مجله نخبگان علوم و مهندسی

Journal of Science and Engineering Elites

جلد ۸- شماره ۴ - سال ۱۴۰۲

www.ElitesJournal.ir

استفاده از الگوریتم های داده کاوی برای پیش بینی خرید مشتری

مهیار عبدالملکی

کارشناسی ارشد مهندسی کامپیوتر-نرم افزار، دانشگاه پیام نور تهران

abdolmalakmahyar@gmail.com

ارسال: مهر ماه ۱۴۰۲ پذیرش: مهر ماه ۱۴۰۲

چکیده

پیش‌بینی خرید مشتریان یکی از موضوعاتی است که در حال حاضرمورد توجه بسیاری از شرکت‌های بزرگ است. همه این شرکت‌ها دوست دارند بدانند چگونه می‌توانند رفتار مشتری‌ها را پیش‌بینی کنند. تا به حال به این موضوع فکر کرده‌اید که اگر بتوانید رفتار مشتریان خودتان را پیش‌بینی کنید چه اتفاقی خواهد افتاد؟ فکر کنید زمانی که با مشتری قصد خرید دارد بدون اینکه یک کلمه حرف بزند شما بلافاصله چیزی که مدنظرش هست را به او بدهید! در این حالت می‌دانید چه اتفاقی خواهد افتاد؟ فروش‌تان چندین برابر خواهد شد و هم چنین نیاز به صرف هزینه‌های بازاریابی هم نخواهید داشت. تکنیک‌های داده کاوی یکی از روش‌ها برای پیش‌بینی خرید مشتریان می‌باشد. متدلوژی: متدلوژی مورد استفاده در این مقاله CRISP می‌باشد و که جامعه آماری ما پایگاه داده الماس با بیش از ۴۰۰۰ رکورد و دارای پارامترهای سن، جنسیت، تحصیلات، شغل، درآمد، استان، الویت خرید، خرید اولیه، خرید نهایی، کالا، برند و.... می‌باشد که مدل سازی،که با استفاده ازنرم‌افزار رپیدماینر مدل‌ها را پیاده‌سازی می‌کنیم که مدل پیشنهادی ما در این تحقیق مدل ترکیبی با درخت تصمیم و کی- نزدیکترین همسایگی که با الگوریتم‌های RuleModel ، Naïve Bayes Random Fores, و Random Tree مقایسه می‌شود. نتایج: تجزیه و تحلیل داده‌ها جمع‌آوری شده که در این مطالعه که در انتها با بررسی دقت و صحت مدل‌ها و مقایسه آنها باهم به این نتیجه رسیدیم که مدل پیشنهادی ما یعنی مدل ترکیبی درخت تصمیم و کی نزدیک‌ترین همسایکی دارای دقت ۹۲.۱۶٪ می‌باشد.

واژگان کلیدی: داده کاوی، پیش‌بینی خرید، درخت تصمیم، K نزدیکترین همسایگی.

۱- مقدمه

با گسترش جمعیت و تامین نیازهای بشری و همچنین ترافیک‌های سنگین شهری و کاهش هزینه‌های جانبی و الکترونیکی شدن اکثر امور از جمله خرید باعث گردیده است مردم بیشتر به سمت خرید از فروشگاه و خرده‌فروشی‌های آنلاین روی بیاورند و همه شرکت‌ها و فروشگاه برآن شده‌اند که برای بهبود کسب کار و سودآوری بیشتر به فکر روش‌های باشند که نیازهای مشتریان را قبل خرید تا حد بالای تشخیص داده و بیشتر محصولات و برندهای مناسب ویژگی و خصوصیات مشتریان، در سبدفروش قراردهند از این رو پژوهشگران به فکر راهکارهای هستند که با استفاده از اطلاعات و داده‌های بدست آمده از مشتریان و استفاده از تکنیک‌های مختلف داده کاوی به مدل‌های دقیق‌تری برای پیش‌بینی خرید مشتریان دست یابند[۱]. امروزه کاربرد الگوریتم‌ها در فرایندهای روزمره به‌مراتب بیشتر و کاربردی‌تر شده است یکی از روش‌های مناسب و دقیق برای مسائل پیش‌بینی، استفاده از ترکیب الگوریتم‌هاست، در این روش با ترکیب الگوریتم‌ها و بهره‌گیری از مزایای الگوریتم‌های متفاوت، دقت بیشتری حاصل خواهد شد. تکنیک‌های داده‌کاوی یکی از بهترین گزینه‌ها برای دانش از حجم زیاد داده‌ها، کشف روابط پنهان، الگوها و تولید قوانین برای

پیش‌بینی و همبستگی داده‌ها است که می‌توان در تصمیم‌گیری سریع‌تر کمک کرد. ساختن یا حتی رسیدن به درجه بالاتر از اعتمادبه‌نفس. داده‌کاوی به معنای جستجوی الگوهای خاص در مجموعه‌های بزرگ داده‌شده است که امکانات زیادی برای کسب‌وکار و تصمیم‌گیری ایجاد می‌کند. با تجزیه‌وتحلیل تحلیل این الگوها، تصمیمات تجاری بهتری می‌توان انجام داد تا کسب‌وکارها به موفقیت مالی و کارآفرینی بیشتر دست یابند [۲].

با شناخت گروه‌های مختلف مشتریان بر اساس ویژگی‌های رفتاری آنها می‌توان به نتیجه رسید که به خرید منجر می‌شود یا خیر؟. هدف بدست آوردن الگوهای‌رفتاری مشتریان و بازدیدکنندگان در وب سایت‌ها است. استخراج شامل، خوشه‌بندی مشتری، الگوهای خرید مشتریان و پیش‌بینی خرید تولید مدل برای پیش‌بینی مشتریان است. از آنجایی که داده‌های مربوط به مشتریان بسیار غیرمستقیم، دانه‌بندی کم و حجیم هستند، از یک روش طبقه‌بندی برای دسته‌بندی استفاده می‌شود که به بررسی الگوهای کشف‌شده کمک می‌کند. این کار برای یک شرکت مفید است که باعث تصمیم‌گیری می‌شود[۳].

شرکت‌ها و سازمان‌ها به جای هدف قراردادن همه مشتریان باید آنها را بر اساس ویژگی‌های رفتاری و فردی و رفتارخرید آنها و عواملی که باعث سودآوری می‌شود مورد هدف قراردهند[۴]. پیش‌بینی، یکی از مهمترین فاکتورها برای کشف اطلاعات درباره خرید مشتریان می‌باشد. شبکه‌های عصبی مصنوعی از جمله تکنیک‌های داده کاوی می‌باشد که در سطحی بالاتر از پیش‌بینی، با درجه بالا از دقت به کار می‌روند. با استفاده از تکنیک‌های داده کاوی و شبکه‌های عصبی به تحلیل رفتار مشتریان پرداخت و اطلاعات نهان موجود در این رفتارها را شناسایی کرد. پیش‌بینی خرید مشتریان و مشتریان جدید یک شرکت ارائه دهنده خدمات، تأثیر بسزایی در روش بازاریابی و میزان سودآوری آن شرکت دارد. در این مقاله از تکنیک داده کاوی برای شناسایی محصولات مورد علاقه مشتریان بر اساس خرید آنها ارایه می‌شود و نتیجه حاصل از آن نیز مورد قرار می‌گیرد[۵].

دانستن رفتار خرید مشتری با ویژگی‌های مختلف شخصیتی و خصوصیات افراد مشخص می‌شود. امروزه داده‌کاوی معمولاً برای بررسی فعالیت‌های مشتری در خرید با استفاده از الگوریتم‌ها و روش‌های مختلف استفاده می‌شود. هر یک از فعالیت‌های یک مشتری به عنوان یک بایت از داده‌ها در یک پایگاه داده ذخیره می‌شود تا اطلاعاتی را جمع‌آوری کند، مانند اینکه مشتری چگونه وقت ارزشمند خود را صرف تصمیم‌گیری خرید می‌کند. اغلب اقلام خریداری شده و مقدار خرید نیز در نظر گرفته می‌شود. در این تحقیق از مجموعه داده برای تجزیه و تحلیل و دسته‌بندی مشتری بر اساس رفتار خرید آنها استفاده شده است. طبقه‌بندی توسط الگوریتم SVM انجام می‌شود. در این کار از مجموعه داده‌های موجودی و مجموعه داده‌های فروش موجود در اینترنت استفاده شده و عملکرد با استفاده از الگوریتم‌ها ارزیابی می‌شود. نتایج تجربی تحلیل می‌شوند و نشان می‌دهند که روش پیشنهادی پیش‌بینی رفتار خرید مشتری را تا چه حد مطلوبی تشخیص می‌دهد [۶]. در تحقیقی که به بررسی پیش‌بینی رفتار خرید مصرف کنندگان یا مشتریان پرداخته است که برای خرید محصولات باتوجه به پارامترهایی مانند: عامل‌محیطی و عامل سازمان و عامل فردی و بین فردی پرداخته شده است [۷].

۲– پیشینه تحقیق

در پژوهشی که از یازده تکنیک طبقه بندی داده‌کاوی استفاده گردیده، بسیار مطلوب بوده است. در میان روش‌های داده کاوی، الگوریتم‌های طبقه‌بندی در مطالعات برای پیش‌بینی مشتریان بالقوه شرکت موردنظر در صنعت مرتبط استفاده می‌شود. در این مطالعه معیارهای دقت، صحت و اندازه‌گیری f برای آزمایش عملکرد مدل‌های طبقه‌بندی استفاده شد. برنامه‌های داده‌کاوی مورد استفاده برای این فرآیندها R، Knime، RapidMiner و WEKA هستند و الگوریتم‌های طبقه‌بندی که معمولاً در این پلتفرم‌ها استفاده می‌شوند عبارتند از k نزدیک‌ترین همسایه، Naive Bayes و درخت تصمیم .C4.5 عملکرد الگوریتم‌های طبقه‌بندی برای پیش‌بینی رفتارهای خرید آنلاین مصرف کننده با استفاده از داده صنعت خرده فروشی آنلاین یکی از بزرگترین و سریع‌ترین صنایع در حال رشد جهان است که حجم عظیمی از داده‌های فروش آنلاین دارد. این داده‌های فروش شامل اطلاعاتی درباره تاریخچه خرید مشتری، کالاها یا خدمات ارائه شده برای مشتریان است. روابط پنهان در داده‌های فروش را می‌توان با استفاده از تکنیک‌های

داده‌کاوی کشف کرد. داده کاوی یک زمینه امیدوارکننده بین رشته‌ای است که بر دسترسی به اطلاعات مفید برای تصمیم‌گیری‌های برای کمک به فروشگاه‌های خرید آنلاین برای شناسایی رفتار مشتری آنلاین است تا محصولات مناسبی را که برای آنها جالب است را بیشتر ارایه دهند [۸].

فروشگاه‌های آنلاین می‌توانند بر اساس رفتارهای خرید و مرور آنلاین روزانه‌شان، درباره ترجیحات مصرف‌کننده بیشتر بفهمند. در این تحقیق برای پیش‌بینی قصد خرید مصرف‌کننده در طول جلسات مرور پیشنهاد شده است. با استفاده از تکنیک‌های داده کاوی انجام گرفته نشان داده است که روش پیشنهادی توانایی پیش‌بینی قوی خود را در مقایسه با سایر مدل‌های سنتی نشان می‌دهد[۹]. ترکیب الگوریتم‌ها و بهره‌گیری از مزایای الگوریتم‌های متفاوت، بهبود بیشتری حاصل خواهد شد که پژوهش‌هایی در این زمینه انجام گرفته است. امروز شرکت‌های بزرگ و فروشگاه‌ها بسیار مشتاق هستند تا در مورد مشتریان خود با استفاده از فناوری‌های داده کاوی اطلاعات کسب کنند. اما موقعیت‌های متنوع چنین شرکت‌هایی، شناخت مؤثرترین الگوریتم برای مسائل داده شده را دشوار می‌سازد. اخیراً، حرکتی به سمت ترکیب طبقه‌بندی‌کننده‌های متعدد برای بهبود نتایج طبقه‌بندی پدید آمده است. در این پژوهش روشی را برای پیش‌بینی رفتار خرید مشتری EC با ترکیب چند الگوریتم طبقه‌بندی این روش با استفاده از داده‌های وب از یک شرکت پیشرو EC آزمایش و ارزیابی شد. همچنین اعتبار رویکرد خود را در مسائل طبقه‌بندی کلی با استفاده از اعداد دست‌نویس آزمایش کرده است. در هر دو مورد، روش الگوریتم ترکیبی عملکرد بهتری نسبت به طبقه‌بندی‌کننده‌های منفرد نشان می‌دهد[۱۰].

خرده‌فروشی آنلاین رشد سریعی داشته است و وب‌سایت‌ها سرشار از داده‌های رفتار کاربر هستند. رفتارهای عملیاتی کاربران در پلت‌فرم تجارت الکترونیکی می‌تواند منعکس کننده ترجیحات کاربر باشد. نحوه استفاده از رفتارهای کاربر برای استخراج ترجیحات کاربر به کانون دانشگاه و صنعت تبدیل شده است و نتایج تحقیقات زیادی به دست آمده است. در بسیاری از موارد، با آموزش ترکیبی دو یا چند الگوریتم مختلف، توانایی تعمیم الگوریتم را می‌توان به طور قابل توجهی بهبود بخشید تا اثر پیش‌بینی بهبود یابد. این مطالعه ترکیبی از رگرسیون لجستیک و الگوریتم‌های ماشین بردار پشتیبان را برای ساخت یک مدل پیش‌بینی ترکیبی برای رفتار خرید مشتریان ترکیب می‌کند و در این مطالعه نتایج تجربی نشان می‌دهد که مدل ترکیبی اثر پیش‌بینی بهتری نسبت به مدل واحد دارد.[۱۱]. پژوهش‌های زیادی جهت اریابی دقت الگوریتم‌های داده کاوی انجام گرفته است که در زیر به تشریح مختصر آنها خواهیم پرداخت:

در پژوهشی که جهت پیش‌بینی رفتار خرید آنلاین مشتریان انجام گرفته است، با هدف ارائه پیشنهادی برای افزایش دقت و تحلیل و شناخت گروه‌های مشتریان و همچنین ارائه مدل و قوانین برای پیش‌بینی رفتار مشتری است. بنابراین از الگوریتم CRISP-DM و K-means برای خوشه‌بندی داده‌ها استفاده می‌شود. در نهایت مدلی با دقت ۶۳.۶٪ را به دست آورده است[۱۲].

تحقیقی دیگر که یک مدل پیش‌بینی را ایجاد کرده است که بر اساس ویژگی‌های رفتار خرید مصرف‌کنندگان است. ویژگی‌های رفتار خرید آنلاین مانند ایمنی تراکنش، در دسترس بودن محصولات نوآورانه و کیفیت محصولات. پنج طبقه‌بندی‌کننده منفرد و مجموعه‌های آنها با Bagging و Boosting بر روی مجموعه داده جمع‌آوری‌شده از یک سایت خرید آنلاین بررسی می‌شوند. نتایج نشان می‌دهد که مدل ساخته شده با استفاده از مجموعه درخت تصمیم با Bagging بهترین پیش‌بینی رفتار مصرف کننده را با دقت ۹۵.۳٪ به دست آورده است[۱۳].

در مطالعه‌ی با تجزیه و تحلیل داده ها و اطلاعات مشتریان داده‌های تجربی خریداران آنلاین را برای ساختن یک مدل پیش‌بینی بهتر برای پیش‌بینی قصد خرید آن‌ها تحلیل کرده است. الگوریتم‌های طبقه‌بندی مختلفی مانند درخت تصمیم، جنگل تصادفی، ساده بیز، SVM را تجزیه و تحلیل کرده‌است تا پیش‌بینی کند که آیا مشتری با بازدید از صفحات وب یک فروشگاه آنلاین، به خرید ختم می‌شود و تحقیقات نشان داد که جنگل تصادفی برای پیش‌بینی قصد خرید مشتری مناسب‌تر است. می‌تواند با بالاترین دقت پیش‌بینی کند که ۹۰.۳۴٪ است[۱۴]. با بهبود سیستم‌های تراکنش آنلاین و بسترهای خرید آنلاین، مشتریان بیشتری خرید آنلاین را انتخاب می‌کنند. با این حال، از آنجایی که مشتریان و بازرگانان نمی‌توانند رو در رو با هم ارتباط برقرار کنند، بازرگانان اطلاعات بسیار

کمی در مورد نیازهای مشتریان خود دارند و نمی‌توانند افکار آنها را به موقع درک کنند. سیستم آنلاین عملیات مصرف‌کننده را ثبت می‌کند و در پژوهشی که داده‌های رفتار مصرف‌کننده را جمع‌آوری می‌کند و امکان پیش‌بینی ترجیحات خرید مصرف‌کنندگان را فراهم می‌کند. داده‌های خرید عدم خرید واقعی پلت فرم تجارت الکترونیک را به عنوان هدف تحقیق در نظر می‌گیرد و از مدل catboostبرای تجزیه و تحلیل و پیش‌بینی اینکه آیا مصرف‌کنندگان محصول خاصی را خریداری خواهند کرد یا خیر، استفاده می‌کند. دقت، صحت و برخی معیارهای دیگر مدل برای ارزیابی عملکرد پیش‌بینی ارائه شده است. اثر بهتری به دست می‌آید. دقت در پیش‌بینی رفتار خرید در این داده‌ها به ۸۸.۵۱٪ می‌رسد[۱۵].

در پژوهشی از الگوریتم‌های طبقه‌بندی‌های برای پیش‌بینی خرید استفاده می‌شوند با استفاده از ماشین‌های رگرسیون لجستیک، بیز ساده و ماشین‌های بردار پشتیبان که الگوریتم‌های طبقه‌بندی هستند، مدل‌هایی با بهترین دقت مقایسه می‌شوند. داده‌های رفتار آماری فیلتر و پردازش می‌شوند. الگوریتم شبکه عصبی بازگشتی (RNN) برای طبقه‌بندی استفاده و نحوه رفتار کاربر به دست می‌باشد. سپس امتیاز به عنوان یک ویژگی جدید در نظر گرفته می‌شود نتایج به دست آمده با استفاده از الگوریتم‌های مطرح شده مقایسه شده است. نتایج آزمون نشان می‌دهد که روش پیشنهادی دارای دقت بالاتری است، دقت پیش‌بینی نیز در مقایسه با یک مدل بیزی ساده بهبود یافته است، که عملکرد بهتری دارد[۱۶].

در تحقیقی که به منظور پیش‌بینی خرید آنلاین از وب‌سایت، قصد خرید بازدیدکننده را پیش‌بینی می‌کند. برای انجام این کار، به اطلاعات فردی مشتریان و بازدیدکننده تکیه می‌کند و طبقه‌بندی‌کننده ساده بیز، درخت تصمیم C4.5 و جنگل تصادفی را بررسی می‌کنیم. نتایج نشان می‌دهد که جنگل تصادفی به طور قابل توجهی دقت بالاتری نسبت به تکنیک‌های دیگر دارد[۱۷].

در پژوهشی دیگر با هدف پیش‌بینی خرید مشتری و اینکه آیا خرید می‌کند یا خیر؟ بنابراین تکنیکی برای پیش‌بینی خرید می‌سازد به همین دلیل است که برخی از الگوریتم‌های شناخته شده را برای دستیابی به دقت بهتر برای خرید در مقاله خود پیشنهاد می‌کند. آن الگوریتم‌ها را در مجموعه داده خود که شامل ۵۰ داده است، اعمال می‌کند. در میان آنها، ماشین بردار پشتیبانی (SVM) بهترین نتیجه را با دقت ۸۶.۷٪ پیش‌بینی کرد. در این مقاله، همچنین نتایج مقایسه‌ای را با استفاده از الگوریتم‌های مختلف دقت، Recall و امتیاز F1 برای همه نمونه‌های داده نشان می‌دهد[۱۸]. تحقیقی دیگرکه با استفاده از الگوریتم‌های طبقه‌بندی بیز ساده، Apiorir، Decision Tree وRandom Forest برای پیش‌بینی خرید و علاقه اینترنتی مشتریان انجام شد. تحقیق انجام شده روی مجموعه داده این پژوهش بالاترین دقت را الگوریتم Apiorir با دقت ۸۸٪ و نایوبیز با دقت ۸۷٪ را نشان داد[۱۹].

۳- روش بررسی

روش انجام این پژوهش در دو مرحله صورت می‌پذیرد، مرحله اول مبتنی بر مطالعه گزارش‌های فنی و پژوهش علمی، مقالات، پایان‌نامه‌ها، کتاب‌ها و پروژه‌های تحقیقاتی و مطالعات موجود در سایت‌های اینترنتی در زمینه پیش‌بینی و رفتار خرید مشتریان با تکنیک‌های مختلف می‌باشد. روش ما در مرحله دوم روش آزمایشگاهی است. به این صورت که از الگوریتم DT و الگوریتم KNN و برای پیش‌بینی خرید مشتریان استفاده خواهد شد برای این منظور ترکیب الگوریتم پیشنهادی با استفاده از نرم‌افزار رپیدماینر پیاده‌سازی می‌گردد. سپس هردو الگوریتم بر روی داده‌هایی الماس که حاوی نمونه‌های استاندارد در زمینه خریدهای اینترنتی است مورد اجرا قرارگرفته و نتیجه آن ضمن مقایسه میان کارایی الگوریتم ترکیبی پیشنهادی در این مقاله با الگوریتم‌های دیگر ارائه‌شده بررسی گردیده و درپایان به جمع‌بندی نهایی و تحلیل نتایج خواهیم پرداخت.

۳-۱- الگوریتمKNN

الگوریتم K Nearest Neighbor در گروه یادگیری تحت نظارت قرار می‌گیرد و برای طبقه‌بندی (رایج‌ترین) و رگرسیون استفاده می‌شود. این یک الگوریتم همه کاره است. الگوریتم K نزدیکترین همسایه (KNN) یک الگوریتم یادگیری ماشین نظارت شده است و از ویژگی‌های آن سادگی و آسانی این الگوریتم برای پیاده‌سازی است که می‌تواند برای حل مسائل طبقه‌بندی و رگرسیون مورد استفاده قرار گیرد. الگوریتم نزدیک‌ترین همسایگی کاربرد فراوانی در داده کاوی دارد و یک الگوریتم بسیار محبوب در این

زمینه است. یکی از دلایل اصلی پرکاربرد بودن الگوریتم‌های طبقه‌بندی آن است که «تصمیم‌گیری» یکی از چالش‌های اساسی موجود در اغلب پروژه‌های تحلیلی است.

۳-۲- الگوریتم درخت تصمیم(Decision Tree)

الگوریتم درخت تصمیم یکی از تکنیک‌های پرکاربرد در داده‌کاوی، سیستم‌هایی است که طبقه‌بندی کننده‌ها را ایجاد می‌کنند در داده‌کاوی، الگوریتم‌های طبقه‌بندی قادر به مدیریت حجم وسیعی از اطلاعات هستند. می‌توان از آن برای ایجاد مفروضات در مورد نام‌های طبقه‌بندی‌شده، طبقه‌بندی دانش بر اساس مجموعه‌های آموزشی و برچسب‌های کلاس و طبقه‌بندی داده‌های تازه به‌دست‌آمده استفاده کرد[۱۳]. الگوریتم درخت تصمیم یکی از تکنیک‌های پرکاربرد در داده‌کاوی، سیستم‌هایی است که طبقه‌بندی‌کننده‌ها را ایجاد می‌کنند. درخت تصمیم یکی از روش‌های قدرتمندی است که معمولاً در زمینه‌های مختلف مانند یادگیری ماشین، پردازش تصویر و شناسایی الگوها استفاده می‌شود. علاوه بر این DT یک مدل طبقه‌بندی معمولاً مورد استفاده در داده کاوی است دهد و هر زیر مجموعه مقداری را تعریف می‌کند که می‌تواند توسط گره گرفته شود. درختان تصمیم به دلیل تجزیه و تحلیل ساده و دقت آنها بر روی فرم‌های داده‌های متعدد، زمینه‌های پیاده سازی بسیاری را پیدا کرده‌اند.

۳-۳- الگوریتم ترکیبی درخت تصمیم وKNN

الگوریتم ترکیبی درخت تصمیم و KNN با استفاده از عملگر vot که از عملگرهای نرم‌افزار رپیدماینر می‌باشد انجام می‌گیرد و این عملگر برای دسته‌بندی از رای اکثریت استفاده می‌کند. در این تحقیق دیتاست شامل چهار هزار رکورد می‌باشد و پارامترهای مورد استفاده در این تحقیق مانند سن، جنس، حقوق، تحصیلات و ... مربوط به پایگاه داده الماس می‌باشد و برای حداکثر استفاده این صفات و پارامترها را بصورت عددی تبدیل کرده‌ایم. جهت ارزیابی معیارها روش‌های مختلفی وجود دارد که در این پژوهش برای ارزیابی معیارها از ماتریس درهم ریختگی (Confusion Matrix) استفاده شده است.

- معیارهای ارزیابی شده
  - Accuracy (دقت)
  - Precision (صحت)
  - Recall (بازخوانی)
  - F Score (حاصل از میانگین هارمونیک دقت و بازخوانی)
- ماتریس درهم ریختگی

TP: نشان دهند تعداد رکوردهایی است که دسته واقعی آن‌ها مثبت بوده و الگوریتم دسته‌بندی نیز دسته آنها را به‌درستی مثبت تشخیص داده است. (پیش‌بینی بله است و آنها خرید داشته‌اند).

TN: نشان دهنده تعداد رکوردهایی است که دسته واقعی آن‌ها منفی بوده و الگوریتم دسته‌بندی نیز دسته آنها را به‌درستی منفی تشخیص داده است. (پیش‌بینی منفی است و آنها خرید نکرده‌اند).

FP: نشان دهنده تعداد رکوردهایی است که دسته واقعی آن‌ها منفی بوده و الگوریتم دسته‌بندی دسته آن‌ها را به اشتباه مثبت تشخیص داده است. (ما پیش‌بینی کردیم آنها خرید داشته‌اند، اما در واقع خرید نکرده‌اند. (همچنین به عنوان "خطای نوع اول" شناخته می‌شود).

FN: نشان دهنده تعداد رکوردهایی است که دسته واقعی آن‌ها مثبت بوده و الگوریتم دسته‌بندی دسته آن‌ها را به اشتباه منفی تشخیص داده است. (ما پیش‌بینی کردیم مشتریان خرید نداشتند، اما در واقع آنها خرید کرده اند. (همچنین به عنوان "خطای نوع دوم" شناخته می‌شود).

جدول ۱- ماتریس درهم ریختگی

پیش‌بینی			
واقعی	**Positive**	**Negative**	**مجموع**
Positive	TP	FN	P=TP+FN
Negative	FP	TN	N=FP+TN
Total	TP+FP	FN+TN	

جدول ۲- نحوه محاسبه معیارهای ارزیابی

Evaluation Methods	**Equations**
Accuracy	$\frac{(TP+TN)}{(TP+FN+FP+TN)}$
Precision	$\frac{TP}{(TP+FP)}$
Recall	$\frac{TP}{(TP+FN)}$
F Score	$\frac{2\times(Precision\times Recall)}{(Precision+Recall)}$

از مقادیر ماتریس درهم ریختگی جهت ارزیابی دسته‌بند استفاده می‌شود. جدول ۲ نحوه محاسبه معیارهای ارزیابی را براساس ماتریس در هم ریختگی نشان می‌دهد. یکی از مهمترین معیارها از بین معیارهای استفاده شده برای کارایی الگوریتم معیار دقت یا نرخ تشخیص است که میزان پیش‌بینی صحیح به کل نمونه‌ها را نشان می‌دهد.

۴- مراحل انجام پژوهش

مراحل انجام تحقیق بصورت استاندارد (CRISP: Cross-Industry Standard Process) به روش زیر می‌باشد:

۴-۱- مرحله اول: جمع‌آوری و پیش پردازش داده‌ها

جمع‌آوری داده از طریق پرسشنامه انجام شده پرسشنامه به روش الکترونیکی بود و پاسخ‌های ثبت شده در پرسشنامه‌های اسکن شده توسط رایانه خوانده شده در این گام به جمع‌آوری داده‌های اولیه، توصیف داده‌ها، بازرسی و بررسی داده‌ها پرداخته‌شده است.

۴-۲-آماده‌سازی داده

در ابتدا برای جمع و آماده‌سازی داده‌ها از کوئری‌های Select، Where، Top و Distinct کوئری‌های Join کردن در جداولی مانند Inner Join و ساخت View در نرم‌افزار SQL، استفاده گردید. نرم‌افزار رپیدماینر مجهز به ابزارهای بسیار قوی است تا بتواند مجموعه داده را در پایگاه داده داخلی یا محلی نرم‌افزار بارگذاری نموده و این مجموعه داده را برای ارائه به عملگرهای یادگیری مدل آماده کند.

۴-۳- مرحله دوم: مدل‌سازی

در مدل سازی روش‌های داده کاوی زیادی وجود دارد. در این مرحله تکنیک‌های مختلف داده‌کاوی به رسم مدل و الگوی بهبود یافته می‌پردازیم.

۴-۴- مرحله سوم: نتایج

در این مرحله پیش‌بینی می‌گردد که دقت هر مدل چند درصد می‌باشد.

۴-۵- مرحله چهارم: ارزیابی

برای رسیدن به نتیجه و هدف در این مرحله مدل ارزیابی می‌شود تا ببینیم آیا به هدف رسیده‌ایم یا نه؟ قسمت‌هایی که نتیجه بخش نبوده و به هدف نرسیده را تکرار می‌کنیم یا بعضی مواقع ممکن است به تغییرهدف تبدیل شود و یا مجبور به تغییر اعداد اولیه شود.

۴-۶- مرحله پنجم: توسعه

پایان یک پروژه ساخت مدل نیست و هدف از کشف دانش و استفاده از این دانش کشف‌شده در آینده است.

۴-۷- مدلسازی با استفاده از الگوریتم درخت تصمیم و K نزدیکترین همسایه

در این مدل‌سازی پارامترهای مختلف با حالات و مقادیر مختلف مورد بررسی قرار گرفت آزمایش و همچنین در وضعیت عدم هرس و هرس کردن، که بهترین و بالاترین دقت بدست آمده از مدل‌سازی با الگوریتم درخت تصمیم و_K نزدیکترین همسایه با پارامترهای ذکر شده در جدول ۴ نشان داده شده است.

جدول۳- پارامترهای استفاده شده در مدل‌سازی با عملگرها

عملگر	پارامتر	مقدار/حالت	دقت مدل
k-NN	مقدارK	۹	٪۹۲.۱۶
Decision Tree	معیار برش[1]	Gain_ratio	
	حداکثر عمق	۲۰	
	نسبت نمونه آزمایش	0.25	
	هرس	وجود هرس	

۵- ارزیابی داده‌ها

جدول۴- ماتریس درهم ریختگی ارزیابی کل داده‌ها

`Random Tree			
مجموع	عدم خرید	خرید	T/P
۳۹۸۴	۷۱۱	۳۲۷۳	خرید
۳۱۴	۲۰۷	۱۰۷	عدم خرید
۴۲۹۸	۹۱۸	۳۳۸۰	مجموع
Random Forest			
مجموع	عدم خرید	خرید	T/P
۴۲۷۴	۸۹۵	۳۳۷۹	خرید
۲۳	۲۲	۱	عدم خرید
۴۲۹۸	۹۱۸	۳۳۸۰	
Naive Bayes			
مجموع	عدم خرید	خرید	T/P
۳۳۶۱	۱۸۱	۳۱۸۰	خرید
۹۳۷	۷۳۷	۲۰۰	عدم خرید
۴۲۹۸	۹۱۸	۳۳۸۰	
Rule Induction			
مجموع	خرید نمی‌کند	خرید می‌کند	T/P
۳۷۵۶	۴۷۵	۳۲۸۱	خرید
۵۴۲	۴۴۳	۹۹	عدم خرید
۴۲۹۸	۹۱۸	۳۳۸۰	

[1] Criterion

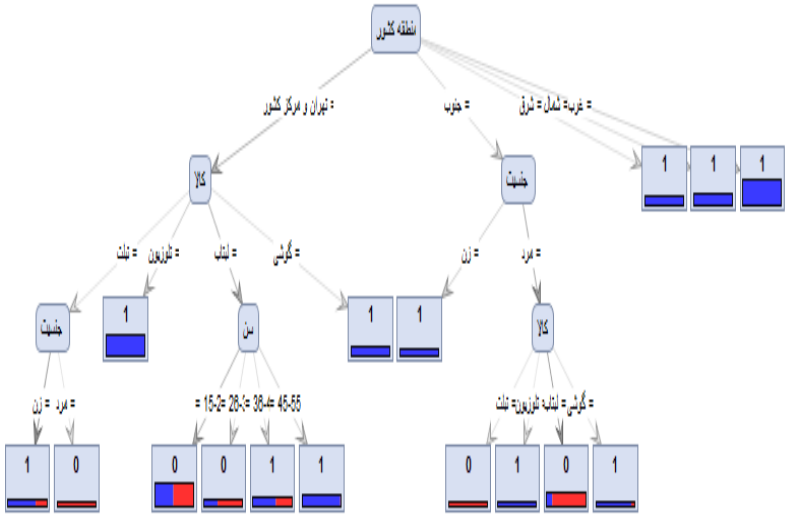
جدول ۵ -پارامترهای عملگر درخت تصمیم و k –NN

عملگر	پارامتر	مقدار/حالت	دقت مدل
k-NN	مقدارK	۹	٪۹۲.۱۶
Decision Tree	معیار برش[1]	Gain_ratio	
	حداکثر عمق	۲۰	
	نسبت نمونه آزمایش	0.25	
	هرس	وجود هرس	

جدول ۶- دقت مدل با استفاده از عملگرهای درخت تصمیم و k -NN

T/P	خرید	عدم خرید	مجموع
خرید	۳۳۰۹	۲۶۶	۳۵۷۵
عدم خرید	۷۱	۶۵۲	۷۲۳
	۳۳۸۰	۹۱۸	۴۲۹۸

از نتایج جدول (۶) چنین استنتاج می‌شود که در دسته واقعی ۱، ۳۳۸۰ مورد از مشاوره‌های انجام شده به خریداران بوده که در دسته ۱ (که در نهایت این مشتریان خرید را انجام داده‌اند قرارگرفته‌اند) را تشخیص داده و مدل نیز ۳۳۰۹ تعداد از دسته آن‌ها را بدرستی و با دقت ۹۸٪ و نزدیک به ۱۰۰ درصد تشخیص داده است و در دسته واقعی (۰) مدل ۹۱۸ مورد از از مشاوره های انجام شده به خریداران بوده که در دسته (۰) (که در نهایت این مشتریان خرید را انجام نداده‌اند قرارگرفته‌اند) را تشخیص داده و مدل نیز ۶۵۲ نفر از دسته آن‌ها را بدرستی و با دقت ۷۱ درصد تشخیص داده است. در قسمت پیش‌بینی دسته ۱ مدل ۳۵۷۵ از مشاوره‌های انجام شده به خریداران بوده که در دسته ۱ (که در نهایت این مشتریان خرید را انجام داده‌اند قرارگرفته‌اند) را مدل پیش‌بینی کرده و مدل نیز ۳۳۰۹ نفر از دسته آن‌ها را بدرستی و با دقت ۹۳٪ پیش‌بینی کرده است و در قسمت پیش‌بینی دسته (۰) مدل از ۷۲۳ مشاوره‌های انجام شده به خریداران بوده که در دسته (۰) (که در نهایت این مشتریان خرید را انجام نداده‌اند قرارگرفته‌اند) را مدل پیش‌بینی کرده و مدل نیز ۶۵۲ نفر از دسته آن‌ها را بدرستی و با دقت ۹۰ درصد پیش‌بینی کرده است. در شکل زیر درخت تصمیم طراحی شده توسط مدل فوق، آمده است:



شکل۱- درخت تصمیم طراحی شده در نرم‌افزار Rapid Miner

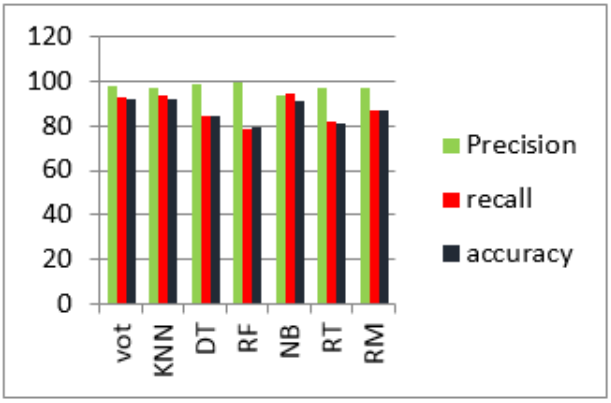
پس از نتیجه‌گیری از روی شکل مدل و از نتایج شکل ۱ چنین استنتاج می‌شود که منطقه کشور، کالا، جنسیت و سن از مهمترین عوامل تاثیرگذار در خرید است.

[1] Criterion

جدول۷- معیارهای ارزیابی الگوریتم‌ها

الگوریتم‌های مورد استفاده	Precision	Recall	Accuracy	F measur
Vot (KNN Decicsion Tree)	97.8%	92.5%	92.1%	95%
Decicson Tree	99.11%	84.6%	85.1%	91.2%
KNN	96.8%	93.6%	92.4%	95.2%
RuleModel	97%	87.3%	86.6%	91.8%
Naïve Bayes	94%	94.6%	91.1%	94.1%
Random Tree	96.8%	82.1%	80.9%	90.4%
Random Fores	99.9%	79%	79.1%	88.2%

معیارهای ارزیابی استفاده شده در جدول۷ نشان داده شده است مطابق نتایج بدست آمده مدل ترکیبی برای پیش‌بینی دارای دقت ۹۲.۱ درصد می‌باشد.



شکل۲- نمودار دقت و صحت الگوریتم‌ها

نمودار معیارهای ارزیابی هرکدام از الگوریتم‌ها را از صفر تا صد نشان می‌دهد نمودارهای رنگ مشکی نماینگر دقت پیش‌بینی هر کدام از مدل‌ها می‌باشد نمودارهای رنگ قرمز نشان دهنده پوشش می‌باشد و نمودارهای به رنگ سبز نشان دهنده معیار صحت هرکدام از مدل‌ها می‌باشد.

۶- نتیجه‌گیری

الگوریتم‌های مختلف و متعددی درخصوص طبقه‌بندی و با هدف پیش‌بینی موجود است ولی آنچه که انتخاب الگوریتم مناسب را سخت کرده است این موضوع است که این الگوریتم‌ها به ساختار داده‌ها وابسته می‌باشند به همین منظور به الگوریتم و ترکیب الگوریتم ارائه‌شده در این زمینه روی آوردند. همچنین با توجه به این‌که تجزیه‌وتحلیل خرید اینترنتی محسوب می‌شود که در شرکت‌ها و فروشگاه یک دیدگاه جدید شکل بگیردکه در این دیدگاه جدید ارتباطات میان افراد و شرکت‌ها بهبود یابد که سبب سودآوری برای شرکت‌ها و فروشگاه‌ها می‌گردد. بر این اساس در این پژوهش از الگوریتم‌های طبقه‌بندی درخت تصمیم و K نزدیکترین همسایگی استفاده شد. سعی شد تا نسبت به برطرف کردن با رفع نمودن چالش‌های مطرح شده بتواند با دقت بیشتری کار کند و اطمینان بیشتری به سود و حفظ و پایداری خود داشته باشند استفاده از الگوریتم ترکیبی درخت تصمیم و K نزدیکترین همسایگی می‌توان پیش‌بینی بهتری برای خرید اینترنتی مشتریان داشت.

۷-مراجع

1. Lalwani P, Mishra MK, Chadha JS. and Sethi P. (2022). Customer churn prediction system: a machine learning approach. Computing, Vol. 104, No. 2, pp. 271-294.

2. Meiyazhagan J, Sudharsan S, Venkatesan A. and Senthilvelan M. (2022). Prediction of occurrence of extreme events using machine learning. The European Physical Journal Plus, Vol. 137, No. 1, pp. 1-20.

3. Mou AD, Saha PK, Nisher SA. and Saha A. (2021). A comprehensive study of machine learning algorithms for predicting car purchase based on customers demands. In 2021 International Conference on Information and Communication Technology for Sustainable Development (ICICT4SD) (pp. 180-184). February 2021. IEEE.

4. Moon NN, Talha IM. and Salehin I. (2021). An advanced intelligence system in customer online shopping behavior and satisfaction analysis. Current Research in Behavioral Sciences, Vol. 2, 100051, November 2021

5. MCA MNZJ. and Gokul M.(2021). Implementation of Customer Purchase Prediction using. International Journal of Computer Techniques, Vol. 8, No. 2, pp. 144-149, March 2021.

6. Qalati SA, Vela EG, Li W, Dakhan SA, Hong Thuy TT. and Merani SH. (2021). Effects of perceived service quality, website quality, and reputation on purchase intention: The mediating and moderating roles of trust and perceived risk in online shopping. Cogent Business & Management, Vol. 8, No. 1, pp. 1869363.

۷. ملائی منیژه و پارسا سودابه .پیش‌بینی رفتار مشتریان با استفاده از تکنیک شبکه‌های عصبی مصنوعی، **۱۳۹۵**.

8. Charbuty B. and Abdulazeez A. (2021). Classification based on decision tree algorithm for machine learning. Journal of Applied Science and Technology Trends, Vol. 02, No. 01, pp. 20-28. Marc 2021.

9. Chang HH. and Meyerhoefer CD. (2021). COVID-19 and the demand for online food shopping services: Empirical Evidence from Taiwan. American Journal of Agricultural Economics, Vol. 103, No. 2, pp. 448-465.

10. Dou X. (2020). Online purchase behavior prediction and analysis using ensemble learning. In 2020 IEEE 5th International conference on cloud computing and big data analytics (ICCCBDA) (pp. 532-536). April 2020. IEEE.

11. Javadi MHM, Dolatabadi HR, Nourbakhsh M, Poursaeedi A. and Asadollahi

12. Hu X, Yang Y, Zhu S. and Chen L. (2020). Research on a hybrid prediction model for purchase behavior based on logistic regression and support vector machine. In 2020 3rd International conference on artificial intelligence and big data (ICAIBD) (pp. 200-204). May 2020. IEEE.

13. Mou AD, Saha PK, Nisher SA. and Saha A. (2021). A comprehensive study of machine learning algorithms for predicting car purchase based on customers demands. In 2021 International Conference on Information and Communication Technology for Sustainable Development (ICICT4SD) (pp. 180-184). February 2021. IEEE.

14. Moon NN, Talha IM. and Salehin I. (2021). An advanced intelligence system in customer online shopping behavior and satisfaction analysis. Current Research in Behavioral Sciences, Vol. 2, 100051, November 2021.

15. MCA MNZJ. and Gokul M.(2021). Implementation of Customer Purchase Prediction using. International Journal of Computer Techniques, Vol. 8, No. 2, pp. 144-149, March 2021.

16. Qayyum R, Rubaab J, Riaz U. and Arif F. (2021). Role of Data Mining and Machine Learning in Software Reusability. In 2021 International Conference on Innovative Computing (ICIC) (pp. 1-8). November 2021. IEEE.

17. Rita P, Oliveira T. and Farisa A. (2019). The impact of e-service quality and customer satisfaction on customer behavior in online shopping. Heliyon, Vol. 5, No. 10, pp. e02690.

18. Sharma P, Meena U. and Sharma GK. (2021). Application of data mining algorithms for tourism industry. In Intelligent Computing and Applications (pp. 481-495). Springer, Singapore.

19. Saura JR. (2021). Using data sciences in digital marketing: Framework, methods, and performance metrics. Journal of Innovation & Knowledge, Vol 9, No 2, pp. 92-102, April–June 2021.