نشریــه علمـــی دانش مالی تحلیل اوراق بهادار
تابستان ۱۴۰۴/دوره ۱۸ / شماره ۶۶
صفحه ۲۳۰ تا ۲۴۳

تشخیص تراکنش‌های مشکوک در کارت‌های اعتباری با استفاده از الگوریتم *Extra Trees Classifier*

عرفان عابدینی[۱]
محمود همت فر[۲*]

تاریخ دریافت: ۱۴۰۳/۱۲/۱۵ تاریخ پذیرش: ۱۴۰۳/۰۵/۲۳

چکیده

در سال‌های گذشته، با توسعه تجارت الکترونیک، کارت‌های اعتباری به طور گسترده تری در تراکنش‌ها به دلایل مختلف از جمله سهولت کارایی مورد استفاده قرار گرفته اند. تقلب و تراکنش های غیرمجاز کارت های اعتباری به دلیل افزایش استفاده از آنها در حال تبدیل شدن به مسئله‌ای رایج و مهم است. بنابراین، مسئولیت بزرگی بر عهده مؤسسات مالی است که مکانیزم موجود خود را به‌روز کنند تا از این اقدامات غیرمجاز جلوگیری کنند. عدم وجود یک الگوی مشخص و همچنین عدم توازن در تعداد تراکنش های مجاز و غیرمجاز، توسعه مدل و ساخت سیستم های تشخیص تقلب را با مشکل مواجه کرده است. هدف اصلی این مطالعه، ایجاد یک راهکار مناسب، دقیق و سریع برای تشخیص تقلب در کارت های اعتباری مبتنی بر یادگیری ماشین است. مدل پیشنهادی در این مقاله، شامل استفاده از روش *SMOTE* برای مدیریت داده‌های نامتوزان و الگوریتم *Extra Trees Classifier* برای شناسایی تراکنش غیرمجاز است. همچنین معیارهای مختلف بررسی دقت و کارایی نشان از عملکرد موثر مدل پیشنهادی میباشد.

کلمات کلیدی: یادگیری ماشین، یادگیری ماشین در تشخیص تقلب، تقلب کارت اعتباری، تقلب آنلاین، کلاهبرداری کارت اعتباری.

[1] دانشجوی کارشناسی ارشد، گروه مدیریت مالی، دانشکده مدیریت و اقتصاد، واحد علوم و تحقیقات، دانشگاه آزاد اسلامی، تهران، ایران
erfan.abedini@srbiau.ac.ir

[۲] دانشیار، گروه حسابداری، واحد بروجرد، دانشگاه آزاد اسلامی، بروجرد، ایران (نویسنده مسئول)*dr.hematfar@yahoo.com**

۱. مقدمه

کلاه‌برداری یعنی فریب دادن کسی به گونه‌ای که عمداً اموال، پول یا هر گونه حقوق قانونی خود را به نفع شخص غیرقانونی واگذار کند که منجر به کسب سود مالی یا شخصی غیرمجاز شود. انواع مختلفی از کلاهبرداری‌ها در زندگی روزمره رخ می‌دهد که از جمله آن‌ها می‌توان به تقلب در کارت‌های اعتباری اشاره کرد. تقلب در تراکنش‌های کارت اعتباری می‌تواند با استفاده از جزئیات کارت توسط فرد غیرمجاز و انجام تراکنش‌های سودمند برای وی رخ دهد. تقلب در کارت اعتباری به یکی از مشکلات اصلی در دنیای امروزی تبدیل شده است. کلاهبرداری‌های از کارت‌های اعتباری رایج‌ترین انواع کلاهبرداری در دنیای بانکداری است و احتمال کلاهبرداری از کارت اعتباری به دلیل استفاده گسترده توسط عموم مردم بالاست. همزمان با تکامل فناوری و ظهور گزینه‌های جدید پرداخت الکترونیکی مانند تجارت الکترونیک و پرداخت‌های موبایلی، معاملات کارت اعتباری محبوبیت بیشتری پیدا کرده است *(Kotsiantis et al., 2006, p. 25-36)*.

از اوایل ماه مارس ۲۰۲۰ میلادی، بسیاری از کشورها تصمیم گرفتند برای کاهش گسترش ویروس کووید ۱۹، قرنطینه اعمال کنند. مغازه‌ها موقتاً تعطیل شدند، اما خرید اینترنتی ادامه پیدا کرد و تعداد تراکنش‌های خرید آنلاین به شدت افزایش یافت. آمازون در وبسایت خود اعلام کرد که طی بحران، خریدهای مشتریان بی‌سابقه افزایش یافته است *(Bhutani, 2021, pp. 649-659)*

(Abdelrhim & Elsayed, 2020). پاندمی کووید ۱۹ همه را مجبور کرد با شرایط سازگار شوند و به سمت خرید آنلاین تغییر جهت دهند که این امر سلامت همه را طی همه‌گیری حفظ کرد. از دوره پاندمی، خرید آنلاین به مناسب‌ترین و قابل اعتمادترین روش خرید تبدیل شده است، زیرا تعاملات فیزیکی را محدود می‌کند و سلامت مشتریان را حفظ می‌کند. طبق یک نظرسنجی در ۹ کشور شامل حدود ۳۷۰۰ مصرف‌کننده انجام شده است که نشان می‌دهد بحران به طور دائم الگوهای خرید آنلاین را تغییر داده است و نتیجه گرفته می‌شود که خرید سنتی پس از کووید ۱۹ هرگز مثل گذشته نخواهد بود و سهم بازرگانی الکترونیک را در کل خرده فروشی افزایش داده است *(UNCTAD, 2020)* *(Guthrie, 2021, p. 102570)*. با استقبال گسترده از این روش، کلاهبرداران بیشتر تمایل به انجام کلاهبرداری دارند *(Meng et al., 2020, p. 052016)*. کلاهبرداران به طور مداوم تاکتیک‌های خود را تغییر می‌دهند تا از دستگیر شدن اجتناب کنند. از آنجا که هر سیستم بانکی دارای نقاط ضعف است *(Kramer, 2011, p. 979)*، امن کردن یک سیستم برای احراز هویت و جلوگیری از کلاهبرداری مشتریان کار دشواری می‌شود. این عملیات کلاهبردارانه می‌توانند بررسی و شناسایی شوند تا از دزدی کارت‌های اعتباری در آینده جلوگیری و آن را کاهش دهند *(Menardi & Torelli, 2014, pp. 92-122)*.

کلاهبرداری از کارت های اعتباری می‌تواند به یکی از اشکال زیر باشد *(Vejalla et al., 2023)*:

۱ – **سرقت ساده:** کارت دزدیده شده

۲ – **کلاهبرداری در درخواست:** افراد با ارائه اطلاعات شخصی غلط به ارائه دهندگان کارت، کارت‌های اعتباری جدید کسب می‌کنند.

۳ – **کلاهبرداری ورشکستگی:** کلاهبرداری ورشکستگی شامل استفاده از کارت در حالی که ورشکسته هستند و خرید کالا در حالی که می‌دانند نمی‌توانند پرداخت کنند. این‌ها می‌توانند با تکنیک‌های طبقه‌بندی کارت جلوگیری شوند .

۴ – **کلاهبرداری داخلی:** وقتی کارکنان بانک جزئیات کارت اعتباری را برای استفاده از راه دور سرقت می‌کنند، کلاهبرداری داخلی نامیده می‌شود.

۵ – **کلاهبرداری تقلبی/رفتاری:** حضور دارنده کارت هنگام ایجاد معاملات از راه دور ضروری نیست. تنها کارت‌های معتبر مورد نیاز است. جزئیات کارت از روش های مختلفی به دست می‌آید. تشخیص این

نوع کلاهبرداری کارت اعتباری زمان و روش‌های ظریف برای شناسایی الگوهای معاملات نیاز دارد.

معاملات کلاهبردارانه تنها بخش کوچکی از کل معاملات را تشکیل می‌دهند و زمانی که با تعداد بسیار اندکی معامله کلاهبردارانه در میان ده‌ها هزار معامله واقعی روبرو هستیم، غربالگری دستی غیرقابل دسترس می‌شود و تشخیص معاملات کلاهبردارانه آنلاین به یک مشکل بسیار چالش‌برانگیز تبدیل می‌شود. در این مرحله روش‌های بسیاری در زمینه تشخیص کلاهبرداری تحقیق شده‌اند و مدل‌های موجود از تکنیک‌های مختلفی مانند یادگیری ماشین، داده‌کاوی و یادگیری عمیق برای تشخیص معاملات کلاهبردارانه استفاده می‌کنند *(Maniraj et al., 2019, pp. 110-115)*.

یکی از مشکلات موجود در این زمینه، عدم توازن در مجموعه داده میباشد. توزیع نامتوازن دو کلاس بدین صورت است که الگوریتم های یادگیری ماشین مانند رگرسیون لجستیک بر اساس کلاس اکثریت عملکرد بسیار خوبی دارند *(Santoso & Wibowo, 2019, pp. 102–108)*، در نتیجه کلاس اقلیت نادیده گرفته می‌شود و الگوریتم آن را به عنوان نویز در نظر می‌گیرد *(Mishra & Ghorpade, 2018)*. این بدان معناست که در فرایند شناسایی تقلب، یادگیری ماشین (*ML*) مدل را بر اساس کلاس اکثریت آموزش می‌دهد. برای غلبه بر این مشکل، تکنیک‌ها و الگوریتم‌های متعددی پیاده‌سازی شده‌اند تا شکاف بین دو کلاس کاهش یابد. برخی روش‌ها برای افزایش تعداد نمونه‌ها در کلاس اقلیت از طریق تکرار تصادفی نمونه‌ها، که به آن *over-sampling* می‌گویند، پیاده‌سازی شده‌اند *(Mishra & Ghorpade, 2018) (Pozzolo et al., 2015, pp. 159–166) (Prasetiyo et al., 2021)* و روش‌های دیگری نیز برای کاهش نمونه‌ها در کلاس اکثریت، که به آن *under-sampling* می‌گویند، به کار گرفته شده‌اند *(Pozzolo et al., 2015, pp. 159–166)*.

در این مقاله به منظور مقابله با عدم توازن داده‌ها، از روش *SMOTE*[3] استفاده شده است. همچنین برای طبقه‌بندی معاملات و تشخیص کلاهبرداری، از الگوریتم *Extra Trees Classifier* بهره گرفته شده است. انتظار می‌رود با کاربرد توأمان *SMOTE* و *Extra Trees Classifier* بتوان به دقت بالایی در شناسایی تراکنش‌های کلاهبردارانه در این مجموعه داده‌ی نامتوازن دست یافت. در ادامه، روند پیاده‌سازی و نتایج این روش‌ها ارائه خواهد شد.

۲. پیشینه تحقیق

شناسایی تقلب در کارت‌های اعتباری به دلیل افزایش تعداد تراکنش‌های آنلاین و افزایش متناظر فعالیت‌های غیرمجاز، حوزه مهمی از تحقیقات بوده است. روش‌های مختلف یادگیری ماشین و یادگیری عمیق برای شناسایی این تقلب‌ها به کار گرفته شده‌اند و عملکرد آن‌ها معمولاً بر اساس معیارهایی مانند دقت (*accuracy*)، صحت (*precision*)، بازخوانی (*recall*) و مساحت زیر منحنی (*AUC*) ارزیابی می‌شود.

Krishna در مقاله خود به بررسی و تحلیل مقایسه‌ای بین الگوریتم جنگل تصادفی (*RandomForest*) و رگرسیون لجستیک (*Regression Logistic*) برای شناسایی تقلب در کارت‌های اعتباری نشان پرداخت و نشان داد که جنگل تصادفی با دقت بالاتر (٪ ۷۶/۲۹) در مقایسه با دقت (٪ ۷۴/۶۵) رگرسیون لجستیک داشته است *(Krishna & Praveenchandar, 2022)*.

Khatri و همکاران چندین الگوریتم یادگیری ماشین از جمله *KNN*[4]، *LR*[5]، *DT*[6]، *RF*[7] و *NB*[8] را برای تشخیص کلاهبرداری کارت اعتباری مقایسه کردند. با استفاده از مجموعه داده ای که از دارندگان کارت اروپایی در سال

3 Synthetic Minority Oversampling Technique
4 K-Nearest Neighbors
5 Linear Regression
6 Decision Tree
7 Random Forest
8 Naïve Bayes

۲۰۱۳ تولید شده بود، این مدل‌ها که در میان آن‌ها *KNN* بهترین حساسیت و دقت در تشخیص کلاهبرداری را داشت *(Khatri et al., 2020, pp. 680-683)*.

Trivedi و همکاران الگوریتم‌های نظارت شده متعددی از جمله *RF* و *GB*[۹] را آزمایش کردند که نتایج تجربی نشان داد *RF* از نظر معیارهای *Accuracy* و *Precision* دقت بهتری دارد *(Trivedi et al., 2020)*.

Rajora و همکاران مطالعه مقایسه‌ای از الگوریتم‌های متعدد یادگیری ماشین برای تشخیص معاملات کلاهبردارانه را انجام دادند و *RF* و *KNN* را مقایسه کردند و نشان دادند که *RF* دقت و *AUC* بهتری دارد، اما این مطالعات مسئله عدم توازن کلاس‌ها که در مجموعه داده مورد استفاده وجود دارد را مد نظر قرار ندادند. چالش طبقه‌بندی این نوع داده ها زمانی که عدم توازن کلاس وجود دارد، موضوع بسیاری از تحقیقات بوده است و معمولا راه حل مسئله استفاده از تکنیک‌های زیرنمونه‌گیری یا بیش‌نمونه‌گیری است (*under-sampling or over-sampling*) *(Rajora et al., 2018)*.

Dornadula و همکاران به منظور مقابله با عدم توازن شدید کلاس‌ها در داده‌ست‌های معاملات غیرمجاز، از تکنیک نمونه‌گیری *SMOTE* استفاده کردند. با استفاده از الگوریتم‌های *DT*، *LR* و *IF* برای مدل‌های تشخیص کلاهبرداری، نتایج نشان داد که *IF* نسبت به *DT* و *LR* دقت بهتری دارد *(Dornadula & Geetha, 2019)*.

*Makki Sara*و همکاران یک مطالعه تجربی با طبقه بندی های نامتوازن ارائه دادند. آنها از هشت تکنیک *ML* مختلف روی مجموعه داده ای با بیش از ده میلیون تراکنش کارت اعتباری استفاده کردند که ٪ ۵/۹۶ آنها غیرمجاز و ٪ ۹۴/۰۴ مجاز هستند. آنها دریافتند *SVM*[۱۰] و *ANN*[۱۱] بالاترین دقت را دارند، بنابراین از رویکردهای نامتوازن برای مجموعه داده استفاده کردند. مطالعه آنها نشان دادند هنگامی که نامتوازنی شدیدی در مجموعه داده وجود داشته باشد، راهکارهایی که برای حل مسائل نامتوازن طراحی شده‌اند ممکن است نتایج نامطلوبی داشته باشند (*Makki et al., 2019*).

Amusan و همکاران ابتدا سعی کردند با نمونه برداری زیرمجموعه‌ای (*under-sampling*) مجموعه داده مورد استفاده خود را متوازن کنند و سپس سعی کردند روی مدل‌های مختلف طبقه‌بندی الگوریتم یادگیری ماشین مانند رگرسیون لجستیک، جنگل تصادفی، *KNN*و درخت تصمیم کار کنند. جنگل تصادفی عملکرد قابل توجهی با بالاترین دقت ٪ ۹۵ نشان داد. اگرچه الگوریتم های دیگر دقت بالای ٪ ۹۰ دریافت کردند اما هدف اصلی تحقیق صرفاً افزایش دقت نبود بلکه تشخیص دقیق‌تر تقلب بود *(Amusan et al., 2021)*.

در پژوهشی دیگر، *Kumar* و همکاران از الگوریتم جنگل تصادفی (*RFA*) در کنار درخت تصمیم برای تمایز تراکنش‌های غیرمجاز از تراکنش‌های عادی استفاده کردند. این الگوریتم یادگیری ماشین نظارت شده‌ای است که از درخت تصمیم برای طبقه‌بندی مجموعه داده استفاده می‌کند. عملکرد مدل که با دقت ٪۹۰ برای این کار به دست آمد، ارائه شده است. جنگل تصادفی نسبت به روش‌های دیگر مزیت دارد زیرا می‌تواند به عنوان طبقه‌بندی و همچنین رگرسیون استفاده شود *(Kumar et al., 2019)*. استفاده از یادگیری عمیق و شبکه های عصبی نیز در این موضوع پیشنهاد شده‌اند، استفاده از تکنیک جدید تبدیل *Text2IMG* برای تولید تصاویر کوچک از داده‌های متنی و تغذیه آن‌ها به یک معماری شبکه عصبی پیچشی *(CNN)*[۱۲]. این رویکرد پیشنهادی، دقت ٪ ۹۹/۸۷ به دست آورده است *(Alharbi et al., 2022)*.

۳. متدلوژی

۳-۱ مجموعه داده

9 Gradient Boosting
10 Support vector machines
11 Artificial Neural Network
12 Convolutional Neural Network

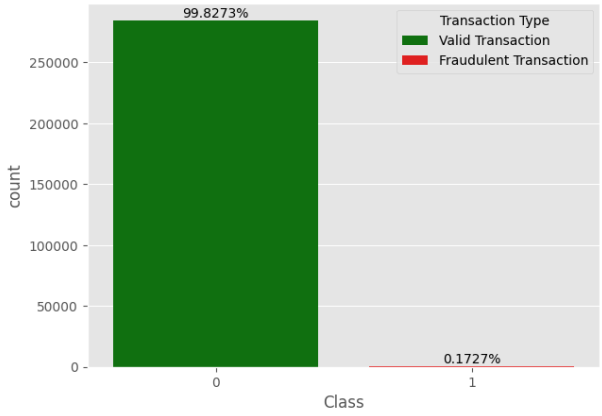
با توجه به عدم دسترسی به دیتاست‌های مرتبط با کارت‌های اعتباری در ایران به دلایل مختلف از جمله حفظ حریم خصوصی، در این پژوهش از یک دیتاست مربوط به قاره اروپا به عنوان یک گزینه مناسب برای تجزیه و تحلیل استفاده شده است. این دیتاست به ما اجازه می‌دهد تا از اطلاعات دقیق و قابل اعتمادی بهره‌مند شویم و همچنین از نقض حریم خصوصی اجتناب کنیم و به احترام حریم خصوصی افراد پایبند باشیم.

با استفاده از یک دیتاست بین‌المللی، می‌توانیم مدل‌هایی برای تشخیص تقلب توسعه دهیم و آزمایش کنیم، در حالی که از حریم خصوصی محافظت می‌کنیم و از نگرانی‌های اخلاقی اجتناب می‌کنیم. تعداد زیاد تراکنش‌های موجود و همچنین ویژگی‌های و متغیر های اصلی، تجزیه و تحلیل عمیق الگوهای تقلب را ممکن می‌سازد. به طور کلی، این دیتاست به عنوان یک جایگزین مناسب که با اهداف تحقیقات تشخیص تقلب همسو است، می‌تواند مورد استفاده قرار گیرد.

این دیتاست[۱۳] شامل تراکنش‌های انجام شده با کارت اعتباری طی دو روز در ماه سپتامبر ۲۰۱۶ می‌باشد. این دیتاست دربرگیرنده ۲۸۴۸۰۷ تراکنش است که ٪ ۰/۱۷۲ آنها تراکنش‌های کلاهبردارانه یا غیرمجاز می‌باشند. "زمان" و "مبلغ تراکنش" بخشی از ۳۰ ویژگی تشکیل دهنده این دیتاست هستند، همراه با ۲۸ متغیر اصلی *V1* تا *V28*. هر ویژگی در دیتاست دارای یک مقدار عددی است. ویژگی‌های *V1* تا *V28* بدون نام‌گذاری هستند تا داده‌ها محرمانه بمانند. همچنین نوع تراکنش در ستون آخر نمایش داده شده است به طوری که مقدار ۰ نشان‌دهنده تراکنش معتبر و مقدار ۱ نشان‌دهنده تراکنش غیرمجاز می‌باشد.

جدول ۱ - ۵ سطر ابتدایی مجموعه داده مورد استفاده

	Time	*V1*	*V2*	*V3*	...	*V26*	*V27*	*V28*	*Amount*	*Class*
0	*0.0*	*-1.359807*	*-0.072781*	*2.536347*	...	*-0.189115*	*0.133558*	*-0.021053*	*149.62*	*0*
1	*0.0*	*1.191857*	*0.266151*	*0.166480*	...	*0.125895*	*-0.008983*	*0.014724*	*2.69*	*0*
2	*1.0*	*-1.358354*	*-1.340163*	*1.773209*	...	*-0.139097*	*-0.055353*	*-0.059752*	*378.66*	*0*
3	*1.0*	*-0.966272*	*-0.185226*	*1.792993*	...	*-0.221929*	*0.062723*	*0.061458*	*123.50*	*0*
4	*2.0*	*-1.158233*	*0.877737*	*1.548718*	...	*0.502292*	*0.219422*	*0.215153*	*69.99*	*0*



نکات قابل توجه در رابطه با این مجموعه داده:

درصدتراکنش‌های مجاز و غیرمجاز در مجموعه داده مورد استفاده

13 (Kaggle, 2018, [Online]. Available: www.kaggle.com/datasets/mlg-ulb/creditcardfraud)

- استفاده از داده‌های پردازش شده به جای داده‌های خام، باعث شده اطلاعات حساس حذف شود که برای حفظ حریم خصوصی مفید است.
- عدم توازن بالای کلاس‌ها در این دیتاست، چالش‌هایی را برای مدل‌سازی با استفاده از الگوریتم های یادگیری ماشین ایجاد می‌کند که نیازمند تکنیک‌هایی برای برقراری توزان می‌باشد.
- وجود تعداد زیاد نمونه و ویژگی‌های پنهان، امکان تحلیل آماری و الگویابی را به خوبی فراهم می‌کند.
- این دیتاست به دلیل حجم بالا و نوع داده‌ها، یک منبع مناسب برای تحقیقات تشخیص تقلب در کارت‌های اعتباری است.

۳-۲ پیش پردازش داده

نکته قابل توجه در این مجموعه داده، عدم توازن شدید در بین نوع تراکنش ها می باشد. در این مقاله برای توزان در مجومع داده و افزایش دقت و کارایی مدل پیشنهادی، از تکنیک *SMOTE* استفاده شده است. روش *SMOTE (Synthetic Minority Over-sampling Technique)* یک تکنیک پرکاربرد در پردازش داده ها است که برای حل مشکل عدم تعادل کلاس‌ها در مجموعه داده استفاده می‌شود. این مشکل معمولاً در مواقعی رخ می‌دهد که تعداد نمونه های یک یا چند ویژگی به طور چشمگیر کمتر از نمونه های ویژگی اکثریت است. این مسئله می‌تواند به مشکلاتی مانند کاهش دقت مدل و افزایش تمرکز بر کلاس اکثریت منجر شود.

نحوه عملکرد *SMOTE* به شرح زیر است:

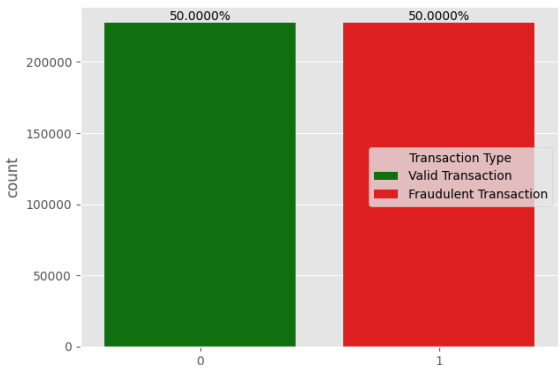
۱ - انتخاب نمونه‌های کلاس اقلیتی: در ابتدا، نمونه های ویژگی اقلیت از مجموعه داده اصلی انتخاب می‌شوند.

۲ - انتخاب همسایگان: برای هر نمونه از ویژگی اقلیت انتخاب شده، نزدیک‌ترین همسایه‌های آن از ویژگی اقلیتی نیز انتخاب می‌شوند. این نمونه‌ها به عنوان همسایگان برای نمونه اصلی انتخاب شده عمل می‌کنند.

۳ - ایجاد نمونه‌های مصنوعی: برای هر نمونه اصلی از کلاس اقلیتی، یکی از همسایگان انتخاب شده و یک نقطه میانی در فضای ویژگی‌ها بین آن نمونه و همسایگانش محاسبه می‌شود. سپس یک مقدار تصادفی بین صفر و یک انتخاب می‌شود و با ضرب این مقدار در فاصله بین نمونه اصلی و نقطه میانی، یک نمونه مصنوعی جدید ایجاد می‌شود.

۴ - اضافه کردن نمونه‌های مصنوعی: نمونه‌های مصنوعی به مجموعه داده اصلی افزوده می‌شوند.

SMOTE به عنوان یک روش موثر برای تعداد نمونه‌های ویژگی اقلیتی استفاده می شود و می‌تواند به بهبود توانایی مدل‌های یادگیری ماشین در تشخیص کلاس اقلیتی کمک کند. این روش برای مسائلی که با عدم تعادل داده‌ها روبه‌رو هستند، بسیار مفید است و در بسیاری از حوزه‌های کاربردی از جمله تصویربرداری پزشکی، تصویرگری ماهواره‌ای، تشخیص تقلب در معاملات مالی و غیره استفاده می‌شود *(Liu et al., 2023, pp. 6224-6227)*.



متوازن سازی مجموعه داده با استفاده از روش SMOTE

در مرحله آخر پیش پردازش داده، مجموعه داده متوازن شده، به دو بخش ویژگی و هدف تقسیم شده است. ابتدا، داده‌های مورد نیاز از مجموعه داده استخراج شده و به صورت یک ماتریس با نام X نگهداری می‌شوند. این ماتریس شامل ویژگی‌های مختلفی از تراکنش‌ها است که شامل ویژگی های $V1$ تا $V28$ و همچنین *Amount* (مقدار تراکنش) استفاده شده است. علاوه بر این، یک ماتریس دیگر به نام y نیز برای نگهداری برچسب‌های کلاس (مجاز یا غیرمجاز) اختصاص داده می‌شود که اطلاعات مربوط به کلاس تراکنش‌ها را در خود جای داده است.

سپس، داده‌ها به دو قسمت آموزش و آزمون تقسیم می‌شوند. در این بخش ۸۰ درصد از داده‌ها برای آموزش مدل استفاده می‌شوند و ۲۰ درصد باقی‌مانده برای ارزیابی عملکرد مدل در مرحله آزمون مورد استفاده قرار می‌گیرد.

۳-۳ الگوریتم مورد استفاده

Classification یا طبقه بندی، یکی از مهم‌ترین وظایف در حوزه یادگیری ماشین است که از آن برای تفکیک داده‌ها به دسته‌های مختلف استفاده می‌شود. این فرآیند به کمک الگوریتم‌های یادگیری ماشین، داده‌های جدید را بر اساس الگوهای موجود در داده‌های آموزشی دسته‌بندی می‌کند. در زمینه مسائل طبقه بندی، الگوریتم‌های مختلفی وجود دارند که هر یک ویژگی‌ها و مزایا و معایب خاص خود را دارند. یکی از این الگوریتم‌ها الگوریتم *Extra Trees Classifier*است که در حوزه‌ی یادگیری ماشین به خصوص در مسائل طبقه بندی و رگرسیون کاربرد دارد *(Akhtar & Feng, 2023, p. 946)*.

Extra Trees Classifier یکی از الگوریتم‌های محبوب درخت تصمیم‌گیری در یادگیری ماشین است که از تکنیک‌های *Ensemble Learning* برای بهبود دقت پیش‌بینی استفاده می‌کند. این الگوریتم، درختان تصمیم‌گیری متعددی را آموزش می‌دهد که هر کدام با تصادفی‌سازی پارامترهایی مانند آستانه برش و ویژگی‌های مورد استفاده در هر شاخه‌بندی، متفاوت هستند. سپس با ترکیب پیش‌بینی‌های این درخت‌ها، پیش‌بینی نهایی انجام می‌شود. مزایای این الگوریتم شامل سرعت بالای آموزش، جلوگیری از *Overfitting* و عملکرد خوب در مجموعه داده‌های با ابعاد بالا است. از *Extra Trees Classifier* معمولاً در مسائلی مانند طبقه‌بندی، رگرسیون و خوشه‌بندی استفاده می‌شود. الگوریتم *Extra Trees Classifier*یک نوع از الگوریتم‌های درخت تصمیم مانند *Random Forest* است. اما تفاوت اصلی بین این دو الگوریتم در روشی است که از آن‌ها برای تولید درخت‌ها و تصمیم‌گیری‌ها استفاده می‌کنند. در الگوریتم *Extra Trees Classifier*، تصمیم‌گیری‌ها به صورت تصادفی انجام می‌شود. به عبارت دیگر، این الگوریتم از تصمیم‌هایی که به صورت تصادفی انجام می‌شوند برای انتخاب بهترین ویژگی‌ها برای جداکردن داده‌ها استفاده می‌کند *(Martiello Mastelini et al., 2023, pp. 6755-6767)*.

طبقه‌بندی با استفاده از این الگوریتم چندین مزیت نسبت به سایر طبقه‌بندها دارد:

۱ - **مقاومت در برابر بیش‌برازش:** *Extra Trees Classifier* به عنوان یک روش طبقه بندی مبتنی بر مجموعه‌ای از طبقه‌بندهای درخت تصمیم‌گیری، به خاطر توانایی جلوگیری از بیش‌برازش شناخته شده است. این مزیت مهمی نسبت به سایر طبقه‌بندهاست، زیرا بیش‌برازش می‌تواند منجر به تعمیم ضعیف به داده‌های دیده نشده شود *(Dhananjay et al., 2021, pp. 402-406)*.

۲ - **پردازش داده‌ها در ابعاد بالا:** این الگوریتم قادر است داده های پربعد را به طور مؤثری پردازش کند. این ویژگی آن را برای انجام تکالیف پیچیده‌ای که شامل تعداد زیادی ویژگی یا متغیر هستند، مناسب می‌سازد *(Padmaja et al., 2020)*.

۳ - **کارایی:** این الگوریتم از نظر محاسباتی کارا است. این الگوریتم به گونه‌ای طراحی شده است که نقاط برش

بهینه را به سرعت به دست می‌آورد و فرایند ساخت مدل را سریع‌تر می‌کند *(Padmaja et al., 2020)*.

۴ - تصادفی‌سازی: *Extra Trees Classifier* سطح اضافی تصادفی‌سازی را نسبت به سایر روش‌های مبتنی بر درخت معرفی می‌کند. این الگوریتم نقاط برش را کاملاً تصادفی انتخاب می‌کند که می‌تواند تنوع مدل را افزایش دهد و همبستگی بین درخت‌ها را کاهش دهد. این امر می‌تواند منجر به بهبود عملکرد مدل شود *(Zafari et al., 2020, pp. 1702-1706)*.

۵ - هزینه محاسباتی پایین: *Extra Trees Classifier* دارای هزینه محاسباتی نسبتاً پایینی نسبت به سایر طبقه‌بندها است. این ویژگی آن را برای مسائل مقیاس بزرگ یا کاربردهایی که در آن‌ها منابع محاسباتی محدود است، انتخاب عملی می‌کند *(Zafari et al., 2020, pp. 1702-1706)*.

۶ - قابلیت تفسیر: این الگوریتم مانند سایر روش‌های مبتنی بر درخت، سطحی از قابلیت تفسیر را ارائه می‌دهد. این الگوریتم بینشی در مورد اهمیت ویژگی‌های مختلف در مدل فراهم می‌کند که می‌تواند در درک الگوهای زیربنایی در داده‌ها با ارزش باشد*(Bombara & Belta, 2021, pp. 1-23)*. در مجموع، *Extra Trees Classifier* ترکیبی از مقاومت در برابر بیش‌برازش، کارایی، تصادفی‌سازی، هزینه محاسباتی پایین، قابلیت تفسیر و انعطاف‌پذیری را ارائه می‌دهد که آن را به الگوریتمی قدرتمند در زمینه یادگیری ماشین تبدیل کرده است.

۴-۳ بهینه سازی پارامترهای الگوریتم

بهینه‌سازی پارامترها نقش مهمی در بهبود کارایی مدلها دارد و می‌تواند بازدهی الگوریتم را بهبود بخشد. در این مقاله برای بهینه‌سازی پارامترها از روش *RandomizedSearchCV* که یکی از روش‌های مؤثر بهینه‌سازی پارامترهای الگوریتم میباشد، استفاده شده است. در روش *RandomizedSearchCV* فضای پارامتری به صورت یک شبکه در نظر گرفته می‌شود و مقادیر مختلفی برای هر پارامتر در نظر گرفته می‌شود. سپس الگوریتم با ترکیب‌های مختلف پارامترها اجرا شده و عملکرد آن‌ها بر روی داده‌های اعتبارسنجی مورد ارزیابی قرار می‌گیرد. در پایان، ترکیب پارامترهایی که بهترین عملکرد را داشته‌اند، به عنوان تنظیمات بهینه انتخاب می‌شوند. با استفاده از روش *RandomizedSearchCV* می‌توان پارامترهای الگوریتم *Extra Trees Classifier* را به صورت کامل بهینه‌سازی نمود و مدلی با بالاترین دقت پیش‌بینی تولید کرد. انتخاب صحیح محدوده پارامترها و فواصل جستجو نیز در موفقیت این روش حائز اهمیت است.

۴. نتایج

۴-۱ انتخاب مناسب ترین الگوریتم

در این پژوهش ، برای انتخاب بهترین مدل، چندین الگوریتم مختلف جهت طبقه‌بندی داده‌های مورد مطالعه قرار گرفت. الگوریتم‌های مورد بررسی عبارتند از *Extra Trees Classifier*، *Random Forest Classifier*، *XGBoost*، *LightGBM* و *Dummy Classifier*. از نظر معیارهای ارزیابی مدل و همچنین زمان آموزش مدل، *Extra Trees Classifier* به نسبت سایر الگوریتم‌ها، عملکرد بهتری از خود نشان داد. به همین دلیل، این الگوریتم به عنوان الگوریتم اصلی برای مدل‌سازی در این تحقیق انتخاب شد. همان‌طور که در جدول ۲ نتایج مقایسه بین الگوریتم ها مشاهده می‌شود، نتایج نشان داد که *Extra Trees Classifier* با دقت ۰/۹۹۹۶، که نشان‌دهنده کمترین پیش‌بینی‌های نادرست است، و بالاترین مقدار *AUC* با ۰/۹۷۴۲، که نشان‌دهنده بهترین توانایی در تفکیک بین دو کلاس (تقلب یا عدم تقلب) بوده است. همچنین حساسیت (*recall*) برابر با ۰/۸۴۰۳ بوده و ٪۸۴ از موارد مثبت (تقلب) را به درستی شناسایی کرده است. دقت (*precision*) با مقدار ۰/۹۰۴۲ نیز نشان‌دهنده این است که حدود ٪۹۰ از مواردی که به عنوان مثبت پیش‌بینی شده بودند، واقعاً مثبت بوده‌اند. امتیاز *F1* نیز با بالاترین مقدار ۰/۸۷۰۴، نشان‌دهنده توازن خوبی بین دقت و حساسیت در این الگوریتم است.

جدول ۲ – مقایسه عملکرد الگوریتم های مختلف

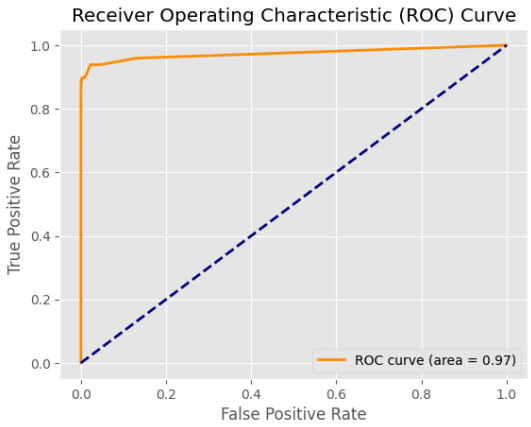
Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	TT (Sec)
Extra Trees Classifier	0.9996	0.9742	0.8403	0.9042	0.8704	0.8702	0.8711	12.9590
Random Forest Classifier	0.9995	0.9713	0.8460	0.8811	0.8623	0.8620	0.8627	81.3090
Extreme Gradient Boosting	0.9995	0.9822	0.8548	0.8594	0.8558	0.8555	0.8562	78.0360
Light Gradient Boosting Machine	0.9992	0.9797	0.8549	0.7406	0.7900	0.7896	0.7935	2.4540
Dummy Classifier	0.9983	0.5000	0.0000	0.0000	0.0000	0.0000	0.0000	0.1370

۲-۴- آموزش الگوریتم *Extra Trees Classifier*

در ابتدا، مدل با استفاده از داده‌های آموزشی، آموزش داده شد. سپس، با استفاده از داده‌های آزمایشی، مدل اولیه مورد ارزیابی قرار گرفته و معیارهای مختلف ارزیابی نتایج محاسبه شدند.

Accuracy:	0.999
precision:	0.911
recall:	0.837
f1:	0.872

جدول ۳ – نتایج اولیه عملکرد مدل Extra Trees Classifier بدون بهینه سازی

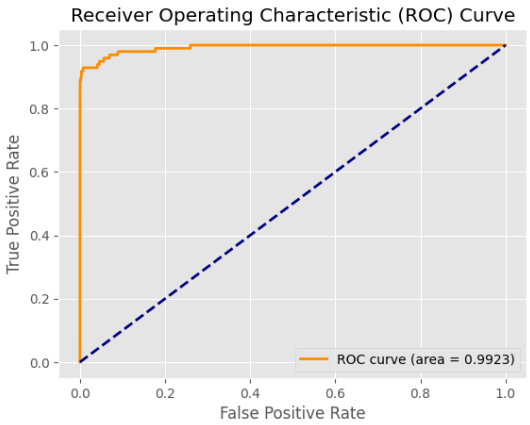


با توجه به جدول ۳، الگوریتم *Extra Trees Classifier* در ابتدا بدون اعمال هیچگونه بهینه‌سازی اجرا شد تا نتایج اولیه به دقت مورد ارزیابی قرار گیرد. در این مرحله، دقت الگوریتم به مقدار ۰/۹۹۹ رسید که نشان‌دهنده دقت بسیار بالای مدل در دسته بندی تراکنش های مجاز و غیرمجاز و تشخیص تقلب در کارت‌های اعتباری است. برای ارزیابی دقیق‌تر عملکرد مدل، از معیارهای دیگری همچون دقت دسته‌بندی *(precision)*، حساسیت *(Recall)*، امتیاز *F1* و مساحت زیر منحنی *(ROC Curve)* استفاده شد. دقت *(precision)* با مقدار ۰/۹۱۱ نیز نشان می‌دهد که الگوریتم در تشخیص موارد مثبت (تقلب) بسیار دقیق عمل کرده است. حساسیت یا *Recall* نیز با مقدار ۰/۸۳۷ نشان‌دهنده توانایی الگوریتم در شناسایی تعداد قابل توجهی از موارد مثبت (تقلب) است. امتیاز *F1* نیز با مقدار ۰/۸۷۲ به عنوان یک معیار ترکیبی از دقت و حساسیت، نشان‌دهنده تعادل مناسب بین این دو است. همچنین، مساحت زیر منحنی *(ROC Curve)* با مقدار ۰/۹۷۴ نشان‌دهنده توانایی مدل در تفکیک بین کلاس‌های مختلف است. این مقدار نشان‌دهنده احتمال مثبت درست تشخیص داده شده توسط مدل است و از اهمیت بالای توانایی تفکیک مدل در مواجهه با داده‌های غیرمجاز خبر می‌دهد.

در نتایج اولیه به دست آمده، مشاهده می‌شود که مدل توانسته است نمونه‌های مثبت (تقلب) را با دقت بالا تشخیص دهد. استفاده از الگوریتم *Extra Trees Classifier* باعث ارائه یک مدل دقیق و قوی و همچنین سریع در تشخیص تقلب در کارت‌های اعتباری شده است که با دقت بالا و معیارهای ارزیابی مناسب، می‌تواند در سیستم‌های حفاظت از تراکنش‌های مالی مورد استفاده قرار گیرد.

۳-۴ بهینه سازی پارامترهای الگوریتم *Extra Trees Classifier*

به بهینه‌سازی مدل *Extra Trees Classifier* با



استفاده از الگوریتم *Randomized Search* پرداخته شد. هدف از این بخش، یافتن پارامترهای بهینه برای مدل بوده و سپس ارزیابی عملکرد مدل با استفاده از این پارامترها انجام شده است. استفاده از این روش بهینه‌سازی این امکان را می‌دهد که با کاهش زمان جستجو و بهینه‌سازی پارامترها، یک مدل قوی‌تر و دقیق‌تر ایجاد کنیم. مجموعه‌ای از پارامترهای مختلف به عنوان فضای جستجو برای *Randomized Search* مشخص شد. این پارامترها شامل تعداد درخت‌ها *(n_estimators)*، عمق حداکثر درخت *(max_depth)*، حداقل تعداد نمونه‌ها برای تقسیم یک گره *(min_samples_split)*، حداقل تعداد نمونه‌های لازم برای یک برگ *(min_samples_leaf)*و تعداد ویژگی‌های مورد استفاده در هر جدول *(max_features)* بودند. پارامترهای ورودی بهینه پس از بهینه سازی مدل به شرح زیر است:

Best Parameters: {'n_estimators': 200, 'min_samples_split': 5, 'min_samples_leaf': 2, 'max_features': 'log2', 'max_depth': None}

مدل با بهینه‌سازی پارامترها، مجددا بر روی داده‌های آزمایشی اعمال شد و عملکرد آن با معیارهای ارزیابی مختلف محاسبه شد. نتایج ارزیابی بر روی داده‌های تست در جدول ۴ نمایش داده شده است:

Accuracy:	1.0
Precision:	0.902
Recall:	0.847
F1 Score:	0.874

جدول ۴ – نتایج عملکرد مدل Extra Trees Classifier براساس پارامترهای بهینه سازی

نتایج

حاصل در این مرحله نشان می‌دهد که با بهینه‌سازی، الگوریتم همچنان دارای دقت بسیار بالاست، اما در عین حال، دقت *precision* به مقدار کمی کاهش یافته است. پیش از بهینه‌سازی، مدل با دقت بسیار بالایی عمل کرده است. با بررسی نتایج پس از بهینه‌سازی، دیده می‌شود که دقت مدل به همان اندازه قبلی باقی‌مانده است و این نشان‌دهنده ادامه توانمندی مدل در ارائه پاسخ‌های صحیح است. اما تغییراتی در دیگر پارامترها به چشم می‌خورد. دقت *precision* با مقدار ۰/۹۰۲ نشان‌دهنده تعداد کمتری از نمونه‌های مثبت پیش‌بینی‌شده به اشتباه افتاده است، اما این تغییر باعث شده تا مقدار *Recall* یا حساسیت مدل به ۰/۸۴۷ افزایش یابد. این نشان می‌دهد که مدل پس از بهینه‌سازی توانسته است بهترین تعادل بین تشخیص درست مثبت و تشخیص درست منفی را برقرار کند. همچنین، امتیاز *F1* پس از بهینه‌سازی به مقدار ۰/۸۷۴ کمی افزایش یافته است که نشان‌دهنده کمی افزایش توازن بین دقت و حساسیت باشد. با توجه به افزایش مساحت زیر منحنی *(ROC Curve)* به ۰/۹۹۲، مشخص است که مدل پس از بهینه‌سازی توانسته است در تشخیص تفاوت‌های میان دسته‌ها بهبود حاصل کند و عملکرد مثبت تشخیص دهی مدل بهبود یابد.

میتوان گفت پس از بهینه‌سازی، مدل با دقت بالا، *Precision* معقول، *Recall* بهتر و بهبودهای قابل توجه در *ROC Curve* به عنوان یک اندازه‌گیری کلی از عملکرد مدل عمل کرده است. این نتایج نشان می‌دهند که بهینه‌سازی مدل باعث افزایش توانمندی در تشخیص تفاوت‌های بین کلاس‌ها شده و مدل را به یک سطح عملکرد بالاتر رسانده است.

۵. نتیجه گیری و پیشنهادات

شناسایی تقلب در تراکنش‌های کارت‌های اعتباری به یک چالش جدی برای موسسات مالی تبدیل شده است. در حال حاضر، چندین رویکرد برای شناسایی تقلب در کارت‌های اعتباری وجود دارد، اما هیچ یک از این رویکردها نمی‌توانند به طور کامل تقلب را قبل از وقوع آن شناسایی کنند. به طور کلی، شناسایی تقلب پس از وقوع تقلب اتفاق می‌افتد، زمانی که بانک‌ها دچار مشکل می‌شوند تا تراکنش را متوقف کنند. این امر منجر به ایجاد مشکلاتی برای بانک‌ها در تلاش برای توقف تراکنش‌های نادرست می‌شود. این وضعیت نه تنها باعث ضرر مالی می‌شود، بلکه به اعتماد مشتریان نیز آسیب می‌زند و می‌تواند تأثیرات منفی قابل‌توجهی بر اعتبار و امنیت سیستم‌های بانکی داشته باشد.

هدف اصلی این پژوهش، توسعه یک مدل دقیق و کارآمد برای تشخیص تقلب در تراکنش‌های کارت اعتباری بود. نتایج مطالعه نشان داد که الگوریتم *Extra Trees Classifier* با دقت ۰/۹۹۹۶ و مقدار *AUC* برابر ۰/۹۷۴۲ بهترین عملکرد را در مقایسه با سایر الگوریتم‌ها داشت. این مدل با استفاده از روش *SMOTE* برای متوازن‌سازی مجموعه داده‌ها توانست حساسیت *(recall)* ۸۴/۰۳ درصد و دقت *(precision)* ۹۰/۴۲ درصدی را ارائه دهد، که نشان‌دهنده توازن مناسب بین دقت و حساسیت است. استفاده از *RandomizedSearchCV* برای بهینه‌سازی پارامترهای مدل، بهبود عملکرد را فراهم کرد و به نتایج قابل قبولی در تفکیک تراکنش‌های مجاز و غیرمجاز دست یافت. با توجه به نتایج بهینه‌سازی، می‌توان اظهار داشت که مدل بهبودیافته، عملکرد خوبی را از خود نشان داده است. این مدل نه تنها دقت بالایی را حفظ کرده، بلکه در شاخص‌های مهمی همچون *recall* نیز پیشرفت قابل توجهی داشته است. همچنین، افزایش قابل ملاحظه از ۹۷ درصد به ۹۹/۲۳ درصد، در مقدار *ROC*، که معیاری جامع برای ارزیابی کارایی مدل است، نشان‌دهنده ارتقای کلی عملکرد سیستم می‌باشد.

با در نظر گرفتن این پیشرفت‌ها، می‌توان گفت که مدل بهینه‌سازی شده گامی مهم در جهت ایجاد سیستم‌های امنیتی قوی‌تر برای تراکنش‌های مالی برداشته است. الگوریتم *Extra Trees Classifier* با بهره‌گیری از قابلیت‌های مقاومت در برابر بیش‌برازش و انعطاف‌پذیری بالا، به‌عنوان یک الگوریتم یادگیری ماشین قدرتمند در زمینه تشخیص تقلب شناخته شد. همچنین این روش دارای سطح پیش‌بینی بالا و هزینه موثر است. بنابراین، این مدل با توانایی بالای خود در تشخیص الگوهای پیچیده تقلب، می‌تواند به طور قابل توجهی به کاهش خسارات ناشی از فعالیت‌های متقلبانه کمک کند. این امر می‌تواند به افزایش اعتماد مشتریان به سیستم‌های بانکی و کاهش ریسک‌های مالی برای موسسات مالی منجر شود.

تحقیقات آینده در زمینه تشخیص تقلب در کارت‌های اعتباری می‌تواند در چندین جهت گسترش یابد. اولاً، استفاده از رویکردهای ترکیبی یادگیری ماشین می‌تواند مورد بررسی قرار گیرد. ادغام الگوریتم *Extra Trees Classifier* با سایر الگوریتم‌های قدرتمند مانند *XGBoost* یا *LightGBM* می‌تواند منجر به ایجاد مدل‌های قوی‌تر و دقیق‌تر شود. ثانیاً، بکارگیری تکنیک‌های یادگیری عمیق، به ویژه شبکه‌های عصبی پیچشی *(CNN)* و شبکه‌های عصبی بازگشتی *(RNN)*، می‌تواند برای شناسایی الگوهای پیچیده‌تر در داده‌های تراکنش مفید باشد. علاوه بر این، استفاده از تکنیک‌های یادگیری تقویتی برای بهبود مداوم مدل در طول زمان می‌تواند مورد بررسی قرار گیرد. این رویکرد می‌تواند به مدل اجازه دهد تا با تغییرات در الگوهای تقلب سازگار شود و عملکرد خود را به طور پویا بهبود بخشد. همچنین توسعه روش‌های پیشرفته‌تر برای مقابله با عدم توازن داده‌ها، می‌تواند به بهبود بیشتر در تشخیص موارد تقلب کمک کند. تکنیک‌هایی مانند *ADASYN* یا روش‌های نوآورانه و ترکیبی *over-sampling* و *under-sampling* می‌توانند در این زمینه مورد آزمایش قرار گیرند.

منابع

and data science[J]. International Journal of Engineering Research, 2019, 8(9): 110-115.

Santoso SHH, Wibowo W. Integration of synthetic minority oversampling technique for imbalanced class. Indones J Electr Eng Comput Sci. 2019;13(1):102–108. doi: 10.11591/ijeecs. v13.i1. pp102-108.

Mishra A, Ghorpade C (2018) Credit card fraud detection on the skewed data using various classification and ensemble techniques. In: 2018 IEEE International Students' Conference on Electrical, Electronics and Computer Science (SCEECS), 2018, pp 1–5. 10.1109/SCEECS.2018.8546939

Pozzolo AD, Caelen O, Johnson RA, Bontempi G (2015) Calibrating probability with undersampling for unbalanced classification. In: 2015 IEEE Symposium series on computational intelligence, 2015, pp 159–166. 10.1109/SSCI.2015.33

Prasetiyo B, Alamsyah A, Muslim MA, Baroroh N. Evaluation performance recall and F2 score of credit card fraud detection unbalanced dataset using SMOTE oversampling technique. J Phys Conf Ser. 2021;1918(4):042002. doi: 10.1088/1742-6596/1918/4/042002.

Krishna, M.Vamsi and J. Praveenchandar. "Comparative Analysis of Credit Card Fraud Detection using Logistic regression with Random Forest towards an Increase in Accuracy of Prediction." 2022 International Conference on Edge Computing and Applications (ICECAA) (2022): 1097-1101.

S. Khatri, A. Arora and A. P. Agrawal, "Supervised machine learning algorithms for credit card fraud detection: A comparison", Proc. 10th Int. Conf. Cloud Comput. Data Sci. Eng. (Confluence), pp. 680-683, Jan. 2020.

N. K. Trivedi, S. Simaiya, U. K. Lilhore and S. K. Sharma, "An efficient credit card fraud detection model based on machine learning methods", Int. J. Adv. Sci. Technol., vol. 29, no. 5, pp. 3414-3424, 2020

S. Rajora, D. L. Li, C. Jha, N. Bharill, O. P. Patel, S. Joshi, et al., "A comparative study of machine learning techniques for credit card fraud detection based on time variance", Proc.

S. Kotsiantis, D. Kanellopoulos, and P. Pintelas," Handling imbalanced datasets: A review," GESTS international transactions on computer science and engineering 30, no. 1, 2006, pp. 25-36.

Bhutani KBH (2021) COVID-19-the inflexion point for E-commerce. Indian J Econ Bus 20(2), PP 649-659.

Abdelrhim M, Elsayed A (2020) The effect of COVID-19 spread on the E-commerce market: the case of the 5 Largest E-Commerce Companies in the World. In: Social Science Research Network, Rochester, NY, SSRN Scholarly Paper ID 3621166, Jun. 2020. 10.2139/ssrn.3621166.

UNCTAD (2020) COVID-19 and e-commerce: Findings from a survey of online consumers in 9 countries. [Online accessed 6-March-2022]. https://unctad.org/system/files/officialdocument/dtlstictinf2020d1_en.pdf

Guthrie C, Fosso-Wamba S, Arnaud JB. Online consumer resilience during a pandemic: an exploratory study of e-commerce behavior before, during and after a COVID-19 lockdown. J Retail Consum Serv. 2021; 61:102570. doi: 10.1016/j.jretconser.2021.102570.

C. Meng, L. Zhou, and B. Liu," A case study in credit fraud detection with SMOTE and XGBoost," In Journal of Physics: Conference Series, vol. 1601, no. 5, p. 052016. IOP Publishing, 2020.

W. Kramer," Wojtek J. Krzanowski and David J. Hand: ROC curves for ¨ continuous data," 2011, p. 979.

G. Menardi, and N. Torelli," Training and assessing classification rules with imbalanced data," Data mining and knowledge discovery 28, no. 1, 2014, pp. 92-122.

I. Vejalla, S. P. Battula, K. Kalluri and H. K. Kalluri, "Credit Card Fraud Detection Using Machine Learning Techniques," 2023 2nd International Conference on Paradigm Shifts in Communications Embedded Systems, Machine Learning and Signal Processing (PCEMS), Nagpur, India, 2023, pp. 1-4, doi: 10.1109/PCEMS58491.2023.10136040.

Maniraj S P, Saini A, Ahmed S, et al. Credit card fraud detection using machine learning

IEEE International Geoscience and Remote Sensing Symposium (2023): 6224-6227.

Akhtar, Muhammad Shoaib, and Tao Feng. "Evaluation of Machine Learning Algorithms for Malware Detection." Sensors (Basel, Switzerland) vol. 23,2 946. 13 Jan. 2023, doi:10.3390/s23020946

Martiello Mastelini, Saulo et al. "Online Extra Trees Regressor." IEEE transactions on neural networks and learning systems vol. 34,10 (2023): 6755-6767. doi:10.1109/TNNLS.2022.3212859

Dhananjay, B. et al. "Cardiac signals classification based on Extra Trees model." 2021 8th International Conference on Signal Processing and Integrated Networks (SPIN) (2021): 402-406.

Padmaja, Budi et al. "A Novel Random Split Point Procedure Using Extremely Randomized (Extra) Trees Ensemble Method for Human Activity Recognition." EAI Endorsed Trans. Pervasive Health Technol. 6 (2020): e5.

Zafari, Azar et al. "Land Cover Classification Using Extremely Randomized Trees: A Kernel Perspective." IEEE Geoscience and Remote Sensing Letters 17 (2020): 1702-1706.

Bombara, Giuseppe and Calin A. Belta. "Offline and Online Learning of Signal Temporal Logic Formulae Using Decision Trees." ACM Transactions on Cyber-Physical Systems 5 (2021): 1 - 23.

IEEE Symp. Comput. Intell. (SSCI), pp. 1958-1963, Nov. 2018.

Dornadula V N, Geetha S. Credit card fraud detection using machine learning algorithms[J]. Procedia computer science, 2019, 165: 631-641.

Sara makki, Zainabassaghir, Yehiataher, Rafiqul haque, Mohand said, Hacid, Hassan zeneddine," An Experimental Study with Imbalanced classification approaches for Credit card Fraud Detection", IEEE, 2019

E. A. Amusan, O. M. Alade, J. O. Emuoyibofarhe, and O. D. Fenwa, "Credit Card Fraud Detection on Skewed Data using Machine Learning Techniques," LAUJCI, 2021. [Online]. Available: www.laujci.lautech.edu.ng

E. A. M.Suresh Kumar, V.Soundarya , S.Kavitha , E.S.Keerthika, CREDIT CARD FRAUD DETECTION USING RANDOM FOREST ALGORITHM. 2019.

Alharbi, Abdullah et al. "A Novel text2IMG Mechanism of Credit Card Fraud Detection: A Deep Learning Approach." Electronics (2022)

"Credit Card Fraud Detection." Kaggle, 23 Mar. 2018, www.kaggle.com/datasets/mlg-ulb/creditcardfraud.

Liu, Xuebo et al. "An Imbalanced Signal Modulation Classification and Evaluation Method Based on Synthetic Minority Over-Sampling Technique." IGARSS 2023 - 2023

Journal of Financial Knowledge of Securities Analysis

Vol. 18 / No. 66, Summer 2025

Detecting suspicious transactions in credit cards using the Extra Trees Classifier algorithm

Erfan Abedini[1]

Mahmud Hematfar*[2]

Abstract

In recent years, with the development of e-commerce, credit cards have been widely used in transactions for various reasons, including the ease and efficiency they provide. Fraud and unauthorized transactions with credit cards have become a common and significant problem due to the increased use of credit cards. Therefore, financial institutions bear a significant responsibility to update their existing mechanisms to prevent such unauthorized actions. The lack of a specific pattern and an imbalance in the number of authorized and unauthorized transactions have posed challenges in developing fraud detection models and systems.

The main objective of this study is to create a suitable, accurate, and fast solution for fraud detection in credit cards based on machine learning. The proposed model in this article includes the use of the SMOTE method to handle imbalanced data and the Extra Trees Classifier algorithm to identify unauthorized transactions. Various metrics examining accuracy and performance indicate the effective performance of the proposed model.

Key words*: Machine Learning, Machine Learning in Fraud Detection, Credit Card Fraud, Online Fraud, Credit Card Scam.*

[1] *Master's Student, Department of Financial Management, Faculty of Management and Economics, Science and Research Branch, Islamic Azad University, Tehran, Iran .erfan.abedini@srbiau.ac.ir*

[2] *Associate Professor, Department of Accounting, Boroujerd Branch, Islamic Azad University, Boroujerd, Iran (Corresponding Author) dr.hematfar@yahoo.com**