گویند. شناسایی داده‌های پرت در بسیاری از کاربردهای عملی مانند شناسایی سو استفاده از کارت اعتباری در داده‌های تراکنش‌های مالی، شناسایی خطاهای اندازه‌گیری در داده‌های علمی و یا آنالیز آماری داده‌های ورزشی مهم است [18]. در این پژوهش، به منظور عملکرد بهتر الگوریتم‌های طبقه‌بندی و جلوگیری از تاثیر داده‌های پرت بر روی نتایج خروجی، داده‌های پرت با استفاده از الگوریتم Local Outlier Factor [19] شناسایی و حذف می‌شوند. این الگوریتم بر مبنای مفهوم چگالی محلی بنا شده و در آن محلی بودن بر اساس k-

نزدیک‌ترین همسایه تعیین می‌شود که فاصله آنها برای تخمین چگالی مورد استفاده قرار می‌گیرد. با مقایسه چگالی محلی یک شی با چگالی‌های همسایه‌های آن می‌توان نواحی دارای چگالی مشابه و نقاطی که چگالی کمتری نسبت به همسایه‌های خود دارند را شناسایی کرد. این نقاط به عنوان داده پرت در نظر گرفته می‌شوند. ترکیب این روش با خروجی‌های روش Z-Score، توان کشف داده پرت در روش پیشنهادی را افزایش می‌دهد.

5.2. متوازن‌سازی داده‌ها

با توجه به نامتوازن بودن مجموعه داده مورد استفاده (10442 تراکنش بدون خرید در مقابل 1908 تراکنش منتج به خرید)، برای عملکرد بهتر الگوریتم‌های طبقه‌بندی، متوازن سازی داده‌ها یک امر ضروری است. الگوریتم‌های طبقه‌بندی کننده عموماً در مجموعه‌های داده متوازن عملکرد خوبی دارند. اما در واقعیت، داده‌های جمع‌آوری‌شده برای آموزش طبقه‌بندها به طور معمول نامتوازن هستند، یعنی تعداد نمونه‌های داده در کلاس‌های مختلف متفاوت است. این تفاوت باعث می‌شود که پیش‌بینی صحیح نمونه‌های کلاس اقلیت برای طبقه‌بندها دشوار شود و نمونه‌های کلاس اقلیت به اشتباه به عنوان نمونه‌های کلاس اکثریت طبقه‌بندی شوند [20]. پس از انجام عملیات متوازن‌سازی داده‌ها، تعداد نمونه‌هایی که در هر یک از کلاس‌های قرار گرفته‌اند، با یکدیگر برابر می‌باشد. در این پژوهش برای بهبود عملکرد الگوریتم‌های طبقه‌بندی از روش Borderline SMOTE [21] برای متوازن‌سازی داده‌ها استفاده شده است.

الگوریتم Borderline SMOTE نمونه توسعه یافته الگوریتم معروف SMOTE می‌باشد که بر خلاف آن به جای تولید نمونه‌های مصنوعی به صورت کورکورانه، نمونه‌های مصنوعی نزدیک به ناحیه‌های مرزی ایجاد می‌کند. نمونه‌هایی که در ناحیه‌های مرزی قرار می‌گیرند بیشتر از نمونه‌های دیگر به اشتباه طبقه‌بندی می‌شوند و در نتیجه برای طبقه‌بندی اهمیت بیشتری دارند.

6.2. پیش‌بینی نتیجه تراکنش

هدف نهایی یک مجموعه فروش برخط، کسب درآمد و سود از فروش محصولات است و بنابراین کشف به هنگام اولین لحظه حقیقت بازدیدکنندگان برای چنین مجموعه‌ای بسیار حیاتی است. با کشف به موقع کاربر متمایل به خرید و هدایت مناسب وی، به راحتی می‌توان نرخ تبدیل موثر وبگاه را افزایش داد. جهت پیش‌بینی قصد خرید بازیدکنندگان در این پژوهش از الگوریتم‌های یادگیری ماشین نظارت شده استفاده شده است. الگوریتم‌های یادگیری ماشین نظارت شده، به دلیل وجود امکان آموزش توام با نظارت، امکان پیش‌بینی با دقت مطلوب را فراهم می‌آورند. از جمله معروف‌ترین الگوریتم‌های یادگیری ماشین با نظارت که در کارهای مشابه گذشته بکار گرفته شده‌اند می‌توان به درخت تصمیم، جنگل تصادفی و شبکه عصبی پرسپترون چندلایه (MLP[1]) اشاره نمود.

در این پژوهش برای پیش‌بینی قصد خرید مشتریان برخط از پنج الگوریتم طبقه‌بندی کننده مختلف به نام‌های جنگل تصادفی، MLP، Bagging، Gradient Boosting و درخت تصمیم استفاده شده است. از آنجا که الگوریتم‌های تلفیقی در تحقیقات مختلف، به دلیل مکانیزم نظردهی ترکیبی، عملکرد بهتری در مقایسه با سایر الگوریتم‌های مشابه از خود نشان داده‌اند، در این تحقیق نیز از انواع مختلف این الگوریتم‌ها استفاده شده است. به استثنا الگوریتم‌های MLP و درخت تصمیم، سایر الگوریتم‌ها مورد استفاده از نوع تلفیقی می‌باشند.

3. ارزیابی نتایج

در این بخش، ابتدا به معرفی پارامترها و سناریوهای ارزیابی مورد استفاده در مقاله می‌پردازیم و سپس نتایج حاصل از ارزیابی راهکار پیشنهادی را از جنبه‌های مختلف مورد بحث و بررسی قرار داده‌ایم.

[1] Multi-Layer Perceptron

1.3. معیارها و سناریوی ارزیابی

برای ارزیابی نتایج در این پژوهش از روش 10-fold cross validation استفاده شده است. معیارهای که برای مقایسه الگوریتم‌های طبقه‌بندی در این پژوهش مورد استفاده قرار گرفته‌اند، به شرح ذیل می‌باشد:

$$Accuracy = \frac{TP + TN}{TP + TN + FN + FP} \quad (4)$$

$$Precision = \frac{TP}{TP + FP} \quad (5)$$

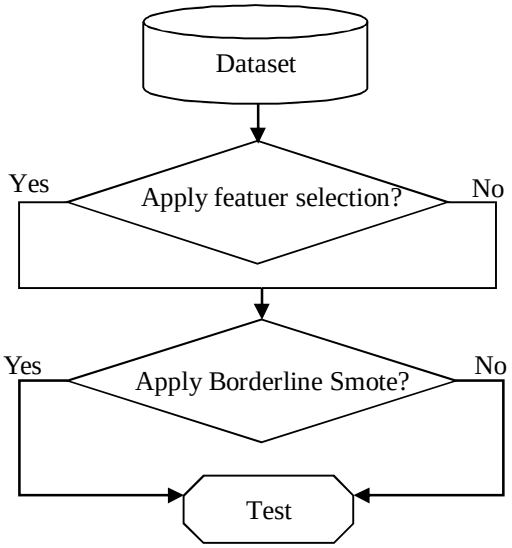
$$Recall = \frac{TP}{TP + FN} \quad (6)$$

$$F\_measure = \frac{2 * Precision * Recall}{Precision + Recall} \quad (7)$$

در روابط بالا، TP معرف تعداد نمونه‌های مثبتی است که به درستی عضو کلاس مثبت تشخیص داده شده‌اند، TN معرف تعداد نمونه‌های منفی است که به درستی عضو کلاس منفی تشخیص داده شده‌اند، FP معرف تعداد نمونه‌های منفی است که به اشتباه به عنوان عضو کلاس مثبت علامت زده شده‌اند و FN معرف تعداد نمونه‌های مثبتی است که به اشتباه به عنوان عضو کلاس منفی در نظر گرفته شده‌اند. پارامتر accuracy درصد صحت پیش‌بینی نتایج توسط طبقه‌بندها را نشان می‌دهد. اما از آنجا که مجموعه داده مورد استفاده نامتوازن است، نیازمند استفاده از پارامترهای Precision، Recall و F-Measure نیز هستیم تا دقت تشخیص طبقه‌بند در مواجه با کلاس‌های مختلف به خوبی ارزیابی شود.

پارامتر Precision یک معیار درستی است و نماینده درصدی از نمونه‌هایی است که به عنوان کلاس مثبت علامت گذاری شده‌اند و واقعا کلاس آنها مثبت است. پارامتر recall نیز یک معیار تمامیت است که نشان‌دهنده درصدی از نمونه‌های مثبت است که به درستی دسته‌بندی شده‌اند. پارامتر F-Measure نیز میانگین هارمونیک این دو پارامتر، در قالب یک پارامتر می‌باشد. همچنین منحنی هزینه-سود ROC[2] نیز به عنوان بخشی از فرآیند ارزیابی میزان دقت طبقه‌بندهای مختلف [17]، رسم شده است.

سناریو ارزیابی روش پیشنهادی در شکل (2) نشان داده شده است. همان‌طور که در این شکل مشخص است، ترکیبی از حالت‌های مختلف استفاده/عدم استفاده از روش متوازن‌سازی Borderline SMOTE، در کنار استفاده/عدم‌استفاده از روش‌های انتخاب ویژگی و ترکیب این حالت‌ها با یکدیگر در سناریو ارزیابی این مقاله قرار گرفته است.

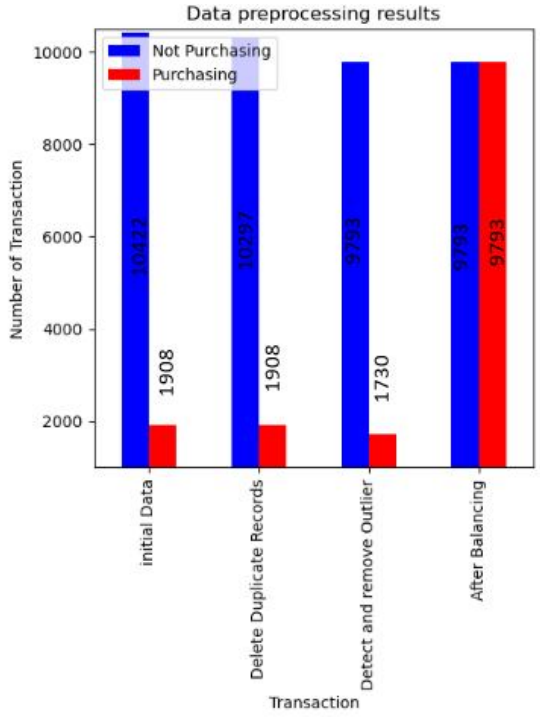


شکل (2) : سناریو ارزیابی روش پیشنهادی

2.3. ارزیابی نتایج حاصل از متوازن‌سازی و عدم متوازن‌سازی داده‌ها

شکل (3) الگوی توزیع داده‌ها در مجموعه داده مورد ارزیابی را در طی مراحل مختلف پیش پردازش نشان می‌دهد. همانطور که در این شکل دیده می‌شود، مجموعه داده مورد استفاده در این مقاله به شدت نامتوازن است و 84/4% رکوردها به کلاس منفی و بقیه به کلاس مثبت تعلق دارند. بنابراین فرایند متوازن‌سازی داده‌ها برای بهبود عملکرد الگوریتم‌های طبقه‌بندی ضروری است. نتایج اعمال الگوریتم‌های طبقه‌بندی بدون/با متوازن‌سازی در جداول (3) و (4) آورده شده است.

[2] Receiver operating characteristic



شکل (3): الگوی توزیع داده‌ها

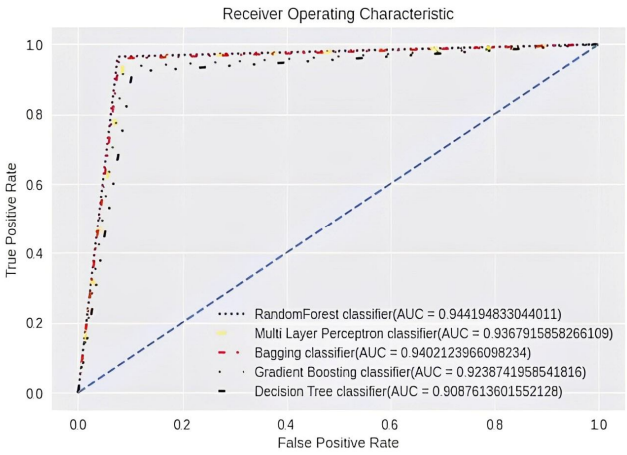
جدول (3): نتایج به دست آمده بدون متوازن‌سازی داده‌ها

Classifier Algorithm	بدون متعادل سازی داده‌ها			
	Accuracy	Precision	Recall	F_measure
Random Forest	0/9042	0/7474	0/5474	0/6313
MLP	0/8690	0/5662	0/5566	0/5591
Bagging	0/9020	0/7154	0/5780	0/6389
Gradient Boosting	0/9047	0/7211	0/5983	0/6532
Decision Tree	0/8649	0/5497	0/5630	0/5556

جدول (4): نتایج به دست آمده با متوازن‌سازی داده‌ها

Classifier Algorithm	بعد از متعادل سازی داده‌ها			
	Accuracy	Precision	Recall	F_measure
Random Forest	0/9442	0/9270	0/9644	0/9453
MLP	0/9368	0/9167	0/9612	0/9383
Bagging	0/9402	0/9240	0/9594	0/9413
Gradient Boosting	0/9239	0/9139	0/9360	0/9248
Decision Tree	0/9088	0/8986	0/9218	0/9100

همان گونه که از جدول (3) مشهود است، بدون متوازن‌سازی داده‌ها، اگرچه مقادیر به دست آمده برای پارامتر Accuracy قابل توجه است، اما مقادیر به دست آمده برای پارامترهای Precision، Recall و F_measure پایین هستند که نشان‌گر عملکرد نامطلوب الگوریتم‌های طبقه‌بندی در مواجه و تشخیص داده‌های کلاس اقلیت می‌باشد. ارزیابی نتایج طبقه‌بندی داده‌های متوازن شده در جدول (4) گویای اثر مثبت استفاده از الگوریتم متوازن‌سازی Borderline SMOTE و رفع این مشکل می‌باشد. شکل (4) نمودار ROC حاصل از اعمال الگوریتم‌های طبقه‌بندی بر روی داده‌های متوازن را نشان می‌دهد. نمودار ROC برای نمایش توانایی یک طبقه‌بند دودویی استفاده می‌شود. در نمودار یک خط مورب و قطری وجود دارد که مربوط به حدس تصادفی می‌باشد. هر چه منحنی ROC یک مدل به این خط قطری نزدیک‌تر باشد، مدل از صحت کمتری برخوردار است. جهت ارزیابی یک مدل مساحت زیر منحنی (AUC)[3] اندازه گیری می‌شود. هر چه مقدار AUC به مقدار 0/5 نزدیک‌تر باشد، مدل مزبور از صحت کمتری برخوردار است. یک مدل ایده‌آل و کامل از نظر صحت، دارای مقدار AUC برابر با 1 می‌باشد.



شکل (4): نمودار ROC حاصل از اعمال الگوریتم‌های طبقه‌بندی بر روی مجموعه داده متوازن

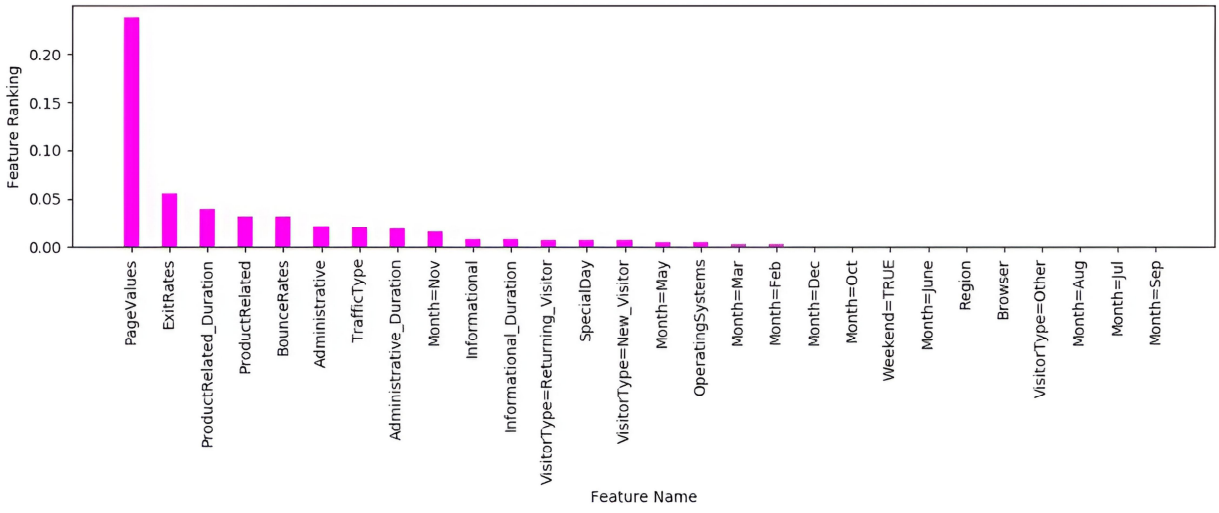
همان‌گونه که از شکل (4) و جدول (4) مشهود است، طبقه‌بند جنگل تصادفی بهتر از دیگر طبقه‌بندها عمل کرده است. شایان

[3] Area Under the Curve

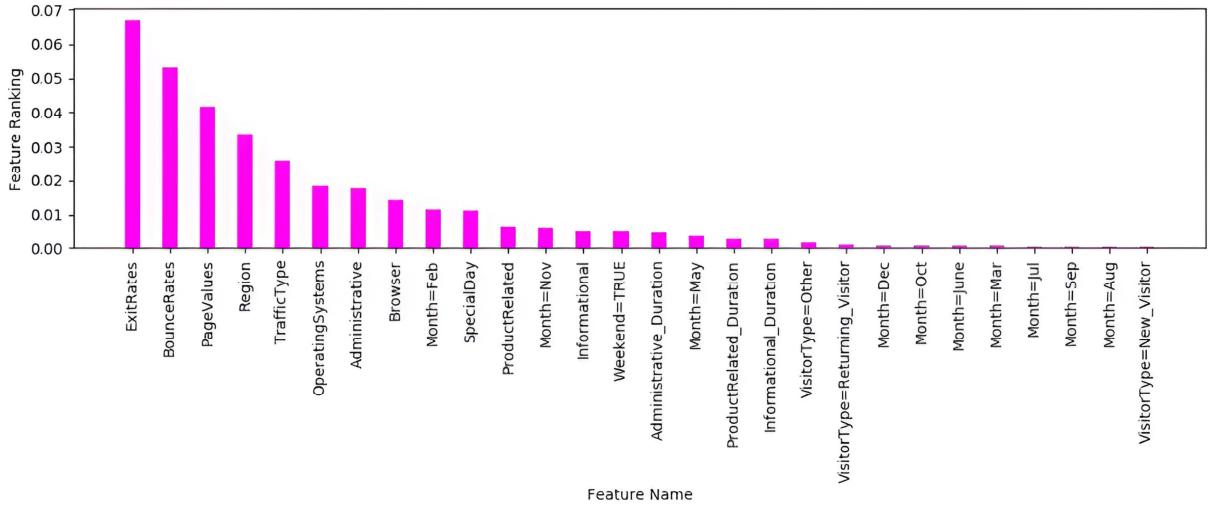
ذکر است که کلیه نتایج این مرحله، بدون اعمال الگوریتم‌های کاهش ویژگی می‌باشد. در بخش بعد، تاثیر الگوریتم‌های کاهش ویژگی نیز بررسی شده است.

3.3. ارزیابی نتایج حاصل از اعمال روش‌های کاهش ویژگی

به منظور کاهش تعداد ویژگی‌های استفاده شده در این پژوهش ما از 2 روش مشهور و متداول ReliefF [22] و Information Gain [22] استفاده می‌کنیم. این دو روش در بین روش‌های انتخاب ویژگی مبتنی بر رتبه هستند. شکل‌های (5- الف) و (5- ب)، ترتیب قرار گرفتن ویژگی‌ها، بر اساس رتبه آنها، با هر یک از روش‌های کاهش ویژگی مورد استفاده شده را نشان می‌دهد. براساس نتایج به دست آمده، ویژگی‌های منتخب پس از دهمین ویژگی با افت شدید رتبه و اهمیت مواجه هستند. بنابراین در این مقاله، 10 ویژگی برتر از هر روش کاهش ویژگی را انتخاب شده است. در ادامه، عملیات‌های پیش‌پردازش داده‌ها، تشخیص و حذف داده‌های پرت، متعادل‌سازی داده‌ها و اعمال الگوریتم‌های طبقه‌بندی بر روی ویژگی‌های انتخاب شده انجام شده است.



الف: روش Information Gain



ب: روش ReliefF

شکل (5): رتبه ویژگی‌های مختلف به کمک دو روش انتخاب ویژگی

نتایج به دست آمده با اعمال روش‌های کاهش ویژگی ذکر شده در جدول (5) نشان داده شده است. همان‌گونه که مشهود است، مقادیر به دست آمده از هر دو روش کاهش ویژگی به صورت تقریبی یکسان می‌باشد. به گونه‌ای که در ابتدای کار که تعداد

10 ویژگی برتر از هر روش انتخاب شده‌اند، فقط 5 ویژگی مشترک در بین آن‌ها وجود دارد ولی مقادیر به دست آمده از هر روش کاهش ویژگی برای پارامتر accuracy و یا دیگر پارامترها بسیار نزدیک به یکدیگر می‌باشند.

جهت مقایسه بهتر، روند ارزیابی با تعداد ویژگی منتخب بیشتر از 10، با گام 3، نیز تکرار شد که نتایج آنها نیز در جدول (5) آورده شده است. همانطور که دیده می‌شود، با اضافه کردن تعداد ویژگی‌های منتخب، تعداد ویژگی‌های مشترک بین هر دو روش در حال افزایش است و نتایج حاصل از هر دو روش کاهش ویژگی در حال نزدیک شدن به یکدیگر می‌باشند. همان‌طور که از نتایج جدول (5) مشهود است، در مجموعه داده مورد استفاده، با افزایش تعداد ویژگی‌های منتخب از 10 عدد (کمترین دقت) به 25 ویژگی منتخب و حتی تمامی ویژگی‌ها (بیشترین دقت)، در بهترین حالت دقت به اندازه 1/5 درصد بهبود یافته است. این درحالی است که کاهش ویژگی در این سطح، در مجموعه‌های داده بزرگ، بسیار حیاتی است و سبب کاهش شدید سربار محاسباتی پردازنده می‌شود. به علاوه همان گونه که مشهود است در بین الگوریتم‌های طبقه‌بندی مورد ارزیابی در این پژوهش طبقه‌بند جنگل تصادفی بهترین عملکرد را دارد و روش‌های Bagging و MLP، با اختلاف در جایگاه‌های دوم و سوم قرار دارند. نتایج حاصل از اعمال روش‌های کاهش ویژگی بدون متعادل سازی داده‌ها در جدول (6) ارائه شده است.

4.3. مقایسه نتایج کسب شده با کارهای مشابه

بیشترین دقت به دست آمده در این پژوهش برابر با 94/42 درصد می‌باشد، که با استفاده از طبقه‌بند جنگل تصادفی بر روی داده‌های متعادل به دست آمده است. همان‌طور که در شکل (6) نشان داده شده است، در مقام مقایسه با بهترین نتایج کارهای مشابه، دقت به دست آمده در این مقاله بیشتر از مقدار دقت به دست آمده از کارهای ذکر شده در بخش مقدمه می‌باشد. به گونه‌ای که کمترین مقدار دقت به دست آمده در این پژوهش (در حالت داده‌های متعادل) با بیشترین مقدار دقت به دست آمده در کارهای پیشینه پژوهش برابری می‌کند.

جدول (5): ارزیابی نتایج حاصل از کاهش ویژگی

نتایج به دست آمده با 10 ویژگی برتر

Classifier Algorithm	نتایج به دست آمده با روش information gain				نتایج به دست آمده با روش ReliefF			
	Accuracy	Precision	Recall	F_measure	Accuracy	Precision	Recall	F_measure
Random Forest	**0/9357**	**0/9129**	**0/9636**	**0/9375**	**0/9291**	0/9096	**0/9530**	**0/9307**
MLP	0/9268	0/9062	0/9530	0/9288	0/9176	0/8934	0/9490	0/9201
Bagging	0/9337	0/9119	0/9603	0/9354	0/9289	**0/9143**	0/9466	0/9301
Gradient Boosting	0/8957	0/8779	0/9196	0/8982	0/8951	0/8976	0/8921	0/8948
Decision Tree	0/8977	0/8893	0/9085	0/8988	0/8923	0/8850	0/9018	0/8933

نتایج به دست آمده با 13 ویژگی برتر

Classifier Algorithm	نتایج به دست آمده با روش information gain				نتایج به دست آمده با روش ReliefF			
	Accuracy	Precision	Recall	F_measure	Accuracy	Precision	Recall	F_measure
Random Forest	**0/9367**	0/9129	**0/9656**	**0/9385**	**0/9396**	**0/9203**	**0/9626**	**0/9410**
MLP	0/9319	0/9065	0/9631	0/9339	0/9291	0/9093	0/9535	0/9308
Bagging	0/9344	**0/9131**	0/9604	0/9361	0/9369	**0/9203**	0/9566	0/9381
Gradient Boosting	0/9015	0/8834	0/9253	0/9038	0/9123	0/8971	0/9316	0/9140
Decision Tree	0/9020	0/8928	0/9138	0/9032	0/9038	0/8955	0/9143	0/9048

ادامه جدول (5): ارزیابی نتایج حاصل از کاهش ویژگی

نتایج به دست آمده با 16 ویژگی برتر

Classifier Algorithm	نتایج به دست آمده با روش information gain				نتایج به دست آمده با روش ReliefF			
	Accuracy	Precision	Recall	F_measure	Accuracy	Precision	Recall	F_measure
Random Forest	**0/9405**	**0/9214**	**0/9632**	**0/9419**	**0/9418**	**0/9267**	0/9597	**0/9429**
MLP	0/9332	0/9176	0/9523	0/9344	0/9343	0/9115	**0/9622**	0/9361
Bagging	0/9383	0/9208	0/9592	0/9396	0/9390	0/9261	0/9543	0/9399
Gradient Boosting	0/9131	0/8976	0/9326	0/9147	0/9156	0/9083	0/9247	0/9164
Decision Tree	0/9052	0/8942	0/9193	0/9065	0/9083	0/9009	0/9176	0/9091

نتایج به دست آمده با 19 ویژگی برتر

Classifier Algorithm	نتایج به دست آمده با روش information gain				نتایج به دست آمده با روش ReliefF			
	Accuracy	Precision	Recall	F_measure	Accuracy	Precision	Recall	F_measure
Random Forest	**0/9418**	**0/9225**	**0/9648**	**0/9432**	**0/9429**	0/9275	**0/9612**	**0/9440**
MLP	0/9374	0/9176	0/9615	0/9389	0/9311	0/9125	0/9542	0/9326
Bagging	0/9365	0/9178	0/9591	0/9379	0/9412	**0/9279**	0/9570	0/9422
Gradient Boosting	0/9089	0/8959	0/9256	0/9104	0/9186	0/9114	0/9275	0/9193
Decision Tree	0/9050	0/8988	0/9129	0/9057	0/9064	0/9009	0/9134	0/9071

نتایج به دست آمده با 22 ویژگی برتر

Classifier Algorithm	نتایج به دست آمده با روش information gain				نتایج به دست آمده با روش ReliefF			
	Accuracy	Precision	Recall	F_measure	Accuracy	Precision	Recall	F_measure
Random Forest	**0/9422**	**0/9230**	**0/9650**	**0/9435**	**0/9430**	**0/9279**	**0/9607**	**0/9440**
MLP	0/9375	0/9186	0/9601	0/9388	0/9347	0/9154	0/9582	0/9362
Bagging	0/9383	0/9212	0/9586	0/9395	0/9413	0/9276	0/9572	0/9422
Gradient Boosting	0/9133	0/8989	0/9315	0/9149	0/9229	0/9105	0/9381	0/9241
Decision Tree	0/9065	0/8986	0/9164	0/9074	0/9066	0/8980	0/9176	0/9077

نتایج به دست آمده با 25 ویژگی برتر

Classifier Algorithm	نتایج به دست آمده با روش information gain				نتایج به دست آمده با روش ReliefF			
	Accuracy	Precision	Recall	F_measure	Accuracy	Precision	Recall	F_measure
Random Forest	**0/9427**	**0/9265**	0/9617	**0/9438**	**0/9435**	**0/9271**	0/9629	**0/9446**
MLP	0/9367	0/9142	**0/9640**	0/9384	0/9339	0/9092	**0/9646**	0/9360
Bagging	0/9411	0/9253	0/9596	0/9422	0/9411	0/9254	0/9598	0/9422
Gradient Boosting	0/9224	0/9113	0/9360	0/9235	0/9244	0/9136	0/9377	0/9255
Decision Tree	0/9091	0/9012	0/9189	0/9099	0/9068	0/8986	0/9171	0/9077

ادامه جدول (5): ارزیابی نتایج حاصل از کاهش ویژگی

نتایج به دست آمده با تمام ویژگی‌ها

Classifier Algorithm	نتایج به دست آمده با روش information gain				نتایج به دست آمده با روش ReliefF			
	Accuracy	Precision	Recall	F_measure	Accuracy	Precision	Recall	F_measure
Random Forest	**0/9442**	**0/9270**	**0/9644**	**0/9453**	**0/9442**	**0/9270**	**0/9644**	**0/9453**
MLP	0/9368	0/9167	0/9612	0/9383	0/9368	0/9167	0/9612	0/9383
Bagging	0/9402	0/9240	0/9594	0/9413	0/9402	0/9240	0/9594	0/9413
Gradient Boosting	0/9239	0/9139	0/9360	0/9248	0/9239	0/9139	0/9360	0/9248
Decision Tree	0/9088	0/8986	0/9218	0/9100	0/9088	0/8986	0/9218	0/9100

جدول (6): ارزیابی نتایج حاصل از کاهش ویژگی

نتایج به دست آمده با 10 ویژگی برتر

Classifier Algorithm	نتایج به دست آمده با روش information gain				نتایج به دست آمده با روش ReliefF			
	Accuracy	Precision	Recall	F_measure	Accuracy	Precision	Recall	F_measure
Random Forest	0/8994	0/7331	0/5808	0/6470	0/8935	0/6966	0/5373	0/6061
MLP	0/8775	0/6343	0/5628	0/5940	0/8691	0/5849	0/4986	0/5370
Bagging	0/8963	0/7185	0/5791	0/6399	0/8912	0/6819	0/5384	0/6010
Gradient Boosting	0/8990	0/7208	0/5987	0/6532	0/8964	0/7035	0/5563	0/6208
Decision Tree	0/8515	0/5334	0/5590	0/5455	0/8481	0/5024	0/5171	0/5087

نتایج به دست آمده با 13 ویژگی برتر

Classifier Algorithm	نتایج به دست آمده با روش information gain				نتایج به دست آمده با روش ReliefF			
	Accuracy	Precision	Recall	F_measure	Accuracy	Precision	Recall	F_measure
Random Forest	0/9000	0/7293	0/5656	0/6364	0/9028	0/7359	0/5696	0/6416
MLP	0/8691	0/5836	0/5435	0/5619	0/8692	0/5857	0/5298	0/5548
Bagging	0/8984	0/7107	0/5838	0/6404	0/8981	0/7032	0/5796	0/6351
Gradient Boosting	0/8988	0/7104	0/5901	0/6438	0/9028	0/7188	0/6000	0/6536
Decision Tree	0/8610	0/5515	0/5656	0/5581	0/8553	0/5286	0/5414	0/5342

نتایج به دست آمده با 16 ویژگی برتر

Classifier Algorithm	نتایج به دست آمده با روش information gain				نتایج به دست آمده با روش ReliefF			
	Accuracy	Precision	Recall	F_measure	Accuracy	Precision	Recall	F_measure
Random Forest	0/9019	0/7353	0/5605	0/6355	0/9021	0/7340	0/5661	0/6387
MLP	0/8761	0/6049	0/5512	0/5753	0/8651	0/5605	0/5517	0/5549
Bagging	0/9006	0/7228	0/5674	0/6352	0/8980	0/7043	0/5749	0/6326
Gradient Boosting	0/9031	0/7194	0/6006	0/6540	0/9018	0/7138	0/5976	0/6502
Decision Tree	0/8587	0/5371	0/5484	0/5422	0/8600	0/5413	0/5561	0/5483

ادامه جدول (6): ارزیابی نتایج حاصل از کاهش ویژگی

نتایج به دست آمده با 19 ویژگی برتر

Classifier Algorithm	نتایج به دست آمده با روش information gain				نتایج به دست آمده با روش ReliefF			
	Accuracy	Precision	Recall	F_measure	Accuracy	Precision	Recall	F_measure
Random Forest	0/9027	0/7411	0/5614	0/6371	0/9036	0/7416	0/5688	0/6428
MLP	0/8766	0/6084	0/5505	0/5746	0/8709	0/5870	0/5399	0/5607
Bagging	0/8992	0/7132	0/5746	0/6350	0/9005	0/7202	0/5754	0/6384
Gradient Boosting	0/9024	0/7224	0/5918	0/6488	0/9015	0/7142	0/5974	0/6502
Decision Tree	0/8608	0/5446	0/5585	0/5509	0/8599	0/5405	0/5682	0/5534

نتایج به دست آمده با 22 ویژگی برتر

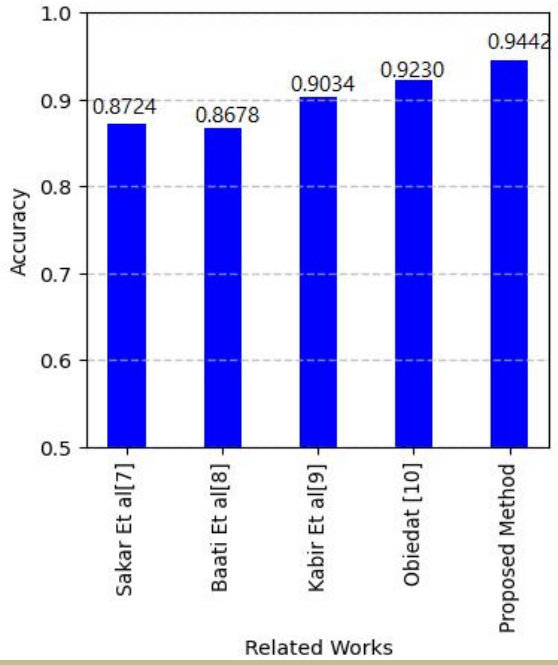
Classifier Algorithm	نتایج به دست آمده با روش information gain				نتایج به دست آمده با روش ReliefF			
	Accuracy	Precision	Recall	F_measure	Accuracy	Precision	Recall	F_measure
Random Forest	0/9030	0/7455	0/5447	0/6292	0/9026	0/7432	0/5596	0/6373
MLP	0/8714	0/5858	0/5261	0/5528	0/8716	0/5881	0/5590	0/5717
Bagging	0/9017	0/7204	0/5722	0/6373	0/9021	0/7248	0/5857	0/6471
Gradient Boosting	0/9040	0/7243	0/5904	0/6498	0/9031	0/7241	0/5974	0/6534
Decision Tree	0/8638	0/5488	0/5576	0/5523	0/8582	0/5347	0/5601	0/5468

نتایج به دست آمده با 25 ویژگی برتر

Classifier Algorithm	نتایج به دست آمده با روش information gain				نتایج به دست آمده با روش ReliefF			
	Accuracy	Precision	Recall	F_measure	Accuracy	Precision	Recall	F_measure
Random Forest	0/9040	0/7485	0/5566	0/6377	0/9035	0/7398	0/5641	0/6392
MLP	0/8735	0/5941	0/5414	0/5656	0/8722	0/5841	0/5624	0/5714
Bagging	0/9019	0/7216	0/5797	0/6424	0/9013	0/7153	0/5850	0/6426
Gradient Boosting	0/9035	0/7218	0/5960	0/6524	0/9041	0/7260	0/5934	0/6520
Decision Tree	0/8614	0/5433	0/5656	0/5539	0/8590	0/5347	0/5602	0/5467

نتایج به دست آمده با تمام ویژگی‌ها

Classifier Algorithm	نتایج به دست آمده با روش information gain				نتایج به دست آمده با روش ReliefF			
	Accuracy	Precision	Recall	F_measure	Accuracy	Precision	Recall	F_measure
Random Forest	0/9042	0/7474	0/5474	0/6313	0/9042	0/7474	0/5474	0/6313
MLP	0/8690	0/5662	0/5566	0/5591	0/8690	0/5662	0/5566	0/5591
Bagging	0/9020	0/7154	0/5780	0/6389	0/9020	0/7154	0/5780	0/6389
Gradient Boosting	0/9047	0/7211	0/5983	0/6532	0/9047	0/7211	0/5983	0/6532
Decision Tree	0/8649	0/5497	0/5630	0/5556	0/8649	0/5497	0/5630	0/5556



شکل (6) : مقایسه بهترین دقت به دست آمده به کمک روش پیشنهادی و سایر کارهای مرتبط

4. نتیجه‌گیری

این مقاله رفتار خریداران برخط را برای پیش‌بینی اینکه آیا آنها یک محصول را در بازدید از وبگاه خریداری می‌کنند یا خیر، مورد مطالعه قرار می‌دهد. افزایش دقت این پیش‌بینی سبب می‌شود تا کاربران با قصد خرید بالا زودتر و بهتر شناسایی شده و با ارائه محتوای سفارشی مناسب‌تر به آن‌ها، احتمال خرید را افزایش دهد. برای این منظور از ترکیبی از روش‌های مختلف پیش‌پردازش داده‌ها، شامل تبدیل داده‌های اسمی به عددی، نرمال‌سازی، تشخیص و حذف داده‌های پرت، انتخاب ویژگی و متوازن‌سازی، به منظور بهبود کیفیت داده‌های ورودی به الگوریتم‌های پیش‌بینی‌کننده استفاده شده است. همچنین با هدف کسب نتایج بهتر، استفاده از الگوریتم‌های طبقه‌بندی تلفیقی نیز به عنوان راه‌کارهای پیش‌بینی کننده مورد توجه قرار گرفته است. نتایج ارزیابی‌ها نشان می‌دهد که الگوریتم‌های طبقه‌بندی بر روی داده‌های متوازن عملکرد بهتری دارند، به گونه‌ای که بیشترین دقت به دست آمده در این پژوهش برابر با 94/42 درصد می‌باشد که با استفاده از طبقه‌بند جنگل تصادفی بر روی داده‌های متعادل به دست آمده است. نتیجه کسب شده به نسبت سایر کارهای مشابه انجام شده، از دقت بالاتری برخوردار است.

تعارض منافع: نویسندگان اعلام می‌کنند که هیچ تعارض منافعی ندارند.

مراجع

[1] I. Kurniawan, M.F. Akbar, D.F. Saepudin, M.S. Azis, and M. Tabrani, “Improving the effectiveness of classification using the data level approach and feature selection techniques in online shoppers purchasing intention prediction,” J. Physics Conf. Series, vol. 1641, no. 012083, pp. 1-8, 2020, doi: 10.1088/1742-6596/1641/1/012083.

[2] I.O. Adam, M.D. Alhassan, and Y. Afriyie, “What drives global B2C Ecommerce? An analysis of the effect of ICT access, human resource development and regulatory environment,” Technol. Anal. Strateg. Manag., vol. 32, no. 7, pp. 835-850, 2020, doi: 10.1080/09537325.2020.1714579.

[3] D. Blanchard, Supply chain management best practices, Third Edition, Wiley, 2021.

[4] J. Wolny, and N. Charoensuksai, “Mapping customer journeys in multichannel decision-making,” J. Direct Data Digit. Mark. Pract., vol. 15, no. 4, pp. 317-326, 2014, doi: 10.1057/dddmp.2014.24.

[5] N. Gudigantala, P. Bicen, and M. Eom, “An examination of antecedents of conversion rates of e-commerce retailers,” Manag. Res. Rev., vol. 39, no. 1, pp. 82-114, 2016, doi: 10.1108/MRR-05-2014-0112.

[6] G. Suchacka, M. Skolimowska-Kulig, and A. Potempa, “A k-nearest neighbors method for classifying user sessions in e-commerce scenario,” J. Telecommun. Inf. Technol., pp. 64-69, 2015.

[7] C.O. Sakar, S.O. Polat, M. Katircioglu, and Y. Kastro, “Real-time prediction of online shoppers’ purchasing intention using multilayer perceptron and LSTM

recurrent neural networks," Neural Comput. Appl., vol. 31, pp. 6893-6908, 2019, doi: 10.1007/s00521-018-3523-0.

[8] UCI Machine Learning Repository, Accessed August 2023, Available: https://archive.ics.uci.edu/ml/datasets/Online+Shoppers+Purchasing+Intention+Dataset

[9] K. Baati, and M. Mohsil, "Real-time prediction of online shoppers' purchasing intention using random forest," in IFIP Int. Conf. Artif. Intell. Appl. Innov. 16th IFIP WG 12.5 Int. Conf. (AIAI), Neos Marmaras, Greece, 2020, Part I 16, pp. 43-51, doi: 10.1007/978-3-030-49161-1_4.

[10] M.R. Kabir, F.B. Ashraf, and R. Ajwad, "Analysis of different predicting model for online shoppers' purchase intention from empirical data," in 22nd Int. Conf. Comput. Inf. Technol. (ICCIT), 2019, pp. 1-6, doi: 10.1109/ICCIT48885.2019.9038521.

[11] R. Obiedat, "A comparative study of different data mining algorithms with different oversampling techniques in predicting online shopper behavior," Int. J. Adv. Trends Comput. Sci. Eng., vol. 9, no. 3, pp. 3575-3583, 2020, doi: 10.30534/ijatcse/2020/164932020.

[12] Z. Sharifi Mehrjard, H. Momeni, and H. Adabi Ardekani, "A review of machine learning algorithms to diagnose autism using EEG signal," Soft Comput. J., vol. 13, no. 1, pp. 2-19, 2024, doi: 10.22052/SCJ.2023.248522.1110 [In Persian].

[13] M. Mousavi, S. Hosseini, and M.R. Omidi, "Improved Deep Neural Network Algorithm for Covid-19 Detection in Internet of Things," Soft Comput. J., vol. 12, no. 2, pp. 54-71, 2024, doi: 10.22052/SCJ.2023.248686.1117 [In Persian].

[14] E. Saberi, E. Radmand, J. Pirgazi, and A. Kermani, "Buying and selling strategy in the Iranian stock market using machine learning models along with feature selection using the Cuckoo Search algorithm," Soft Comput. J., vol. 12, no. 2, pp: 130-145, 2024, doi: 10.22052/SCJ.2023.252793.1144 [In Persian].

[15] E.H.A. Rady and A.S. Anwar, "Prediction of kidney disease stages using data mining algorithms," Inf. Med. Unlocked, vol. 15, pp: 1-7, 2019, doi: 10.1016/j.imu.2019.100178.

[16] S.A. Alasadi and W.S. Bhaya, "Review of data preprocessing techniques in data mining," J. Eng. Appl. Sci., vol. 12, no. 16, pp. 4102-4107, 2017, doi: 10.36478/jeasci.2017.4102.4107.

[17] J. Han, M. Kamber, and J. Pei, Data mining: concepts and techniques, Third Edition, Morgan Kaufmann, 2012, doi: 10.1016/C2009-0-61819-5.

[18] A. Zimek and P. Filzmoser, "There and back again: Outlier detection between statistical reasoning and data mining algorithms," Wiley Interdisciplinary Reviews: Data Mining Knowl. Discov., vol. 8, no. 6, pp. 1-37 ,2018, doi: 10.1002/widm.1280.

[19] M.M. Breunig, H.P. Kriegel, R.T. Ng, and J. Sander, "LOF: identifying density-based local outliers," in Proc. 2000 ACM SIGMOD Int. Conf. Manag. Data, 2000, pp. 93-104, doi: 10.1145/342009.335388.

[20] C.F. Tsai, W.C. Lin, Y.H. Hu, and G.T. Yao, "Under-sampling class imbalanced datasets by combining clustering analysis and instance selection," Inf. Sci., vol. 477, pp. 47-54, 2019, doi: 10.1016/j.ins.2018.10.029.

[21] H. Han, W.Y. Wang, and B.H. Mao, "Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning," in Int. Conf. Intell. Comput., 2005, pp. 878-887, doi: 10.1007/11538059_91.

[22] P. Yildirim, "Filter based feature selection methods for prediction of risks in hepatitis disease," Int. J. Mach. Learn. Comput., vol. 5, no. 4, pp. 258-263, 2015, doi: 10.7763/IJMLC.2015.V5.517.