# تشخیص اولین لحظه حقیقت در خرید آنلاین با استفاده از روش‌های پیش‌پردازش داده‌ها و طبقه‌بندهای تلفیقی

محسن امیرافضلی [۱]، دانشجوی دکتری، حسین غفاریان [۲*]، استادیار

[۱] دانشکده فنی و مهندسی ـ دانشگاه اراک ـ اراک ـ ایران ـ m.a.afzali1991@gmail.com

[۲] دانشکده فنی و مهندسی ـ دانشگاه اراک ـ اراک ـ ایران ـ h-ghaffarian@araku.ac.ir

**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 آنلاین، نموده است. این روزها مردم تمایل به کاوش آنلاین برای پیدا کردن اقلام مورد نیاز خود و خرید از طریق تراکنش‌های آنلاین دارند. با این حال، تعداد خریداران واقعی هنوز در مقایسه با تعداد کل بازدیدکنندگان از این سایت‌ها بسیار کم است. تحلیل رفتاری، پیش‌بینی و شناسایی زودهنگام بازدیدکنندگانی که قصد خرید از فروشگاه آنلاین را دارند، زمینه ارائه محتوای سفارشی مناسب‌تر برای آن‌ها را فراهم می‌آورد. از دیدگاه مدیریتی به این زمان اصطلاحاً اولین لحظه حقیقت گفته می‌شود. مزیت اصلی این پیشبینی کاهش ریسک 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 داده-های 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 به کار گرفته شده است. سپس داده‌های اصلاح شده به مجموعه‌ای 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 ۹۴/۴۲ ٪ رسیده است که در مقام مقایسه با بهترین نتایج کارهای پیشین، دقت تشخیص افزایش داشته است.**

**واژه‌های کلیدی: خرید آنلاین، اولین لحظه حقیقت، داده‌کاوی، پیش‌پردازش، طبقه‌بند تلفیقی**

* **حسین غفاریان**، h-ghaffarian@araku.ac.ir

# *Detecting the First Moment Of Truth in Online Shopping Using Data Preprocessing Methods and Ensemble Classifiers*

Mohsen Amir Afzali [1], PHD Student, Hossein Ghaffarian [2*], Assistant Professor

[1] Department of Computer Engineering, Faculty of Engineering, Arak University, Arak, Iran, m.a.afzali1991@gmail.com

[2*] Department of Computer Engineering, Faculty of Engineering, Arak University, Arak, Iran, h-ghaffarian@araku.ac.ir

***Abstract***: In this article, we present a strategy with the aim of increasing the accuracy of early detection of buyers from customers browsing in an online store. Nowadays, people tend to explore online to find the items they need and buy through online transactions. However, the number of actual buyers is still very low compared to the total number of visitors to these sites. Behavioral analysis, prediction and early identification of visitors who intend to buy from the online store provides the basis for providing more suitable customized content for them. From a managerial point of view, this time is called the First Moment of Truth (FMoT). The main advantage of this precedent is reducing the risk of losing users with high purchase probability and increasing the conversion rate. Due to the consistency of the prediction and diagnosis framework in data mining, the focus of this article is on the optimal use of pre-processing methods, with the aim of improving the quality of input data to classification algorithms. For this reason, in the proposed strategy, we use a set of algorithms for converting nominal content into numerical, normalization, outlier data detection, feature selection and balancing. Then, we give the modified data to a set of different classification algorithms, including C4.5 decision tree and multi-layer perceptron, and combined classification algorithms of random forest, bagging and gradient boosting. The evaluation of the results shows that the highest amount of accuracy obtained in this research by using ensemble classifiers has reached 94.42%, which compared with the best results of previous works, the accuracy of diagnosis has increased.

**Keywords:** Online Shopping ،First Moment of Truth ،Data Mining, Pre-processing, Ensemble Classifier

*** Hossein Ghaffarian, h-ghaffarian@araku.ac.ir**

## ۱. 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 آنلاین می‌باشد. این اقدام یا نگرش به دلیل ویژگی‌هایی از قبیل راحتی، دسترسی آسان، عدم وجود جمعیت و صرفه‌جویی در زمان که سایت‌های تجارت الکترونیک در اختیار کاربران قرار می‌دهند، زندگی انسان‌ها را راحت‌تر و آسان‌تر کرده است. به گونه‌ای که در سال‌های ۲۰۱۵ تا ۲۰۱۷ درصد مشتریانی که هنوز خرید از فروشگاه‌های فیزیکی را ترجیح می‌دهند از ۸۵ درصد به ۷۰ درصد کاهش یافته است [۱]. در سال ۲۰۱۳ حدود یک میلیارد نفر در سراسر جهان کالاهای خود را به صورت آنلاین خریداری می‌کردند. این عدد در سال ۲۰۱۸ به ۱/۸ میلیارد نفر افزایش یافت که بیش از یک پنجم جمعیت جهان است. علاوه بر این، در سال ۲۰۱۷ فروش تجارت الکترونیکی در سراسر جهان بالغ بر ۲/۳ تریلیون دلار بوده است [۲].

پس از اینکه جان کارلزون[1] در اوایل دهه ۱۹۸۰ عبارت لحظه حقیقت (MOT[2]) 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 [۳]. اولین زمان اخذ تصمیم مشتری برای خرید به اولین لحظه حقیقت (FMoT[3]) معروف است. با رشد اینترنت کم کم مفاهیم جدید دیگری نظیر تصمیم‌گیری چندکاناله[4] در مباحث فروش‌های آنلاین و آفلاین مورد توجه قرار گرفت. جیم لسینسکی[5] از گوگل مفهوم لحظه صفر حقیقت (ZMOT[6]) را در ابتدای قرن جدید مطرح نمود [۴]. 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 آنلاین باید متوجه بشویم که یک کاربر آیا به واقع با قصد و نیت خرید مراجعه کرده است و یا صرفاً 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 آنلاین، 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 وب سایت است که به خرید منجر شده‌اند [۵]. هر چه نرخ تبدیل رشد داشته باشد به معنای موفقیت سایت می‌باشد. با این حال، تعداد خریداران واقعی هنوز در مقایسه با تعداد کل بازدیدکنندگان از این سایت‌ها بسیار کم است. از آنجا که در خرید آنلاین هیچ تعامل فردی بین خریدار و فروشنده صورت نمی‌گیرد، با بررسی تاریخچه رفتار کاربران می‌توان رفتار بعدی آن‌ها را پیش‌بینی کرد. در یک وب سایت، دسترسی به داده‌های گذشته کاربران در هر جلسه که شامل، صفحاتی که

[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

دسترسی داشته‌اند، محصولاتی که مشاهده کرده و/یـا خریـداری کرده‌اند، کلیک‌هایی که انجام داده‌اند، زمانی که صرف کرده‌اند و بسیاری موارد دیگر آسان است. ویژگی‌های جلسـه و اطلاعـات رفتاری کاربر به عنوان دنباله‌ای از درخواست‌های HTTP ارسال شده توسط مشتری در فایل‌های log سرور وب ذخیره می‌شوند [۶]. بنابراین می‌توان با تجزیه و تحلیل داده‌های قبلی خریداران، قصد خرید آن‌ها را پیش‌بینی کرد، که به یک حـوزه نوظهـور از تحقیقات در حوزه محاسبه و داده‌کاوی تبدیل شده است.

در زمینه تشخیص رفتار خریداران آنلاین کارهای متفاوتی انجام شده است. در [۷] ساکار[8] و همکارانش با استفاده از مجموعه داده Online Shoppers Intention [۸]، یک سیستم تجزیه و تحلیل رفتار خریداران آنلاین در زمان بلادرنگ طراحی کردند. سیستم پیشنهادی متشکل از دو ماژول است که قصد خرید بازدید کننده و احتمال ترک سایت را پیش‌بینی می‌کند. در ماژول اول، قصد خرید بازدید کننده با استفاده از اطلاعات صفحه نمایش جمع آوری شده در طول بازدید همراه با برخی از اطلاعات جلسه و کاربر پیش‌بینی می‌شود. در ماژول دوم، نویسندگان از یک شبکه عصبی مکرر مبتنی بر حافظه طولانی کوتاه مدت بر اساس اطلاعات جریان کلیک متوالی استفاده کردند تا احتمال ترک سایت توسط بازدید کننده، بدون انجام معامله را پیش‌بینی کنند. همچنین در این مقاله از روش‌های نمونه‌برداری بیش از حد[9] و کاهش ویژگی برای بهبود عملکرد و و مقایس‌پذیری مجموعه‌ای از الگوریتم‌های یادگیری ماشین نظارت شده، استفاده شده است.

در [۹] باتی[10] و همکارش یک سیستم پیش‌بینی رفتار خریداران آنلاین پیشنهاد کردند تا به محض بازدید از وب سایت، 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 [۸] 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 [۱۰] کبیر[11] و همکارش عملکرد الگوریتم-های نظارت شده و روش‌های تلفیقی[12] را بر روی مجموعه داده [۸] بررسی نموده‌اند. هدف از اینکار شناسایی مدلی مناسب است که بتواند با دقت بیشتری قصد خرید یک خریدار که از صفحات وب یک فروشگاه آنلاین 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 [۱۱] اوبیدات[13]، عملکرد پنج روش نمونه‌برداری بیش از حد مختلف را بر روی مجموعه داده [۸] 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 آنلاین بـا افـرادی که فقط به عنوان بازدیدکننده در سایت به گشـت و گـذار مـی-

8 Sakar
9 oversampling
10 Baati

11 Kabir
12 Ensemble
13 Obiedat

پردازند، متفاوت است، این مقاله، با هدف افزایش دقت در کشف اولین لحظه حقیقت، رفتار خریداران آنلاین را برای پیش‌بینی اینکه آیا آن‌ها یک محصول را در بازدید از وب سایت 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 داده‌کاوی پیشنهاد داده‌ایم. تکنیک‌های داده-کاوی و یادگیری ماشین در دیگر زمینه‌ها نظیر [۱۴-۱۳-۱۲] استفاده گسترده‌ای دارند. با توجه به اینکه فرایند کشف و پیش-بینی در حوزه داده کاوی و یادگیری ماشین، دارای چارچوب و الگوریتم‌های مشخص می‌باشد، 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 عموماً نتایج بهتری نسبت به سایر روش‌ها ارائه می‌دهند. نتایج پیاده‌سازی ارزیابی استراتژی پیشنهادی به زبان پایتون، نشانگر بهبود دقت کشف خریداران آنلاین به نسبت کارهای انجام شده قبلی می‌باشد.

## ۲. 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.

### ۱،۲. توصیف مجموعه داده

همانند کارهای مشابه بررسی شده در بخش مقدمه، در این پژوهش نیز از مجموعه داده Online Shoppers Intention [۸] استفاده شده است. این مجموعه داده دارای ۱۲۳۳۰ رکورد می‌باشد که هر رکورد متعلق به یک کاربر متفاوت است. این مجموعه داده در یک دوره زمانی یک ساله جمع آوری 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 ۱۲۳۳۰ رکورد ۸۴/۵ درصد (تعداد ۱۰۴۴۲ رکورد) متعلق به کلاس منفی می‌باشند که خریدی را انجام نداده‌اند و ۱۵/۵ درصد باقیمانده ( تعداد ۱۹۰۸ رکورد) متعلق به کلاس مثبت می‌باشند که با خرید خاتمه یافته-اند. مجموعه داده از ۱۰ ویژگی عددی و ۸ ویژگی اسمی تشکیل شده است. توضیحات هر یک از این ویژگی‌های اسمی و عددی به ترتیب در جداول ۱ و ۲ ارائه شده است.

**جدول ۱: 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. |

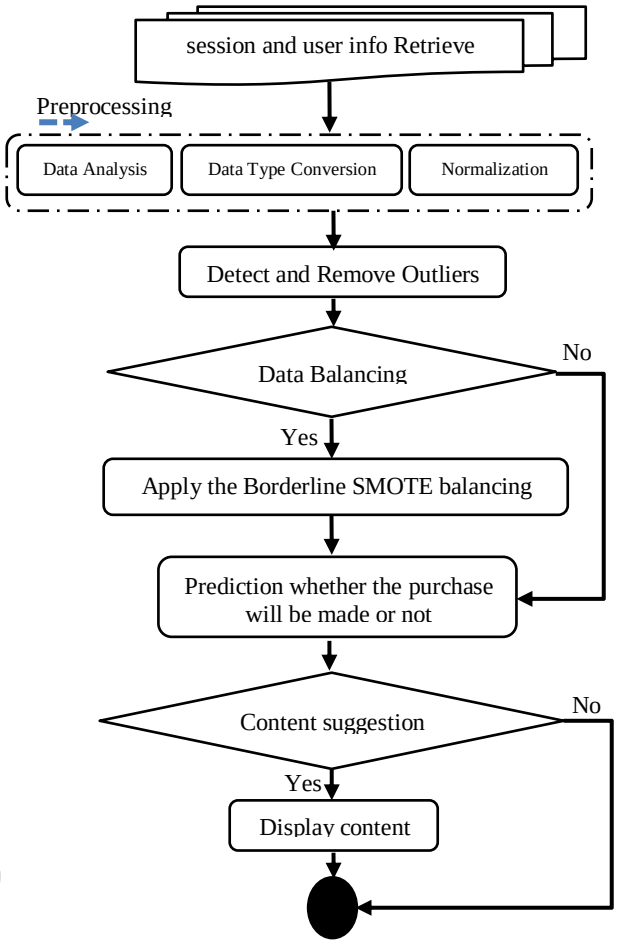
| | |
|---|---|
| Month | ماهی که بازدید در آن صورت گرفته است |
| Revenue | بر چسب کلاس که نشان می‌دهد که مشاهده سایت توسط بازدید کننده به تراکنش منجر شده است یا خیر. |

**جدول ۲: 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. | ۰ | ۲۷ | ۲/۳۱۵ | ۳/۳۲۲ |
| Administrative duration | کل زمانی (بر حسب ثانیه) که توسط بازدیدکننده در صفحات مربوط به مدیریت حساب صرف شده است. | ۰ | ۳۳۹۸/۷۵ | ۸۰/۸۱۹ | ۱۷۶/۷۷۹ |
| Informational | تعداد صفحاتی که بازدیدکننده در مورد وب سایت، ارتباطات و اطلاعات آدرس سایت خرید، ملاقات کرده است. | ۰ | ۲۴ | ۰/۵۰۴ | ۱/۲۷ |
| Informational duration | کل زمانی (بر حسب ثانیه) که توسط بازدیدکننده در صفحات اطلاعاتی صرف شده است. | ۰ | ۲۵۴۹/۳۷۵ | ۳۴/۴۷۲ | ۱۴۰/۷۴۹ |
| Product related | تعداد صفحاتی که بازدیدکننده درباره صفحات مرتبط با محصول ملاقات کرده است. | ۰ | ۷۰۵ | ۳۱/۷۳۱ | ۴۴/۴۷۶ |
| Product related duration | کل زمانی (بر حسب ثانیه) که توسط بازدیدکننده در صفحات مرتبط با محصول صرف شده است. | ۰ | ۶۳۹۷۳/۵۲۲ | ۱۱۹۴/۷۴۶ | ۱۹۱۳/۶۶۹ |
| Bounce rate | مقدار میانگین نرخ پرش صفحاتی که توسط بازدیدکننده، بازدید شده است. | ۰ | ۰/۲ | ۰/۰۲۲ | ۰/۰۴۸ |
| Exit rate | مقدار میانگین نرخ خروج از صفحاتی که توسط بازدیدکننده بازدید شده است. | ۰ | ۰/۲ | ۰/۰۴۳ | ۰/۰۴۹ |
| Page value | میانگین مقدار صفحه از صفحاتی که توسط بازدیدکننده بازدید شده است. | ۰ | ۳۶۱/۷۶۴ | ۵/۸۸۹ | ۱۸/۵۶۸ |
| Special day | نزدیک بودن زمان بازدید از سایت به یک روز خاص | ۰ | ۱ | ۰/۰۶۱ | ۰/۱۹۹ |

## ۲،۲. 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 دانش-های 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 [۱۵]. 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 شماره ۱ نشان داده شده‌اند. در ادامه جزئیات هر یک از این گام‌ها بیان شده است.




شکل شماره ۱ : مراحل روش پیشنهادی

### ۳،۲. پیش پردازش داده‌ها

کیفیت نتایج خروجی یک الگوریتم رابطه مستقیمی با کیفیت داده‌های ورودی آن دارد. هر چه داده‌های با کیفیت-تری 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.

آماده‌سازی و انتقال داده‌ها به یک شکل مناسب برای فرآیند استخراج دانش از داده‌ها پیش پردازش گفته می‌شود که یکی از مهم‌ترین وظایف داده‌کاوی است [۱۶]. برای این منظور، در این مقاله گام‌های زیر به عنوان مراحل پیش-پردازش 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 بر 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.

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 نرمال-سازی Z-Score [۱۷] استفاده می‌شود. روابط مورد استفاده در این روش نرمال‌سازی به شکل زیر می-باشد:

$$x_i = \frac{v_i - \mu}{\sigma} \quad (۱)$$

$$\mu = \frac{1}{n}\sum_{i=1}^{n} v_i \quad (۲)$$

$$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(v_i - \mu)^2} \quad (۳)$$

که در آن $v_i$ مقدار iام ویژگی، $\mu$ میانگین و $\sigma$ 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 تست‌های 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.

### ۴،۲. تشخیص و حذف داده‌های پرت

در مجموعه داده‌ها ممکن است نمونه‌هایی وجود داشته باشد که با دیگر نمونه‌ها تفاوت چشمگیری دارند. این نمونه-ها 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 داده‌هایی 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 [۱۸]. در این پژوهش به جهت عملکرد بهتر الگوریتم‌های طبقه‌بندی و جلوگیری از تاثیر داده‌های پرت بر روی نتایج خروجی، داده‌های پرت با استفاده از الگوریتم Local Outlier Factor [۱۹] 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.

### ۵،۲. متوازن سازی داده‌ها

با توجه به نامتوازن بودن مجموعه داده مورد استفاده (۱۰۴۴۲ تراکنش بدون خرید در مقابل ۱۹۰۸ 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 معمولاً 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 [۲۰]. 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 [۲۱] 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.

### ۶،۲. پیش‌بینی نتیجه تراکنش

هدف نهایی یک مجموعه فروش آنلاین، 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 سایت را افزایش داد. جهت پیش‌بینی قصد خرید بازدیدکنندگان 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 به کار گرفته شده‌اند می‌توان به درخت تصمیم، جنگل تصادفی و شبکه عصبی پرسپترون چندلایه(MLP[۱۴]) اشاره نمود.

در این پژوهش برای پیش‌بینی قصد خرید مشتریان آنلاین 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 استثناء الگوریتم‌های MLP و درخت تصمیم، سایر الگوریتم‌ها مورد استفاده از نوع تلفیقی می‌باشند.

## ۳. 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.

### ۱،۳. 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:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FN + FP} \quad (۴)$$

$$\text{Precision} = \frac{TP}{TP + FP} \quad (۵)$$

14 Multi-Layer Perceptron

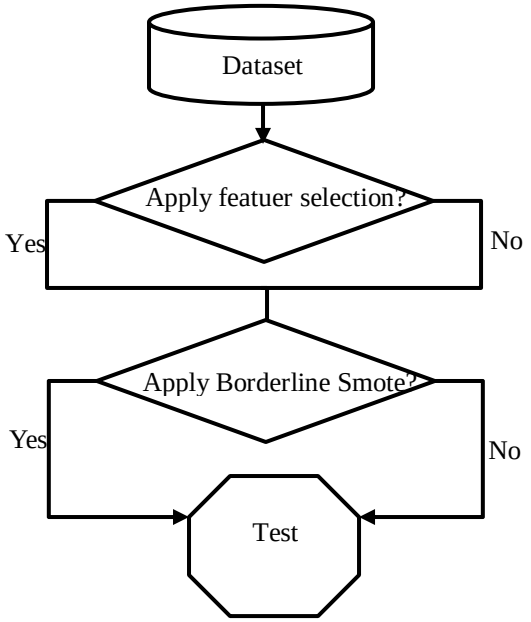
$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (۶)$$

$$\text{F\_measure} = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \quad (۷)$$

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 واقعاً کلاس آنها مثبت است. پارامتر recall نیز یک معیار تمامیت است که نشان‌دهنده درصدی از نمونه-های 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[15] نیز به عنوان بخشی از فرایند ارزیابی میزان دقت طبقه‌بندهای مختلف [۱۷]، رسم شده است.

سناریو ارزیابی روش پیشنهادی در شکل ۲ نشان داده شده است. همانطور که در شکل مشخص است، ترکیبی از حالت‌های مختلف استفاده/عدم استفاده از روش متوازن‌سازی Borderline SMOTE، در کنار استفاده/عدم استفاده از روش‌های انتخاب ویژگی و ترکیب این حالت‌ها با یکدیگر در سناریو ارزیابی این مقاله قرار گرفته است.

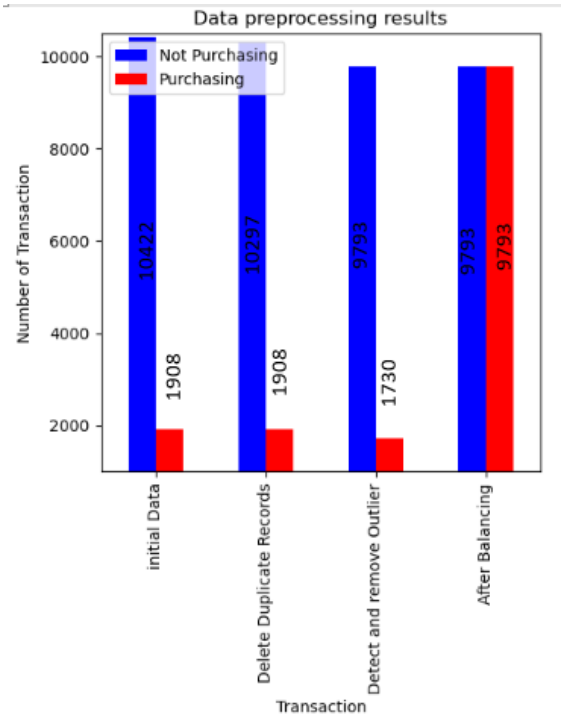



شکل شماره ۲ : سناریو ارزیابی روش پیشنهادی

### ۱،۳. ارزیابی نتایج حاصل از متوازن‌سازی و عدم متوازن‌سازی داده‌ها

شکل شماره ۳ 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 ۸۴/۴٪ رکوردها به کلاس منفی و بقیه به کلاس مثبت تعلق دارند. بنابراین فرایند متوازن‌سازی داده‌ها برای بهبود عملکرد الگوریتم‌های طبقه‌بندی ضروری است. نتایج اعمال الگوریتم-های طبقه‌بندی بدون/با متوازن‌سازی در جداول ۳ و ۴ آورده شده است.

15 receiver operating characteristic




**شکل شماره ۳: الگوی توزیع داده‌ها**

همان گونه که از جدول شماره ۳ 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 ۴ گویای اثر مثبت استفاده از الگوریتم متوازن‌سازی Borderline SMOTE و رفع این مشکل می‌باشد.

**جدول شماره ۳: نتایج به دست آمده بدون متوازن‌سازی داده‌ها**

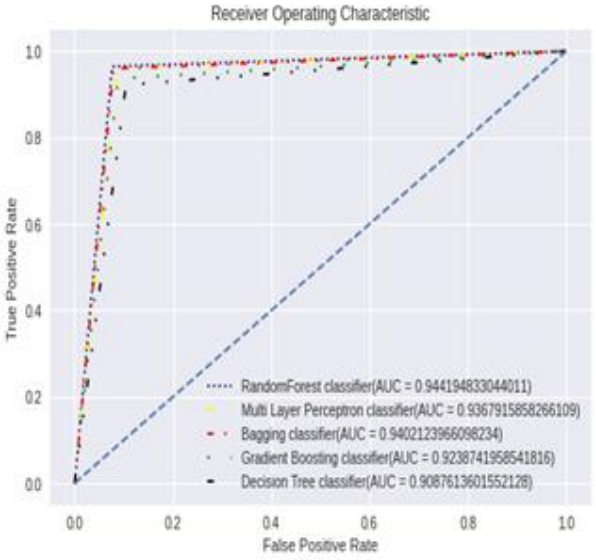
| Classifier Algorithm | بدون متعادل سازی داده‌ها | | | |
|---|---|---|---|---|
| | Accuracy | Precision | Recall | F_measure |
| Random Forest | ۰/۹۰۴۲ | ۰/۷۴۷۴ | ۰/۵۴۷۴ | ۰/۶۳۱۳ |
| MLP | ۰/۸۶۹۰ | ۰/۵۶۶۲ | ۰/۵۵۶۶ | ۰/۵۵۹۱ |
| Bagging | ۰/۹۰۲۰ | ۰/۷۱۵۴ | ۰/۵۷۸۰ | ۰/۶۳۸۹ |
| Gradient Boosting | ۰/۹۰۴۷ | ۰/۷۲۱۱ | ۰/۵۹۸۳ | ۰/۶۵۳۲ |
| Decision Tree | ۰/۸۶۴۹ | ۰/۵۴۹۷ | ۰/۵۶۳۰ | ۰/۵۵۵۶ |

**جدول شماره ۴: نتایج به دست آمده با متوازن‌سازی داده‌ها**

| Classifier Algorithm | بعد از متعادل سازی داده‌ها | | | |
|---|---|---|---|---|
| | Accuracy | Precision | Recall | F_measure |
| Random Forest | **۰/۹۴۴۲** | **۰/۹۲۷۰** | **۰/۹۶۴۴** | **۰/۹۴۵۳** |
| MLP | ۰/۹۳۶۸ | ۰/۹۱۶۷ | ۰/۹۶۱۲ | ۰/۹۳۸۳ |
| Bagging | ۰/۹۴۰۲ | ۰/۹۲۴۰ | ۰/۹۵۹۴ | ۰/۹۴۱۳ |
| Gradient Boosting | ۰/۹۲۳۹ | ۰/۹۱۳۹ | ۰/۹۳۶۰ | ۰/۹۲۴۸ |
| Decision Tree | ۰/۹۰۸۸ | ۰/۸۹۸۶ | ۰/۹۲۱۸ | ۰/۹۱۰۰ |

شکل شماره ۴ نمودار ROC حاصل از اعمال الگوریتم-های 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[۱۶]) اندازه گیری می‌شود. هر چه مقدار AUC به مقدار ۰/۵ نزدیک‌تر باشد، مدل مزبور از صحت کمتری برخوردار است. یک مدل ایده‌آل و کامل از نظر صحت، دارای مقدار AUC برابر با ۱ می‌باشد. همان گونه که از این شکل و جدول شماره ۴ مشهود است، طبقه‌بند جنگل تصادفی بهتر از دیگر طبقه‌بندها عمل کرده است. شایان ذکر است که کلیه نتایج این مرحله، بدون اعمال الگوریتم‌های کاهش ویژگی می‌باشد. در بخش بعد، تاثیر الگوریتم‌های کاهش ویژگی نیز بررسی شده است.

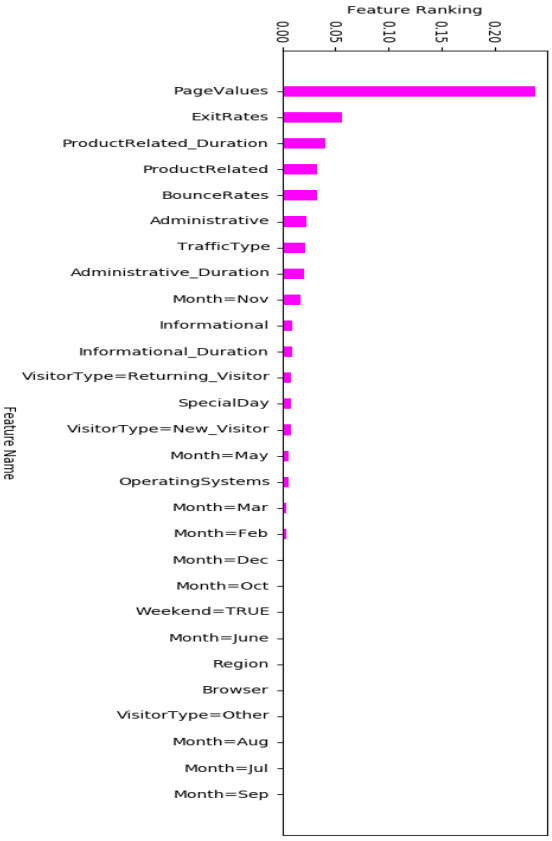
16 Area Under the Curve



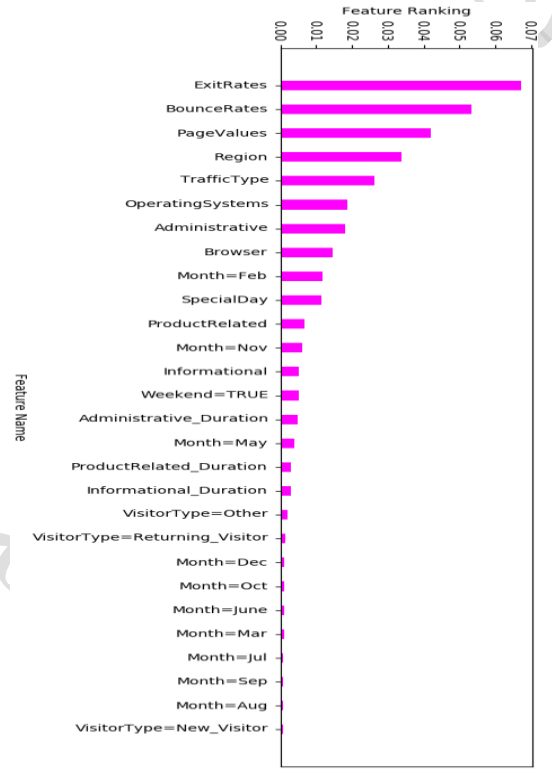

شکل شماره ۴: نمودار ROC حاصل از اعمال الگوریتم‌های طبقه‌بندی بر روی مجموعه داده متوازن

## ۲،۳. ارزیابی نتایج حاصل از اعمال روش‌های کاهش ویژگی

به منظور کاهش تعداد ویژگی‌های استفاده شده در این پژوهش ما از ۲ روش مشهور و متداول ReliefF [۲۲] و Information Gain [۲۲] استفاده می‌کنیم. این دو روش در بین روش‌های انتخاب ویژگی مبتنی بر رتبه هستند. شکل‌های شماره ۵- الف و ۵- 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، ۱۰ ویژگی برتر از هر روش کاهش ویژگی را انتخاب شده است. در ادامه، عملیات‌های پیش‌پردازش داده‌ها، تشخیص و حذف داده‌های پرت، متعادل سازی داده‌ها و اعمال الگوریتم‌های طبقه‌بندی بر روی ویژگی‌های انتخاب شده انجام شده است.




الف: روش Information Gain




ب: روش ReliefF

شکل شماره ۵: رتبه ویژگی‌های مختلف به کمک دو روش انتخاب ویژگی

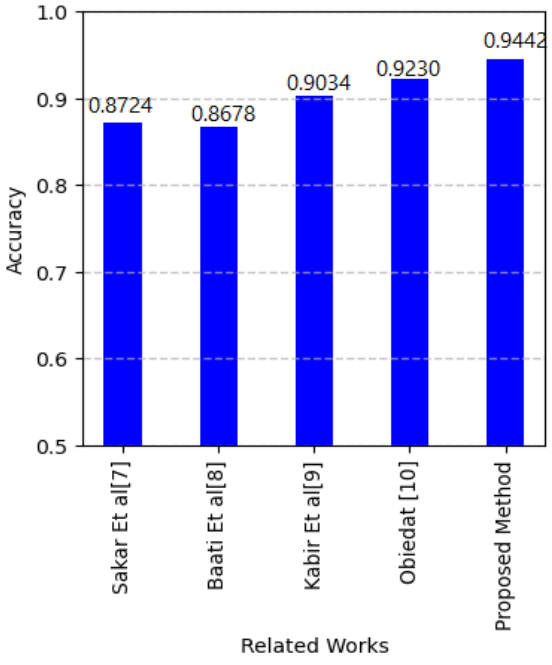
نتایج به دست آمده با اعمال روش‌های کاهش ویژگی ذکر شده در جدول ۵ نشان داده شده است. همان گونه که مشهود است، مقادیر به دست آمده از هر دو روش کاهش ویژگی تقریباً یکسان می‌باشد. به گونه‌ای که در ابتدای کار که تعداد ۱۰ ویژگی برتر از هر روش انتخاب شده‌اند، فقط ۵ ویژگی مشترک در بین آن‌ها وجود دارد ولی مقادیر 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 ۱۰، با گام ۳، نیز تکرار شد که نتایج آنها نیز در جدول ۵ 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 ۵ مشهود است، در مجموعه داده مورد استفاده، با افزایش تعداد ویژگی-های منتخب از ۱۰ عدد (کمترین دقت) به ۲۵ ویژگی منتخب و حتی تمامی ویژگی‌ها (بیشترین دقت)، در بهترین حالت دقت به اندازه ۱/۵ 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. بعلاوه 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 ۶ در بخش ضمایم ارائه شده است.

### ۳،۳. مقایسه نتایج کسب شده با کارهای مشابه

بیشترین دقت به دست آمده در این پژوهش برابر با ۹۴/۴۲ درصد می‌باشد، که با استفاده از طبقه‌بند جنگل تصادفی بر روی داده‌های متعادل به دست آمده است. همانطور که در شکل ۶ 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.




شکل شماره ۶ : مقایسه بهترین دقت به دست آمده به کمک روش پیشنهادی و سایر کارهای مرتبط

**جدول شماره ۵: ارزیابی نتایج حاصل از کاهش ویژگی**

| نتایج به دست آمده با ۱۰ 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 | **۰/۹۳۵۷** | **۰/۹۱۲۹** | **۰/۹۶۳۶** | **۰/۹۳۷۵** | **۰/۹۲۹۱** | ۰/۹۰۹۶ | **۰/۹۵۳۰** | **۰/۹۳۰۷** |
| MLP | ۰/۹۲۶۸ | ۰/۹۰۶۲ | ۰/۹۵۳۰ | ۰/۹۲۸۸ | ۰/۹۱۷۶ | ۰/۸۹۳۴ | ۰/۹۴۹۰ | ۰/۹۲۰۱ |
| Bagging | ۰/۹۳۳۷ | ۰/۹۱۱۹ | ۰/۹۶۰۳ | ۰/۹۳۵۴ | ۰/۹۲۸۹ | **۰/۹۱۴۳** | ۰/۹۴۶۶ | ۰/۹۳۰۱ |
| Gradient Boosting | ۰/۸۹۵۷ | ۰/۸۷۷۹ | ۰/۹۱۹۶ | ۰/۸۹۸۲ | ۰/۸۹۵۱ | ۰/۸۹۷۶ | ۰/۸۹۲۱ | ۰/۸۹۴۸ |
| Decision Tree | ۰/۸۹۷۷ | ۰/۸۸۹۳ | ۰/۹۰۸۵ | ۰/۸۹۸۸ | ۰/۸۹۲۳ | ۰/۸۸۵۰ | ۰/۹۰۱۸ | ۰/۸۹۳۳ |
| **نتایج به دست آمده با ۱۳ 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 | **۰/۹۳۶۷** | ۰/۹۱۲۹ | **۰/۹۶۵۶** | **۰/۹۳۸۵** | **۰/۹۳۹۶** | **۰/۹۲۰۳** | **۰/۹۶۲۶** | **۰/۹۴۱۰** |
| MLP | ۰/۹۳۱۹ | ۰/۹۰۶۵ | ۰/۹۶۳۱ | ۰/۹۳۳۹ | ۰/۹۲۹۱ | ۰/۹۰۹۳ | ۰/۹۵۳۵ | ۰/۹۳۰۸ |
| Bagging | ۰/۹۳۴۴ | **۰/۹۱۳۱** | ۰/۹۶۰۴ | ۰/۹۳۶۱ | ۰/۹۳۶۹ | **۰/۹۲۰۳** | ۰/۹۵۶۶ | ۰/۹۳۸۱ |
| Gradient Boosting | ۰/۹۰۱۵ | ۰/۸۸۳۴ | ۰/۹۲۵۳ | ۰/۹۰۳۸ | ۰/۹۱۲۳ | ۰/۸۹۷۱ | ۰/۹۳۱۶ | ۰/۹۱۴۰ |
| Decision Tree | ۰/۹۰۲۰ | ۰/۸۹۲۸ | ۰/۹۱۳۸ | ۰/۹۰۳۲ | ۰/۹۰۳۸ | ۰/۸۹۵۵ | ۰/۹۱۴۳ | ۰/۹۰۴۸ |
| **نتایج به دست آمده با ۱۶ 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 | **۰/۹۴۰۵** | **۰/۹۲۱۴** | **۰/۹۶۳۲** | **۰/۹۴۱۹** | **۰/۹۴۱۸** | **۰/۹۲۶۷** | ۰/۹۵۹۷ | **۰/۹۴۲۹** |
| MLP | ۰/۹۳۳۲ | ۰/۹۱۷۶ | ۰/۹۵۲۳ | ۰/۹۳۴۴ | ۰/۹۳۴۳ | ۰/۹۱۱۵ | **۰/۹۶۲۲** | ۰/۹۳۶۱ |
| Bagging | ۰/۹۳۸۳ | ۰/۹۲۰۸ | ۰/۹۵۹۲ | ۰/۹۳۹۶ | ۰/۹۳۹۰ | ۰/۹۲۶۱ | ۰/۹۵۴۳ | ۰/۹۳۹۹ |
| Gradient Boosting | ۰/۹۱۳۱ | ۰/۸۹۷۶ | ۰/۹۳۲۶ | ۰/۹۱۴۷ | ۰/۹۱۵۶ | ۰/۹۰۸۳ | ۰/۹۲۴۷ | ۰/۹۱۶۴ |
| Decision Tree | ۰/۹۰۵۲ | ۰/۸۹۴۲ | ۰/۹۱۹۳ | ۰/۹۰۶۵ | ۰/۹۰۸۳ | ۰/۹۰۰۹ | ۰/۹۱۷۶ | ۰/۹۰۹۱ |
| **نتایج به دست آمده با ۱۹ 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 | **۰/۹۴۱۸** | **۰/۹۲۲۵** | **۰/۹۶۴۸** | **۰/۹۴۳۲** | **۰/۹۴۲۹** | ۰/۹۲۷۵ | **۰/۹۶۱۲** | **۰/۹۴۴۰** |
| MLP | ۰/۹۳۷۴ | ۰/۹۱۷۶ | ۰/۹۶۱۵ | ۰/۹۳۸۹ | ۰/۹۳۱۱ | ۰/۹۱۲۵ | ۰/۹۵۴۲ | ۰/۹۳۲۶ |
| Bagging | ۰/۹۳۶۵ | ۰/۹۱۷۸ | ۰/۹۵۹۱ | ۰/۹۳۷۹ | ۰/۹۴۱۲ | **۰/۹۲۷۹** | ۰/۹۵۷۰ | ۰/۹۴۲۲ |
| Gradient Boosting | ۰/۹۰۸۹ | ۰/۸۹۵۹ | ۰/۹۲۵۶ | ۰/۹۱۰۴ | ۰/۹۱۸۶ | ۰/۹۱۱۴ | ۰/۹۲۷۵ | ۰/۹۱۹۳ |
| Decision Tree | ۰/۹۰۵۰ | ۰/۸۹۸۸ | ۰/۹۱۲۹ | ۰/۹۰۵۷ | ۰/۹۰۶۴ | ۰/۹۰۰۹ | ۰/۹۱۳۴ | ۰/۹۰۷۱ |
| **نتایج به دست آمده با ۲۲ 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 | **۰/۹۴۲۲** | **۰/۹۲۳۰** | **۰/۹۶۵۰** | **۰/۹۴۳۵** | **۰/۹۴۳۰** | **۰/۹۲۷۹** | **۰/۹۶۰۷** | **۰/۹۴۴۰** |
| MLP | ۰/۹۳۷۵ | ۰/۹۱۸۶ | ۰/۹۶۰۱ | ۰/۹۳۸۸ | ۰/۹۳۴۷ | ۰/۹۱۵۴ | ۰/۹۵۸۲ | ۰/۹۳۶۲ |
| Bagging | ۰/۹۳۸۳ | ۰/۹۲۱۲ | ۰/۹۵۸۶ | ۰/۹۳۹۵ | ۰/۹۴۱۳ | ۰/۹۲۷۶ | ۰/۹۵۷۲ | ۰/۹۴۲۲ |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Gradient Boosting | ۰/۹۱۳۳ | ۰/۸۹۸۹ | ۰/۹۳۱۵ | ۰/۹۱۴۹ | ۰/۹۲۲۹ | ۰/۹۱۰۵ | ۰/۹۳۸۱ | ۰/۹۲۴۱ |
| Decision Tree | ۰/۹۰۶۵ | ۰/۸۹۸۶ | ۰/۹۱۶۴ | ۰/۹۰۷۴ | ۰/۹۰۶۶ | ۰/۸۹۸۰ | ۰/۹۱۷۶ | ۰/۹۰۷۷ |

**نتایج به دست آمده با ۲۵ 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 | **۰/۹۴۲۷** | **۰/۹۲۶۵** | ۰/۹۶۱۷ | **۰/۹۴۳۸** | **۰/۹۴۳۵** | **۰/۹۲۷۱** | ۰/۹۶۲۹ | **۰/۹۴۴۶** |
| MLP | ۰/۹۳۶۷ | ۰/۹۱۴۲ | **۰/۹۶۴۰** | ۰/۹۳۸۴ | ۰/۹۳۳۹ | ۰/۹۰۹۲ | **۰/۹۶۴۶** | ۰/۹۳۶۰ |
| Bagging | ۰/۹۴۱۱ | ۰/۹۲۵۳ | ۰/۹۵۹۶ | ۰/۹۴۲۲ | ۰/۹۴۱۱ | ۰/۹۲۵۴ | ۰/۹۵۹۸ | ۰/۹۴۲۲ |
| Gradient Boosting | ۰/۹۲۲۴ | ۰/۹۱۱۳ | ۰/۹۳۶۰ | ۰/۹۲۳۵ | ۰/۹۲۴۴ | ۰/۹۱۳۶ | ۰/۹۳۷۷ | ۰/۹۲۵۵ |
| Decision Tree | ۰/۹۰۹۱ | ۰/۹۰۱۲ | ۰/۹۱۸۹ | ۰/۹۰۹۹ | ۰/۹۰۶۸ | ۰/۸۹۸۶ | ۰/۹۱۷۱ | ۰/۹۰۷۷ |

**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 | **۰/۹۴۴۲** | **۰/۹۲۷۰** | **۰/۹۶۴۴** | **۰/۹۴۵۳** | **۰/۹۴۴۲** | **۰/۹۲۷۰** | **۰/۹۶۴۴** | **۰/۹۴۵۳** |
| MLP | ۰/۹۳۶۸ | ۰/۹۱۶۷ | ۰/۹۶۱۲ | ۰/۹۳۸۳ | ۰/۹۳۶۸ | ۰/۹۱۶۷ | ۰/۹۶۱۲ | ۰/۹۳۸۳ |
| Bagging | ۰/۹۴۰۲ | ۰/۹۲۴۰ | ۰/۹۵۹۴ | ۰/۹۴۱۳ | ۰/۹۴۰۲ | ۰/۹۲۴۰ | ۰/۹۵۹۴ | ۰/۹۴۱۳ |
| Gradient Boosting | ۰/۹۲۳۹ | ۰/۹۱۳۹ | ۰/۹۳۶۰ | ۰/۹۲۴۸ | ۰/۹۲۳۹ | ۰/۹۱۳۹ | ۰/۹۳۶۰ | ۰/۹۲۴۸ |
| Decision Tree | ۰/۹۰۸۸ | ۰/۸۹۸۶ | ۰/۹۲۱۸ | ۰/۹۱۰۰ | ۰/۹۰۸۸ | ۰/۸۹۸۶ | ۰/۹۲۱۸ | ۰/۹۱۰۰ |

## ۴. جمع بندی

این مقاله رفتار خریداران آنلاین را برای پیش‌بینی اینکه آیا آن‌ها یک محصول را در بازدید از وب سایت 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 پیش پردازش 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 پیش‌بینی کننده 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 ۹۴/۴۲ 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.

## 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." *In Journal of Physics: Conference Series*, vol.1641, no.012083, pp.1-8, 2020. https://doi.org/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." *Technology Analysis & Strategic Management*, vol. 32, no.7, pp.835-850, 2020. https://doi.org/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." *Journal of Direct, Data and Digital Marketing Practice*, vol. 15, no.4, pp. 317-326, 2014. https://doi.org/10.1057/dddmp.2014.24

[5] N. Gudigantala, P. Bicen, and M. Eom. "An examination of antecedents of conversion rates of e-commerce retailers." *Management Research Review*, vol. 39, no.1, pp: 82-114, 2016. https://doi.org/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.", *journal of Telecommunications and Information Technology*. 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 Computing and Applications*, vol. 31, pp: 6893-6908, 2019. https://doi.org/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 International Conference on Artificial Intelligence Applications and Innovations* 16th IFIP WG 12.5 International Conference, AIAI 2020, Neos Marmaras, Greece, June 5–7, 2020, Proceedings, Part I 16, Pp: 43-51, 2020. https://doi.org/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 2019 22nd International Conference on Computer and Information Technology (ICCIT).* Pp:1-6, 2019. https://doi.org/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.", *International Journal of Advanced Trends in Computer Science and Engineering*, vol.9, no. 3, pp: 3575-3583 ,2020. https://doi.org/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 Computing Journal*, pp: 1-21, 2023. https://doi.org/10.22052/SCJ.2023.248522.1110

[13] M. Mousavi, S. Hosseini, and M. R. Omidi. "Improved Deep Neural Network Algorithm for Covid-19 Detection in Internet of Things." *Soft Computing Journal*, pp: 1-19, 2023. http://doi.org/10.22052/SCJ.2023.248686.1117

[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 Computing Journal*, pp: 1-21, 2023. http://doi.org/10.22052/SCJ.2023.252793.1144

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

[16] S. A. Alasadi, and W. S. Bhaya. "Review of data preprocessing techniques in data mining." *Journal of Engineering and Applied Sciences.* Vol. 12, no. 16, pp: 4102-4107 ,2017. https:// doi.org/10.36478/jeasci.2017.4102.4107

[17] J.Han, M.Kamber and J.Pei. *Data mining: concepts and techniques. Third Edition*, Morgan Kaufmann , 2012. https://doi.org/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 and Knowledge Discovery*. Vol. 8, no. 6, pp:: 1-37 ,2018,. https://doi.org/10.1002/widm.1280

[19] M. M. Breunig, H. P. Kriegel, R. T. Ng, and J. Sander. "LOF: identifying density-based local outliers." *In Proceedings of the 2000 ACM SIGMOD international conference on Management of data*. Pp: 93-104,2000. https://doi.org/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." *Information Sciences.* vol. 477, pp: 47-54, 2019. https://doi.org/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 International conference on intelligent computing.* Pp:878-887, 2005. https://doi.org/10.1007/11538059_91

[22] P. Yildirim, "Filter based feature selection methods for prediction of risks in hepatitis disease." *International Journal of Machine Learning and Computing*, vol. 5, no. 4, pp: 258-263, 2015. https://doi.org/10.7763/IJMLC.2015.V5.517

## ضمایم

**جدول شماره ۶: ارزیابی نتایج حاصل از کاهش ویژگی**

| نتایج به دست آمده با ۱۰ 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 | ۰/۸۹۹۴ | ۰/۷۳۳۱ | ۰/۵۸۰۸ | ۰/۶۴۷۰ | ۰/۸۹۳۵ | ۰/۶۹۶۶ | ۰/۵۳۷۳ | ۰/۶۰۶۱ |
| MLP | ۰/۸۷۷۵ | ۰/۶۳۴۳ | ۰/۵۶۲۸ | ۰/۵۹۴۰ | ۰/۸۶۹۱ | ۰/۵۸۴۹ | ۰/۴۹۸۶ | ۰/۵۳۷۰ |
| Bagging | ۰/۸۹۶۳ | ۰/۷۱۸۵ | ۰/۵۷۹۱ | ۰/۶۳۹۹ | ۰/۸۹۱۲ | ۰/۶۸۱۹ | ۰/۵۳۸۴ | ۰/۶۰۱۰ |
| Gradient Boosting | ۰/۸۹۹۰ | ۰/۷۲۰۸ | ۰/۵۹۸۷ | ۰/۶۵۳۲ | ۰/۸۹۶۴ | ۰/۷۰۳۵ | ۰/۵۵۶۳ | ۰/۶۲۰۸ |
| Decision Tree | ۰/۸۵۱۵ | ۰/۵۳۳۴ | ۰/۵۵۹۰ | ۰/۵۴۵۵ | ۰/۸۴۸۱ | ۰/۵۰۲۴ | ۰/۵۱۷۱ | ۰/۵۰۸۷ |
| **نتایج به دست آمده با ۱۳ 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 | ۰/۹۰۰۰ | ۰/۷۲۹۳ | ۰/۵۶۵۶ | ۰/۶۳۶۴ | ۰/۹۰۲۸ | ۰/۷۳۵۹ | ۰/۵۶۹۶ | ۰/۶۴۱۶ |
| MLP | ۰/۸۶۹۱ | ۰/۵۸۳۶ | ۰/۵۴۳۵ | ۰/۵۶۱۹ | ۰/۸۶۹۲ | ۰/۵۸۵۷ | ۰/۵۲۹۸ | ۰/۵۵۴۸ |
| Bagging | ۰/۸۹۸۴ | ۰/۷۱۰۷ | ۰/۵۸۳۸ | ۰/۶۴۰۴ | ۰/۸۹۸۱ | ۰/۷۰۳۲ | ۰/۵۷۹۶ | ۰/۶۳۵۱ |
| Gradient Boosting | ۰/۸۹۸۸ | ۰/۷۱۰۴ | ۰/۵۹۰۱ | ۰/۶۴۳۸ | ۰/۹۰۲۸ | ۰/۷۱۸۸ | ۰/۶۰۰۰ | ۰/۶۵۳۶ |
| Decision Tree | ۰/۸۶۱۰ | ۰/۵۵۱۵ | ۰/۵۶۵۶ | ۰/۵۵۸۱ | ۰/۸۵۵۳ | ۰/۵۲۸۶ | ۰/۵۴۱۴ | ۰/۵۳۴۲ |
| **نتایج به دست آمده با ۱۶ 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 | ۰/۹۰۱۹ | ۰/۷۳۵۳ | ۰/۵۶۰۵ | ۰/۶۳۵۵ | ۰/۹۰۲۱ | ۰/۷۳۴۰ | ۰/۵۶۶۱ | ۰/۶۳۸۷ |
| MLP | ۰/۸۷۶۱ | ۰/۶۰۴۹ | ۰/۵۵۱۲ | ۰/۵۷۵۳ | ۰/۸۶۵۱ | ۰/۵۶۰۵ | ۰/۵۵۱۷ | ۰/۵۵۴۹ |
| Bagging | ۰/۹۰۰۶ | ۰/۷۲۲۸ | ۰/۵۶۷۴ | ۰/۶۳۵۲ | ۰/۸۹۸۰ | ۰/۷۰۴۳ | ۰/۵۷۴۹ | ۰/۶۳۲۶ |
| Gradient Boosting | ۰/۹۰۳۱ | ۰/۷۱۹۴ | ۰/۶۰۰۶ | ۰/۶۵۴۰ | ۰/۹۰۱۸ | ۰/۷۱۳۸ | ۰/۵۹۷۶ | ۰/۶۵۰۲ |
| Decision Tree | ۰/۸۵۸۷ | ۰/۵۳۷۱ | ۰/۵۴۸۴ | ۰/۵۴۲۲ | ۰/۸۶۰۰ | ۰/۵۴۱۳ | ۰/۵۵۶۱ | ۰/۵۴۸۳ |
| **نتایج به دست آمده با ۱۹ 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 | ۰/۹۰۲۷ | ۰/۷۴۱۱ | ۰/۵۶۱۴ | ۰/۶۳۷۱ | ۰/۹۰۳۶ | ۰/۷۴۱۶ | ۰/۵۶۸۸ | ۰/۶۴۲۸ |
| MLP | ۰/۸۷۶۶ | ۰/۶۰۸۴ | ۰/۵۵۰۵ | ۰/۵۷۴۶ | ۰/۸۷۰۹ | ۰/۵۸۷۰ | ۰/۵۳۹۹ | ۰/۵۶۰۷ |
| Bagging | ۰/۸۹۹۲ | ۰/۷۱۳۲ | ۰/۵۷۴۶ | ۰/۶۳۵۰ | ۰/۹۰۰۵ | ۰/۷۲۰۲ | ۰/۵۷۵۴ | ۰/۶۳۸۴ |
| Gradient Boosting | ۰/۹۰۲۴ | ۰/۷۲۲۴ | ۰/۵۹۱۸ | ۰/۶۴۸۸ | ۰/۹۰۱۵ | ۰/۷۱۴۲ | ۰/۵۹۷۴ | ۰/۶۵۰۲ |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Decision Tree | ۰/۸۶۰۸ | ۰/۵۴۴۶ | ۰/۵۵۸۵ | ۰/۵۵۰۹ | ۰/۸۵۹۹ | ۰/۵۴۰۵ | ۰/۵۶۸۲ | ۰/۵۵۳۴ |

**نتایج به دست آمده با ۲۲ 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 | ۰/۹۰۳۰ | ۰/۷۴۵۵ | ۰/۵۴۴۷ | ۰/۶۲۹۲ | ۰/۹۰۲۶ | ۰/۷۴۳۲ | ۰/۵۵۹۶ | ۰/۶۳۷۳ |
| MLP | ۰/۸۷۱۴ | ۰/۵۸۵۸ | ۰/۵۲۶۱ | ۰/۵۵۲۸ | ۰/۸۷۱۶ | ۰/۵۸۸۱ | ۰/۵۵۹۰ | ۰/۵۷۱۷ |
| Bagging | ۰/۹۰۱۷ | ۰/۷۲۰۴ | ۰/۵۷۲۲ | ۰/۶۳۷۳ | ۰/۹۰۲۱ | ۰/۷۲۴۸ | ۰/۵۸۵۷ | ۰/۶۴۷۱ |
| Gradient Boosting | ۰/۹۰۴۰ | ۰/۷۲۴۳ | ۰/۵۹۰۴ | ۰/۶۴۹۸ | ۰/۹۰۳۱ | ۰/۷۲۴۱ | ۰/۵۹۷۴ | ۰/۶۵۳۴ |
| Decision Tree | ۰/۸۶۳۸ | ۰/۵۴۸۸ | ۰/۵۵۷۶ | ۰/۵۵۲۳ | ۰/۸۵۸۲ | ۰/۵۳۴۷ | ۰/۵۶۰۱ | ۰/۵۴۶۸ |

**نتایج به دست آمده با ۲۵ 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 | ۰/۹۰۴۰ | ۰/۷۴۸۵ | ۰/۵۵۶۶ | ۰/۶۳۷۷ | ۰/۹۰۳۵ | ۰/۷۳۹۸ | ۰/۵۶۴۱ | ۰/۶۳۹۲ |
| MLP | ۰/۸۷۳۵ | ۰/۵۹۴۱ | ۰/۵۴۱۴ | ۰/۵۶۵۶ | ۰/۸۷۲۲ | ۰/۵۸۴۱ | ۰/۵۶۲۴ | ۰/۵۷۱۴ |
| Bagging | ۰/۹۰۱۹ | ۰/۷۲۱۶ | ۰/۵۷۹۷ | ۰/۶۴۲۴ | ۰/۹۰۱۳ | ۰/۷۱۵۳ | ۰/۵۸۵۰ | ۰/۶۴۲۶ |
| Gradient Boosting | ۰/۹۰۳۵ | ۰/۷۲۱۸ | ۰/۵۹۶۰ | ۰/۶۵۲۴ | ۰/۹۰۴۱ | ۰/۷۲۶۰ | ۰/۵۹۳۴ | ۰/۶۵۲۰ |
| Decision Tree | ۰/۸۶۱۴ | ۰/۵۴۳۳ | ۰/۵۶۵۶ | ۰/۵۵۳۹ | ۰/۸۵۹۰ | ۰/۵۳۴۷ | ۰/۵۶۰۲ | ۰/۵۴۶۷ |

**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 | ۰/۹۰۴۲ | ۰/۷۴۷۴ | ۰/۵۴۷۴ | ۰/۶۳۱۳ | ۰/۹۰۴۲ | ۰/۷۴۷۴ | ۰/۵۴۷۴ | ۰/۶۳۱۳ |
| MLP | ۰/۸۶۹۰ | ۰/۵۶۶۲ | ۰/۵۵۶۶ | ۰/۵۵۹۱ | ۰/۸۶۹۰ | ۰/۵۶۶۲ | ۰/۵۵۶۶ | ۰/۵۵۹۱ |
| Bagging | ۰/۹۰۲۰ | ۰/۷۱۵۴ | ۰/۵۷۸۰ | ۰/۶۳۸۹ | ۰/۹۰۲۰ | ۰/۷۱۵۴ | ۰/۵۷۸۰ | ۰/۶۳۸۹ |
| Gradient Boosting | ۰/۹۰۴۷ | ۰/۷۲۱۱ | ۰/۵۹۸۳ | ۰/۶۵۳۲ | ۰/۹۰۴۷ | ۰/۷۲۱۱ | ۰/۵۹۸۳ | ۰/۶۵۳۲ |
| Decision Tree | ۰/۸۶۴۹ | ۰/۵۴۹۷ | ۰/۵۶۳۰ | ۰/۵۵۵۶ | ۰/۸۶۴۹ | ۰/۵۴۹۷ | ۰/۵۶۳۰ | ۰/۵۵۵۶ |