



ORIGINAL RESEARCH PAPER

## Presenting an ensemble model for identifying claims of suspicious damages in agricultural insurance

Y. Ahmadi<sup>1</sup>, A. Pourebrahimi<sup>2,\*</sup>, J. Tanha<sup>3</sup>, A. Rajabzade<sup>4</sup>

<sup>1</sup> Department of Information Technology Management, Science and Research Branch, Islamic Azad University, Tehran, Iran

<sup>2</sup> Department of Industrial Management, Karaj Branch, Islamic Azad University, Karaj, Iran

<sup>3</sup> Department of Information Technology Engineering, Faculty of Electrical and Computer Engineering, University of Tabriz, Tabriz, Iran

<sup>4</sup> Department of Industrial Management, Faculty of Management and Economics, Tarbiat Modares University, Tehran, Iran

### ARTICLE INFO

#### Article History:

Received 30 May 2022

Revised 13 July 2022

Accepted 24 August 2022

#### Keywords:

Agricultural insurance

Anomaly detection

Data mining

Ensemble learning

Unsupervised machine learning

### ABSTRACT

**BACKGROUND AND OBJECTIVES:** It is very difficult and maybe impossible to identify suspicious damage claims in agricultural insurance using traditional methods and using the opinions of experts among a multitude of claims. In the current research, a model for discovering suspicious damage claims in agricultural insurance using data mining techniques has been presented to help the agricultural insurance fund in identifying such claims.

**METHODS:** The research method in the present research is applied in terms of intention and descriptive-post-event in terms of quiddity. One of the applications of data mining is anomaly detection. In the present study, a technique for detecting anomalies in the data using ensemble machine learning models is carried out. To enforcement this method, real data on compensation paid for wheat insurance (irrigated and rainfed) for one year in Khuzestan province was used. Because of differences in the process of determining damages of irrigated and rainfed wheat insurance policies, their anomalies were analyzed separately and a number of suspicious claims were acquired for each.

**FINDINGS:** The analysis of the results showed 5 types of suspicious behavior in claiming damages. The ratio of suspicious claims to the total (percentage of anomalies) was estimated using the histogram of anomalous scores and the opinion of insurance fund experts about 1.5%. Suspicious and unusual cases were examined by experts and the final accuracy of the model in correctly identifying suspicious cases was 72% for irrigated wheat insurance and 68% for dryland wheat insurance.

**CONCLUSION:** Based on the obtained results, the presented model can be used to detect suspicious claims in wet and dry wheat insurance policies. Since most of the unusual cases are caused by not providing sufficient documentation, it can be due to the presentation of forging insurance policies or the existence of collusion between the insured, the agent or the assessor. Therefore, more care should be taken in the payment process. The present study was conducted on the product and can be used for other crops as well.

\*Corresponding Author:

Email: [a.pourebrahimi@kiau.ac.ir](mailto:a.pourebrahimi@kiau.ac.ir)

Phone: +9821 44210808

ORCID: [0000-0001-5741-0260](https://orcid.org/0000-0001-5741-0260)

DOI: [10.22056/ijir.2023.01.06](https://doi.org/10.22056/ijir.2023.01.06)

This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).





## مقاله علمی

### ارائه یک مدل ترکیبی برای تشخیص ادعاهای مشکوک خسارت در بیمه کشاورزی

یعقوب احمدلو<sup>۱</sup>، علیرضا پوراابراهیمی<sup>۲\*</sup>، جعفر تنها<sup>۳</sup>، علی رجب زاده<sup>۴</sup>

<sup>۱</sup> گروه مدیریت فناوری اطلاعات، دانشکده مدیریت و اقتصاد، واحد علوم و تحقیقات، دانشگاه آزاد اسلامی، تهران، ایران

<sup>۲</sup> گروه مدیریت صنعتی، دانشکده مدیریت و حسابداری، دانشگاه آزاد اسلامی، کرج، ایران

<sup>۳</sup> گروه مهندسی فناوری اطلاعات، دانشکده مهندسی برق و کامپیوتر، دانشگاه تبریز، تبریز، ایران

<sup>۴</sup> گروه مدیریت صنعتی، دانشکده مدیریت و اقتصاد، دانشگاه تربیت مدرس، تهران، ایران

#### چکیده:

**پیشینه و اهداف:** شناسایی ادعاهای مشکوک خسارت در بیمه کشاورزی با استفاده از روش‌های سنتی، با بهره‌گیری از کارشناسان در میان انبوه ادعاها بسیار دشوار و شاید غیرممکن باشد. در پژوهش حاضر، مدلی برای کشف ادعاهای مشکوک خسارت در بیمه کشاورزی با استفاده از تکنیک‌های داده‌کاوی ارائه شده است تا به صندوق بیمه کشاورزی در شناسایی این‌گونه ادعاها کمک نماید.

**روش‌شناسی:** روش تحقیق در این پژوهش از نظر هدف، کاربردی و از نظر ماهیت، توصیفی - پس‌رویدادی است. یکی از کاربردهای داده‌کاوی، تشخیص ناهنجاری است. در مطالعه حاضر، روشی برای تشخیص ناهنجاری‌ها در داده‌ها با استفاده از مدل‌های ترکیبی یادگیری ماشین ارائه شده است. برای اجرای این روش، از داده‌های واقعی خسارت پرداختی به بیمه‌گندم (آبی و دیم) به مدت یک سال در استان خوزستان استفاده شده است. با توجه به تفاوت در روند تعیین خسارت بیمه‌نامه‌های گندم آبی و دیم، تحلیل ناهنجاری آنها به تفکیک انجام شده و برای هر کدام تعداد ادعاهای مشکوک به صورت جداگانه به دست آمد.

**یافته‌ها:** تجزیه و تحلیل نتایج، ۵ نوع رفتار مشکوک را در ادعای خسارت نشان داد. نسبت ادعاهای مشکوک به کل (درصد ناهنجاری‌ها) با استفاده از هیستوگرام نمرات ناهنجاری و نظر کارشناسان صندوق بیمه حدود ۱٫۵ درصد برآورد شد. موارد مشکوک و غیرعادی توسط کارشناسان مورد بررسی قرار گرفت و دقت نهایی مدل در تشخیص صحیح موارد مشکوک برای بیمه‌نامه گندم آبی و دیم به ترتیب ۷۲ و ۶۸ درصد به دست آمد.

**نتیجه‌گیری:** بر اساس نتایج به دست آمده، می‌توان از مدل ارائه‌شده برای شناسایی ادعاهای مشکوک خسارت در بیمه‌نامه‌های گندم آبی و دیم استفاده نمود. از آنجا که بیشتر موارد غیرعادی ناشی از عدم ارائه مستندات کافی می‌باشد، علت این موضوع می‌تواند به دلیل ارائه بیمه‌نامه‌های صوری یا وجود تبانی بین بیمه‌گذار، نماینده بیمه‌گر یا ارزیاب باشد. بنابراین، بایستی در روند پرداخت خسارت دقت و بررسی بیشتری صورت گیرد. مطالعه حاضر بر روی محصول گندم انجام شده و قابل استفاده برای سایر محصولات زراعی می‌باشد.

#### اطلاعات مقاله

##### تاریخ‌های مقاله:

تاریخ دریافت: ۰۹ خرداد ۱۴۰۱  
تاریخ داوری: ۲۲ تیر ۱۴۰۱  
تاریخ پذیرش: ۰۲ شهریور ۱۴۰۱

##### کلمات کلیدی:

بیمه کشاورزی  
تشخیص موارد غیرنرمال  
داده‌کاوی  
یادگیری ترکیبی  
یادگیری ماشینی بدون نظارت

##### نویسنده مسئول:

ایمیل: [A.pourebrahimi@kiau.ac.ir](mailto:A.pourebrahimi@kiau.ac.ir)

تلفن: ۰۸۰۸۴۴۲۱۰۹۸۲۱+

ORCID: 0000-0001-5741-0260

DOI: 10.22056/ijir.2023.01.06

توجه: مدت زمان بحث و انتقاد برای این مقاله تا ۱ آوریل ۲۰۲۳ در وب سایت IJIR در «نمایش مقاله» باز می‌باشد

## مبانی نظری پژوهش

یکی از مهم‌ترین مشکلات صنعت بیمه، کلاهبرداری است که به موازات رشد و گسترش این صنعت راه‌های کلاهبرداری و کسب درآمد نامشروع در این حوزه نیز رو به افزایش است (Rostami and Monazamitabar, 2021). از کل حجم هزار میلیارد دلاری صنعت بیمه دنیا، ۲۵ درصد آن را کلاهبرداری تشکیل می‌دهد (Hosseini and Rezaei, 2019). طبق آمار منتشره از سوی ائتلاف ضد کلاهبرداری بیمه، سالانه حدود ۸۰ میلیارد دلار در آمریکا از طریق کلاهبرداری بیمه از مصرف‌کنندگان سرقت می‌شود. Gill et al. (2005) کلاهبرداری بیمه‌ای را به این صورت تعریف کرده‌اند: «ادعای ساختگی خسارت به صورت آگاهانه، اعلام خسارت بیش از میزان واقعی آن، یا به هر نحوی که بیمه‌گذار مبلغی بیش از آنچه مستحق آن است به دست آورد». بانک‌ها و مؤسسات بیمه به دلیل هزینه‌های بسیار زیاد کلاهبرداری به دنبال تحلیل روش‌های کلاهبرداران در جهت شناخت و مقابله با آن هستند (Roholamini et al., 2020).

تقلب در انواع بیمه‌ها از جمله بیمه کشاورزی دیده می‌شود. طبق تخمین، کلاهبرداری بیمه محصولات زراعی تقریباً پنج درصد از کل آن را تشکیل می‌دهد که عدد قابل ملاحظه‌ای است (Marzen, 2013). تقلب در بیمه کشاورزی می‌تواند از طریق اظهارات نادرست، خسارت ساختگی، تبانی بین عوامل شرکت بیمه‌گر و بیمه‌گذار و جابجایی در اعلام میزان بازده/عملکرد محصول رخ دهد (Crop insurance, 2006; Marzen, 2013). به عنوان مثال، در پرونده‌ای مربوط به کلاهبرداری بیمه کشاورزی در انبار دخیانیات مونت استرلینگ، کنتاکی که در سال ۲۰۱۴ رخ داده، ۴۰ نفر متهم دارد. در این پرونده، رایج‌ترین وضعیت مربوط به کشاورزانی بود که محصول تنباکوی خوبی تولید کرده بودند اما با نمایندگان بیمه و تنظیم‌کنندگان خسارت تبانی کردند تا ادعا کنند محصول در اثر طوفان یا آفات نباتی آسیب دیده است. بدین صورت که کشاورز ادعای بیمه می‌کرد و غرامت به او پرداخت می‌شد. برای انجام این کار، نمایندگان بیمه و تنظیم‌کنندگان خسارت رشوه دریافت می‌کردند (Clayton, 2022).

با توجه به کاربرد گسترده داده‌کاوی در بازارهای مالی و بیمه، می‌تواند به عنوان ابزاری برای تشخیص رفتارهای غیرنرمال در بیمه کشاورزی پتانسیل داشته و برای کشف تقلب در بیمه محصولات کشاورزی کمک کند (Rejesus et al., 2004). همچنین تکنیک‌های داده‌کاوی به منظور تعمیم مدل‌های کشف ادعاهای تقلبی و ارائه پیش‌بینی مورد استفاده قرار می‌گیرند (Goodarzi and Janatbabaie, 2017) در حال حاضر، داده‌کاوی در حال تبدیل شدن به یک حوزه پرکاربرد راهبردی است که برای یافتن الگوهای پنهان و کشف روابط ناشناخته در مجموعه داده‌ها کمک می‌کند (Roholamini et al., 2020). یکی از رایج‌ترین تکنیک‌های داده‌کاوی که برای یافتن سوابق تقلبی استفاده می‌شود، تشخیص ناهنجاری بوده (Kirlidog and Asuk, 2012) و به دنبال یافتن الگوهای متفاوت در داده‌هاست که با رفتار مورد انتظار و عادی

در بیشتر مناطق با درآمد پایین، مهم‌ترین دغدغه دولت‌ها توسعه بخش کشاورزی و روستایی است. بر اساس گزارش بانک جهانی، ۴۳/۸ درصد از جمعیت جهان و ۷۵ درصد از فقیران جهان در مناطق روستایی زندگی می‌کنند که کشاورزی به عنوان منبع اصلی درآمد و اشتغال آنها است (The World Bank, 2022). تأکید بر توسعه بخش کشاورزی نه فقط به دلیل ارزش ابزاری آن برای رشد سریع بخش‌های دیگر اقتصاد و تأمین مواد غذایی مورد نیاز کشورها، بلکه در جهت افزایش درآمد واقعی کشاورزان و ساکنین روستا است و به همین دلیل برای دولت‌ها از اولویت برخوردار است. خطرات متعددی در فرایند تولید و بازاریابی محصولات کشاورزی وجود دارد. بیمه محصولات کشاورزی، به عنوان یکی از ابزارهای مدیریت ریسک، با ارائه بیمه‌نامه‌هایی که چنین نوساناتی را تحت پوشش قرار می‌دهد، می‌تواند تا حدی ثبات و امنیت مالی را برای کسانی که در این حوزه فعالیت می‌کنند در پی داشته باشد (Cole and Xiong, 2017).

نیاز به یک منبع جبران خسارت به نام «بیمه»، به دلیل پیچیدگی فعالیت‌های اقتصادی در جهان کنونی برای جبران ضرر و زیان ناشی از این فعالیت‌ها امری اجتناب‌ناپذیر است (Rostami and Monazamitabar, 2021). صندوق بیمه کشاورزی به عنوان تنها نهاد بیمه‌گر محصولات کشاورزی در ایران بوده و منابع آن از طریق بانک کشاورزی تأمین شده و در نهایت دولت هزینه‌های پرداخت‌شده را پرداخت می‌کند. ارائه خدمات بیمه‌ای برای محصولات کشاورزی در کشوری مانند ایران که با تغییرات اقلیمی بالا در چند سال اخیر مواجه بوده، دارای ریسک بسیار بالایی می‌باشد (پیشینه بیمه کشاورزی ایران). در کنار این موضوع، یکی از چالش‌هایی که صندوق بیمه کشاورزی با آن مواجه است، ادعاهای غیرواقعی برای دریافت خسارت می‌باشد؛ به عبارت دیگر، پرداخت خسارت به حوادثی که اتفاق نیفتاده‌اند و به عنوان کلاهبرداری مطرح می‌باشد.

از آنجا که کلاهبرداران تمام تلاش خود را برای قانونی جلوه دادن رفتار خود به کار می‌برند و نیز تعداد پرونده‌های قانونی بسیار بیشتر از تعداد پرونده‌های کلاهبرداری است، بنابراین کشف آنها، به عنوان یک مسئله دشوار در بیمه مطرح می‌باشد. مکانیزم‌های دستی و سنتی و حتی سامانه‌های بیمه که هوشمند نیستند، قادر به تشخیص ادعاهای مشکوک نیستند و باعث می‌شود پرداخت‌های قابل توجهی به حوادث ساختگی و خسارت‌های غیرواقعی انجام شود. بنابراین، کشف تقلب به یکی از بهترین کاربردهای داده‌کاوی و یادگیری ماشین در صنعت و دولت تبدیل شده است (Bouazza and Ameer, 2018). کلاهبرداران به مرور از روش‌های جدیدی برای تقلب استفاده می‌کنند و همیشه از لحاظ اطلاعات غنی‌تر از بیمه‌گر هستند. بنابراین، هدف در این پژوهش، که برای اولین بار در حوزه بیمه کشاورزی ایران انجام می‌شود، این است که با استفاده از تکنیک‌های یادگیری ماشین بدون نظارت مدلی ارائه شود تا بتوان با استفاده از داده‌های موجود، ادعاهای مشکوک و روش‌های کلاهبرداری را شناسایی و از پرداخت غرامت جلوگیری به عمل آورد.

بیمه های اتومبیل شامل: رگرسیون لجستیک، درخت تصمیم و دسته‌بندی بیزین را مورد استفاده قرار داده‌اند. داده های مورد نیاز را از یک شرکت بزرگ بیمه‌ای اخذ کرده بودند. آنها داده‌ها را به دو بخش تقسیم کردند؛ از بخش نخست برای ساخت مدل و از بخش دوم برای دسته‌بندی استفاده کردند. یافته‌های این مطالعه نشان داده که مدل رگرسیون لجستیک دقت بیشتری برای پیش‌بینی کل ادعاها، اعم از قلبی و غیر قلبی، نسبت به دو مدل دیگر یعنی درخت تصمیم و روش بیزین ساده داشته است.

(Hosseini and Rezaei (2019) در پژوهشی با عنوان «کشف

تقلب و راهکارهای مقابله با آن در سازمان‌های بیمه‌ای با استفاده از داده‌کاوی» که به‌صورت مطالعه موردی در سازمان تأمین اجتماعی ایران انجام دادند، سه روش داده کاوی درخت تصمیم، ک-نزدیکترین همسایه و شبکه عصبی مصنوعی را مورد استفاده قرار دادند که به ترتیب درخت تصمیم با دقت ۹۹/۶۴ درصد، شبکه عصبی با دقت ۹۹/۰۷ درصد و ک-نزدیکترین همسایه با دقت ۹۶/۸۴ درصد دقت در تشخیص صحیح مواد کلاهبرداری موفق بودند.

(Rejesus et al. (2004) از تکنیک های داده کاوی برای

تشخیص شهرهایی با جریب های برداشت شده غیرنرمال استفاده کردند. آنها شهرهایی که درصد برداشت پایینی (در حد ۵ درصد و پایین تر) داشتند را پرچم گذاری کردند. آنها در کار خود از روش تشخیص موارد پرت با استفاده از امتیاز استاندارد Z بهره بردند.

(Ngai et al. (2011) در مطالعه خود با دسته بندی کاربرد

تکنیک های داده کاوی در کشف کلاهبرداری در حوزه های مختلف مالی از جمله بیمه نشان دادند ۲۶ تکنیک برای این منظور مورد استفاده قرار می گیرد. همچنین از روش های داده کاوی که بیشترین کاربرد را در کشف تقلب داشتند، طبقه بندی ارائه نمود که عبارتند از: رگرسیون، طبقه بندی، خوشه بندی، پیش‌بینی، کشف موارد پرت و بصری‌سازی. یافته‌های این پژوهش نشان می‌دهد که تکنیک‌های داده کاوی به‌طور گسترده برای کشف تقلب استفاده شده است.

(Verma et al. (2017) در کار خود برای تشخیص تقلب در

بیمه سلامت از قوانین انجمنی (Association rules) و خوشه بندی ک-میانگین (k-means) برای شناسایی الگوهای پرتکرار تقلب که در داده های بیمه سلامت نهفته بودند، استفاده نموده و موارد غیرنرمال را با بهره گیری از تابع توزیع گاوسی مشخص کردند.

(Bauder et al. (2018) از دو روش جنگل ایزوله و جنگل

تصادفی بدون نظارت برای شناسایی کلاهبرداری بیمه سلامت استفاده کردند، آنها این دو روش را همراه با روش‌های متداول‌تر شامل عامل دورافتاده محلی (Local outlier factor)، خودرمزگذار (Autoencoder) و ک-نزدیک‌ترین همسایه‌ها (k-nearest neighbors) مورد بررسی قرار دادند. به منظور اعتبارسنجی عملکرد تشخیص تقلب در هر روش، از پایگاه داده فهرست افراد/شرکت‌های مستثنی‌شده استفاده کردند که حاوی اطلاعات ارائه‌دهندگان خدمات سلامت مستثنی‌شده بود. نتایج بررسی نشان داد که عامل دورافتاده محلی، بهترین روش تشخیص پرت/ناهنجاری است و ک-نزدیکترین همسایه با ۵ همسایه و رمزگذارهای خودکار بدترین

مطابقت نداشته (Chandola et al., 2009) و اغلب به منظور کشف ناهنجاری‌های بیمه، کارت بانکی و تشخیص نفوذ شبکه مورد استفاده قرار می‌گیرد (Zhao et al., 2019).

برخلاف اکثر نرم افزارهای رایانه ای، سیستم های داده کاوی وقوع یک رویداد را با دقت ریاضی نشان نمی دهند، بلکه موارد مشکوک به کلاهبرداری را به کمک سوابق گذشته تجزیه و تحلیل کرده، امتیازی به‌عنوان امتیاز ناهنجاری به هر نمونه اختصاص می‌دهد؛ سپس کارشناسان می‌توانند تحقیقات دقیق تری روی مواردی که دارای امتیازهای بالایی هستند، انجام دهند (Kirlidog and Asuk, 2012). تکنیک های یادگیری ماشینی به طور فزاینده ای به عنوان یکی از رویکردهای تشخیص ناهنجاری استفاده می‌شود. یادگیری ماشینی تلاشی است برای «خودکار کردن فرایند کسب دانش از نمونه‌ها» (Bose and Mahapatra, 2001). بنابراین، تشخیص ناهنجاری را می‌توان بر اساس نوع داده آموزشی مورد استفاده برای ساخت مدل، به سه دسته کلی تقسیم کرد: تشخیص ناهنجاری نظارت شده که در آن نمونه‌های عادی و غیرعادی دارای برچسب هستند، تشخیص ناهنجاری بدون نظارت که در آن نمونه‌ها بدون برچسب هستند و تشخیص ناهنجاری نیمه نظارت‌شده که آموزش در اینجا فقط شامل نمونه‌های عادی است؛ بنابراین، هر چیزی که نمی‌تواند به عنوان عادی طبقه بندی شود، به عنوان غیرعادی مشخص می‌شود (Nassif et al., 2021).

به گفته (Gomes et al. (2021) از آنجا که الگوهای رفتاری کلاهبرداری مدام در حال تغییر هستند، تکنیک‌های یادگیری ماشینی بدون نظارت نقش مهمی نسبت به تکنیک‌های نظارت شده در کشف کلاهبرداری ایفا می‌کند.

روش‌های تشخیص ناهنجاری بدون نظارت مستعد ایجاد نرخ‌های مثبت و منفی کاذب بالایی هستند (Das et al., 2016). طبق گفته (Zhao et al. (2019) برای بهبود دقت و پایداری مدل در سناریوهای تشخیص ناهنجاری، محققان رویکردهای ترکیبی (Ensemble) را برای تشخیص ناهنجاری در پیش‌بینی می‌گیرند. در این نوع یادگیری، چندین تخمین‌گر پایه (Base estimator) را برای دستیابی به عملکرد و قابلیت اطمینان بالاتر در مقایسه با یک تخمین‌گر ترکیب می‌کنند.

### مروری بر پیشینه پژوهش

در پژوهشی که توسط (Soltani Halvaie and Akbari (2014) انجام شده، از سیستم ایمنی مصنوعی به منظور شناسایی تقلب کارت‌های اعتباری استفاده کرده‌اند. آنها مدل جدیدی به نام مدل تشخیص تقلب مبتنی بر سیستم ایمنی مصنوعی که از سیستم ایمنی الهام گرفته شده، برای کشف تقلب ارائه نموده‌اند. مدل آنها از نوع نظارت‌شده بود و برای جبران زمان آموزش زیاد از رایانش ابری استفاده کردند.

در مطالعه ای که توسط (Goodarzi and Janatbabaie (2017) برای بررسی عوامل مؤثر بر تقلبات بیمه اتومبیل با کاربرد مدل های داده کاوی تدوین شده است، تکنیک های رایج در کشف تقلب

اخذ شده از آژانس فضایی اروپا برای مزارع یکی از شهرهای فرانسه بود. در ۹۵/۵ درصد موارد زمین های زراعی مشکوک به درستی تشخیص داده شدند. آنها در کار خود روش های دیگر تشخیص ناهنجاری ماشین بردار پشتیبان تک کلاسه، خودمزمگذار و حلقه احتمالات پرت محلی را نیز بررسی کردند، ولی به دلیل پیچیدگی ابرپارامترهای تنظیمی برای سایر تکنیک های مورد بررسی از به کارگیری آنها صرف نظر کردند.

وجه تمایز این تحقیق با تحقیقات انجام شده در وهله اول استفاده از داده های واقعی بیمه کشاورزی می باشد که بعد از پیش پردازش مورد استفاده قرار گرفته و در ادامه استفاده از روش تشخیص ناهنجاری ترکیبی برای کشف ادعاهای مشکوک درخواست خسارت در بیمه کشاورزی ایران می باشد.

#### سؤال های پژوهش

با توجه به هدفی که از انجام این تحقیق دنبال می شود، می توان سؤال های زیر را مطرح نمود:

- چه میزان از ادعاهای خسارت بیمه نامه های گندم غیرواقعی هستند؟
- تکنیک مؤثر یادگیری ماشین برای تشخیص ادعاهای مشکوک خسارت بیمه گندم چگونه می باشد؟

#### روش شناسی پژوهش

پژوهش حاضر از لحاظ هدف، کاربردی و از لحاظ ماهیت، داده محور و از نوع توصیفی - پس رویدادی می باشد. در این تحقیق از فرایند استاندارد صنعتی متقابل برای داده کاوی (CRISP) برای ساخت مدل یادگیری ماشین استفاده شده و شامل: ۱- جمع آوری داده ۲- آماده سازی و پیش پردازش داده ۳- انتخاب الگوریتم و مدل سازی و ۴- ارزیابی مدل و نتایج می باشد. در شکل ۱ چارچوب کلی فرایند تحلیل کشف کلاهبرداری بیمه نشان داده شده است.

یادگیری ماشین بررسی می کند که چگونه رایانه های توانمند براساس داده ها یاد بگیرند یا عملکرد خود را بهبود بخشند (Han et al., 2011). در تحقیق حاضر برای ساخت مدل یادگیری ماشین بدون نظارت، از الگوریتم های تشخیص ناهنجاری پایه و ترکیبی استفاده شده است. الگوریتم های پایه مورد استفاده شامل جنگل ایزوله، ماشین بردار پشتیبان تک کلاسه و تکنیک مورد استفاده برای ترکیب نتایج الگوریتم های پایه، ترکیب انتخابی محلی در مجموعه های پرت موازی

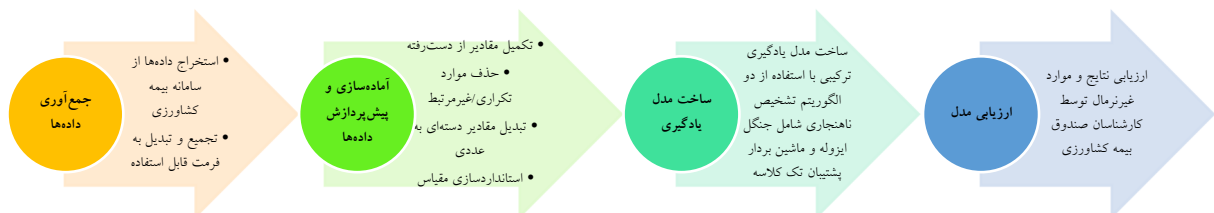
روش ها در تشخیص تقلب بیمه سلامت برای مجموعه داده مورد بررسی هستند.

(Randhawa et al. 2018) در تحقیق خود بر روی تشخیص موارد غیرعادی تراکنش های کارت اعتباری، از روش آدابوست (Adaboost) و رأی حداکثری (Majority vote) استفاده کردند. آنها در کار خود مدل های طبقه بندی از قبیل ناوی بیزین، درخت تصمیم، جنگل تصادفی، شبکه عصبی و رگرسیون لجستیک را به صورت ترکیبی مورد استفاده قرار داده و با بهره گیری از روش های ترکیبی آدابوست و رأی حدکثری، به نتایج بهتری دست پیدا کردند. (Hassan and Abraham 2019) در پژوهش خود از روش تشخیص ناهنجاری ترکیبی به منظور تشخیص تقلب در بیمه خودرو استفاده کردند. آنها از سه روش درخت تصمیم، ماشین بردار پشتیبان و شبکه عصبی مصنوعی به عنوان تخمین گرهای پایه استفاده کردند. برای آزمایش، روش اعتبارسنجی متقاطع ده - برابری مورد بهره برداری قرار گرفت. نتایج بررسی نشان از بهبود دقت مدل ترکیبی نسبت به هر کدام از مدل ها را داشت.

(Kim et al. 2020) در تحقیق خود برای تشخیص نارسایی در موتور کشتی از روش یادگیری ماشین ترکیبی استفاده کردند. آنها با استفاده از ۳۰ مقدار مختلف برای پارامتر تعداد همسایگان (K) برای الگوریتم عامل دورافتاده محلی که به عنوان تخمین گر پایه به کار گرفته شده بود، توانستند نتایج قوی تری به دست آورند.

(Sahni et al. 2020) با بهره گیری از اینترنت اشیا، یادگیری عمیق و بینایی ماشین روشی برای شناسایی تقلب در بیمه کشاورزی ارائه دادند که قادر بود ادعاهای مشکوک خسارت ناشی از آتش سوزی را تشخیص دهد. مدل پیشنهادی آنها ورودی را از حسگرهای مادون قرمز و دمای دستگاه اینترنت اشیا می خواند، و به محض عبور مقادیر سنسور از آستانه های خاص، تصاویر زمین زراعی را جمع آوری می کرد. سپس تصاویر جمع آوری شده به یک مدل تشخیص آتش که با استفاده از طبقه بندی کننده های مختلف برای مقایسه عملکرد آموزش داده شده بودند، وارد می شدند. نتایج نشان می دهد که راه حل پیشنهادی آنها دارای ۹۷ درصد دقت است.

(Mouret et al. 2021) با استفاده از الگوریتم های تشخیص ناهنجاری جنگل ایزوله توانستند زمین های زراعی که رشد غیرنرمال داشتند را شناسایی کنند. آنها با استفاده از این روش موارد مشکوک را شناسایی و سپس با کمک کارشناسان زراعی نسبت به اعتبارسنجی نتایج اقدام کردند. داده های مورد استفاده، تصاویر ماهواره ای



شکل ۱: چارچوب کلی فرایند تحلیل ناهنجاری

(LSCP) می باشد.

آن این است که هر مجموعه داده به صورت خاص و متفاوت برای هر پروژه یادگیری ماشین تهیه شده است. بسیاری از ویژگی های مجموعه داده در شرکت های بیمه غیرمفید، نامربوط و بی تاثیر هستند (Roholamini et al., 2020). ویژگی هایی که برای آموزش مدل های یادگیری ماشینی استفاده می شوند، تأثیر زیادی بر عملکرد مدل خواهد داشت، در نتیجه ویژگی های نامربوط یا تا حدی مرتبط می تواند بر عملکرد مدل تأثیر منفی بگذارد (Brownlee, 2020)؛ همچنین اگر داده های ورودی به دقت غربال نشوند، مدل به دست آمده روی داده های جدید عملکرد خوبی نخواهد داشت (Kim et al., 2020). لذا در این پژوهش، چندین روش آماده سازی و پیش پردازش برای بهبود کیفیت مجموعه داده به کار گرفته شدند.

#### پاک سازی داده ها

در ابتدا ویژگی هایی که مقادیر از دست رفته داشتند تکمیل شده و متناسب با نوع ویژگی مقادیر مناسب بر اساس مقادیر مشابه انتخاب و تکمیل شدند. برای مثال، برای بیمه گذاری که دارای چندین بیمه نامه بود و در برخی از نمونه ها ویژگی نام نماینده خالی بود، با

برای توسعه مدل های یادگیری ماشین، اغلب از مجموعه داده های آماده که در منابع اینترنتی موجود هستند و توسط دانشمندان داده آماده شده، استفاده می شود؛ اما در پژوهش حاضر از داده های واقعی استفاده شده و داده های مورد نیاز از صندوق بیمه کشاورزی ضمن حفظ محرمانگی مشخصات بیمه گذاران اخذ شده است. داده ها مربوط به مشخصات بیمه نامه های محصول گندم آبی و دیم برای سال زراعی ۱۳۹۸-۱۳۹۹ استان خوزستان می باشد که برای آنها خسارت پرداخت شده است. داده ها به صورت گزارش در قالب فایل اکسل از سیستم استخراج و پالایش شدند. مشخصات مجموعه داده استخراجی شامل تمام ویژگی قابل استفاده به همراه ویژگی میانگین خسارت منطقه که یک فیلد مشتق شده می باشد، در جدول ۱ لیست شده است.

مجموعه داده به دست آمده شامل ۲۱،۰۴۳ نمونه بیمه نامه گندم آبی و دیم می باشد که در سال زراعی ۱۳۹۸-۱۳۹۹ خسارت دریافت کرده اند. آماده سازی و پیش پردازش داده ها ممکن است یکی از دشوارترین مراحل در هر پروژه یادگیری ماشین باشد؛ دلیل

جدول ۱: مشخصات مجموعه داده

ردیف	نام ویژگی	نوع داده	نوع مقیاس	توضیح
۱	شناسه قلم بیمه شده	عدد صحیح	اسمی	شناسه یکتا که برای هر قطعه زمین بیمه شده در سیستم تولید می شود.
۲	شناسه بیمه نامه	عدد صحیح	اسمی	شناسه یکتا که برای هر بیمه نامه صادر شده در سیستم تولید می شود.
۳	نوع کشت	رشته ای	اسمی	آبی یا دیم
۴	استان	رشته ای	اسمی	نام یکی از ۳۱ استان کشور
۵	شهر	رشته ای	اسمی	نام شهر
۶	شعبه	رشته ای	اسمی	نام شعبه بانک کشاورزی
۷	نام دهستان	رشته ای	اسمی	نام دهستان
۸	محل	رشته ای	اسمی	نام محل کشت
۹	میزان تعهد کلی	عدد صحیح	نسبی	حداکثر تعهد بیمه گر
۱۰	نوع بیمه	رشته ای	اسمی	پایه، تکمیلی
۱۱	مبلغ غرامت پرداختی	عدد صحیح	نسبی	میزان غرامت پرداخت شده
۱۳	تاریخ وقوع	تاریخ	فاصله ای	تاریخ وقوع خسارت
۱۴	نام بیمه گذار	رشته ای	اسمی	
۱۵	نام صادرکننده	رشته ای	اسمی	
۱۶	نماینده بیمه گر	رشته ای	اسمی	نام بیمه گر
۱۷	مساحت بیمه شده	عدد اعشاری	نسبی	میزان هکتار بیمه شده
۱۸	مساحت خسارت دیده	عدد صحیح	نسبی	میزان هکتار خسارت دیده شده
۱۹	حق بیمه کشاورز	عدد صحیح	نسبی	مبلغ حق بیمه سهم کشاورز
۲۰	نوع طرح	رشته ای	اسمی	شامل: انفرادی، عمومی، پایه، تجمعی، ...
۲۱	عامل خسارت	رشته ای	اسمی	شامل: طوفان، تگرگ، سیل، سرما و یخبندان، خشکسالی، باد گرم
۲۲	ارزیاب	رشته ای	اسمی	نام ارزیاب
۲۳	میانگین مساحت خسارت دیده شده منطقه	عدد اعشاری	نسبی	فیلد مشتق شده از ویژگی میزان مساحت خسارت دیده که نشان دهنده میانگین خسارت منطقه می باشد.

تبدیل شده و ارقام آن رشته باینری به ستون های جداگانه تقسیم می شوند. در این روش داده ها در ابعاد کمتری نسبت به کدگذاری وان هات (One hot) تبدیل می شوند. برای کدگذاری ویژگی تاریخ وقوع خسارت، ابتدا از قالب شمسی به میلادی، سپس به قالب یونیکس تبدیل گردید. در این فرمت، تاریخ وقوع از مبدأ آن (1970/01/01) که معادل با صفر می باشد بر اساس ثانیه محاسبه شده و معادل عددی آن جایگزین تاریخ وقوع در جدول نهایی می شود. در نهایت، بعد از تبدیل تمام ویژگی های غیر عددی، تعداد نهایی ویژگی ها به 38 عدد رسید.

بسیاری از الگوریتم های یادگیری ماشین زمانی که متغیرهای ورودی آن عدد باشد، اگر در محدوده استاندارد مقیاس بندی شوند، عملکرد تفریق میانگین از داده ها (Brownlee, 2020) بهتری خواهند داشت را مرکزیت و تقسیم بر انحراف معیار را مقیاس بندی می نامند. لذا، این روش گاهی اوقات مقیاس بندی استاندارد مرکزی نیز نامیده می شود. برای تبدیل مقیاس متغیرها، از رابطه زیر استفاده می شود

$$y = \frac{x - \text{mean}(x)}{\text{sqrt}(\text{var}(x))} \quad (1)$$

که در آن میانگین به صورت زیر محاسبه می شود:

$$\text{mean}(x) = \frac{1}{N} \sum_{i=1}^N x_i \quad (2)$$

مقیاس بندی، میانگین داده ها را صفر و انحراف معیار آن را یک می کند. در مقیاس بندی، شکل توزیع داده های مقیاس بندی شده فرقی با توزیع داده های اصلی نخواهد داشت و فقط مقیاس محور افقی تغییر می کند.

تجزیه و تحلیل داده ها

در این بخش، به تشریح اعمال الگوریتم تشخیص کلاهبرداری بر روی مجموعه داده پیش پردازش شده پرداخته می شود. همان طور که در بخش های قبلی نیز ذکر شد، این مطالعه یک رویکرد بدون نظارت را اتخاذ می کند، زیرا هیچ متغیر هدفی در مورد نرمال یا غیرنرمال بودن درخواست خسارت در مجموعه داده وجود ندارد. یک مدل بدون نظارت، روابط و الگوهای درون مجموعه داده بدون برچسب را یاد می گیرد، بنابراین، اغلب برای کشف روندهای ذاتی و نهان در یک مجموعه داده معین استفاده می شود. همچنین به دلیل هزینه بالای برچسب گذاری اغلب برای تشخیص ناهنجاری از رویکرد بدون نظارت استفاده می شود (Zhao et al., 2021). رویکرد بدون نظارت فرض می کند که تمام داده های آموزشی وضعیت نرمال دارند، بنابراین، اگر یک مشاهده جدید تفاوت زیادی را داشته باشد، یک ناهنجاری محسوب می شود. هدف از تشخیص ناهنجاری در بیمه کشاورزی، شناسایی نمونه های غیرنرمال خسارت های پرداخت شده مشکوک می باشد. به طور دقیق تر اگر بیان شود، یک ناهنجاری را

استفاده از اطلاعات سایر بیمه نامه های آن بیمه گذار تکمیل گردید؛ طبق گفته کارشناسان صندوق بیمه، بیمه نامه ها برای هر بیمه گذار در یک محل توسط یک نماینده صادر می شود. همچنین نمونه هایی که امکان تکمیل مقادیر از دست رفته آنها به هیچ وجه میسر نبود، حذف شدند.

مجموعه داده دارای نمونه های تکراری بودند، که ناشی از عدم درج حدود اربعه و مختصات جغرافیایی محل بیمه شده بود. لذا در ادامه تعداد 3023 نمونه تکراری از مجموعه داده حذف شدند. وجود نمونه های تکراری تأثیر مثبتی در فرایند یادگیری نداشته و می تواند در عملکرد یادگیری مدل تأثیر منفی داشته باشد (Brownlee, 2020). بعد از حذف موارد تکراری، در نهایت تعداد 18020 نمونه در مجموعه داده نهایی باقی ماندند، که از این تعداد 11293 نمونه مربوط به ادعاهای خسارت بیمه گندم آبی و 6727 نمونه مربوط به گندم دیم بود.

برای انتخاب ویژگی در یادگیری نظارت شده، اغلب از روش فیلتر برای فیلتر کردن ویژگی های مرتبط به برچسب استفاده شده و در یادگیری بدون نظارت، اغلب از روش همبستگی برای انتخاب ویژگی استفاده می شود (Brownlee, 2020). طبق گفته Brownlee, (2020) انتخاب ویژگی در یادگیری بدون نظارت به تکنیک های کاهش ابعاد مرتبط است، زیرا هر دو روش به دنبال متغیرهای ورودی کمتر برای یک مدل یادگیری هستند. از روش های کاهش بعد می توان به تحلیل مؤلفه های اصلی و خودرمزگذار اشاره کرد، که اگر خودرمزگذار با تابع فعال سازی خطی مورد استفاده قرار گیرد، معادل تحلیل مؤلفه های اصلی خواهد بود. در پژوهش حاضر به دلیل عدم وجود متغیر برچسب، نمی توان از روش فیلتر برای انتخاب ویژگی استفاده کرد، در مواردی که تعداد ویژگی ها کم است به جای استفاده از روش های کاهش ابعاد می توان زیرمجموعه هایی از ویژگی ها را انتخاب و مدل را ارزیابی کرد (Brownlee, 2020). لذا پس از استخراج اولیه، با توجه به تعداد ویژگی ها که تعدادشان به 23 رسید، زیرمجموعه ای از ویژگی های تأثیرگذار در تقبل با نظر کارشناسان حوزه بازرسی صندوق بیمه انتخاب و بقیه از مجموعه داده حذف شدند، که در نهایت 13 ویژگی انتخاب شد.

برای مثال، با انتخاب ویژگی «محل» که زمین زراعی بیمه شده در آنجا قرار گرفته بود، سایر ویژگی های نشان دهنده محل از قبیل استان، شهر، دهستان و شعبه حذف شدند؛ چرا که ویژگی «محل» نمایان گر همه این ویژگی ها می باشد. شناسه ها نیز از مجموعه داده نهایی حذف شدند. تأکید شده که برای عملکرد بهینه مدل، بایستی ویژگی هایی که دارای مقادیر یکتا هستند حذف شوند (Brownlee, 2020).

آماده سازی داده ها

از آنجا که اکثر مدل های یادگیری ماشین مقادیر عددی را به عنوان ورودی قبول می کنند، برای این منظور عملیات تبدیل بر روی ویژگی های اسمی (دسته ای) انجام گرفت. برای این کار، از روش کدگذاری باینری استفاده شد. در این تکنیک، ابتدا دسته ها به صورت ترتیبی عددگذاری می شوند، سپس آن اعداد به کد باینری

موجود در مجموعه داده  $X$ .

ب: انتخاب کمترین و بیشترین مقدار در ویژگی  $q$  به نامهای

$max$  و  $min$

ج: انتخاب تصادفی یک مقدار  $p$  بین  $max$  و  $min$

د: انجام آزمون  $q < p$ :

- ساخت یک گره داخلی  $X_i \subseteq X$  به عنوان فرزند چپ درخت،

طوری که ویژگی  $q$  تمامی عناصر مجموعه  $X_i$  کمتر از  $p$  باشد و

فراخوانی مجدد الگوریتم بر روی این مجموعه از داده ها

- ساخت یک گره داخلی  $X_r \subseteq X$  به عنوان فرزند راست درخت،

طوری که ویژگی  $q$  تمامی عناصر مجموعه  $X_r$  بزرگتر یا مساوی  $p$

باشد و فراخوانی مجدد الگوریتم بر روی این مجموعه از داده ها

ارتفاع مسیر نیز به این صورت تعریف می شود (Liu et al.,

2012):

ارتفاع مسیر نقطه  $X(h(x))$  یا طول مسیر با تعداد یال هایی که

از گره ریشه  $X$  یک درخت ایزوله شروع و تا یک گره خارجی ادامه

می یابد، اندازه گیری می شود. طول مسیر به عنوان معیاری برای

میزان حساسیت به ایزوله بودن به کار گرفته می شود:

- طول مسیر کوتاه به معنای حساسیت زیاد به ایزوله،

- طول مسیر طولانی به معنای حساسیت کم به ایزوله است.

از آنجایی که درخت ایزوله ساختاری معادل درخت جستجوی

باینری دارد، لذا تخمین میانگین  $h(x)$  برای خاتمه گره های خارجی

مانند جستجوهای ناموفق در درخت جستجوی دودویی است؛

در یک نمونه داده  $\psi$  میانگین طول مسیر جستجوهای ناموفق در

درخت جستجوی باینری به صورت زیر می باشد:

$$c(\psi) = \begin{cases} 2H(\psi - 1) - \frac{2(\psi - 1)}{n} & \text{for } \psi > 2, \\ 1 & \text{for } \psi = 2, \\ 0 & \text{otherwise.} \end{cases} \quad (3)$$

که در آن  $H(i)$  عدد هارمونیک است و می توان آن

را با  $0.5772156649 + \ln(i)$  (ثابت اویلر) تخمین زد. از

آنجا که  $c(i)$  میانگین  $h(x)$  داده شده است، می توان از آن

برای نرمال سازی  $h(x)$  استفاده کرد. امتیاز ناهنجاری یک

نمونه  $X$  به صورت زیر تعریف می شود (Liu et al., 2012):

$$s(x, \psi) = 2 \frac{E(h(x))}{c(\psi)} \quad (4)$$

که در آن  $E(h(x))$  میانگین  $h(x)$  از مجموع درخت ایزوله است. با

استفاده از امتیاز ناهنجاری  $s$ ، می توان ارزیابی زیر را انجام داد:

- اگر مقدار  $s$  برای نمونه ها بسیار نزدیک به ۱ باشند، قطعاً آن ها

ناهنجاری هستند؛

- اگر مقدار  $s$  بسیار کوچک تر از  $0.5$  باشند، قطعاً عادی هستند؛ و

- اگر برای همه نمونه ها مقدار تقریباً  $0.5$  را برگرداند، در کل

نمونه ها هیچ ناهنجاری مشخصی وجود ندارد.

به عنوان نمونه داده ای تعریف می کنیم که توسط فرایندی تولید

می شود که با فرایند تولید نمونه های داده نرمال متفاوت است.

ناهنجاری ها معمولاً درصد کمی (معمولاً یک درصد یا پایین تر) از

کل داده ها را تشکیل می دهند (Das et al., 2016).

برای کاهش ریسک استفاده از یک مدل تشخیص ناهنجاری،

متخصصان ترجیح می دهند مجموعه ای از مدل های تشخیص ناهنجاری

پایه با ابرپارامترهای مختلف بسازند؛ به عنوان مثال، الگوریتم های

مختلف با پارامترهای متفاوت و ترکیبی متمایز از ویژگی های

مجموعه داده (Aggarwal and Sathe, 2017). رویکرد تشخیص

ناهنجاری گروهی، که چندین الگوریتم پایه را در تشخیص ناهنجاری

ترکیب می کند، به عنوان یک استراتژی برای بهبود دقت و پایداری

مدل در نظر گرفته می شود، زیرا می توان با اجرای چندین بار مدل،

تأثیر واریانس در دقت مدل سازی را کاهش داد (Aggarwal, 2013)

در ادامه، الگوریتم های پایه و الگوریتم ترکیبی مورد استفاده در

تحقیق، مورد بررسی قرار می گیرند.

#### الگوریتم های پایه تشخیص ناهنجاری

در این مطالعه از دو الگوریتم که در تشخیص ناهنجاری کاربرد

دارند به عنوان الگوریتم های پایه مدل ترکیبی استفاده شده است.

الگوریتم ها طوری انتخاب شدند تا بتوان ناهنجاری های مجموعه داده

را براساس روش های مختلف تحلیل کرد. این الگوریتم ها شامل

جنگل ایزوله و ماشین بردار پشتیبان تک کلاسه می باشند.

#### الگوریتم جنگل ایزوله

الگوریتم جنگل ایزوله یک روش مبتنی بر درخت

تصمیم (Liu et al., 2008) و فاصله (Ruff et al., 2023)

است که برای اولین بار توسط (Liu et al., 2008)

ارائه گردید. ایده اصلی آن ها برای طراحی این الگوریتم، وجود

دو ویژگی متداول در داده های ناهنجار بود: (۱) آن ها در اقلیت

هستند و از تعداد نمونه های کمتری تشکیل شده اند و (۲) مقادیر

نمونه های ناهنجار بسیار متفاوت از نمونه های عادی هستند. این

الگوریتم مجموعه ای از درختان ایزوله را برای یک مجموعه داده

می سازد، سپس نمونه هایی را که طول مسیر متوسط کوتاهی در

درختان ایزوله دارند به عنوان نمونه های ناهنجار شناسایی می کند.

در این روش تنها دو متغیر وجود دارد: تعداد درخت برای ساخت

و اندازه زیرنمونه. این الگوریتم رویکرد متفاوتی نسبت به سایر

الگوریتم ها ارائه می دهد و آن شناسایی ناهنجاری ها با جداسازی

نمونه ها بدون تکیه بر فاصله و چگالی می باشد (Liu et al., 2008).

در الگوریتم جنگل ایزوله، درخت ایزوله به صورت زیر ساخته

می شود (Liu et al., 2012):

فرض کنید مجموعه داده  $X = \{x_1, \dots, x_n\}$  را داشته باشیم:

۱- اگر  $|X| \leq 1$  باشد یا ارتفاع درخت به یک حد از پیش تعریف

شده  $l$  رسیده باشد یک گره برگ ساخته شود.

۲- در غیر این صورت:

الف: انتخاب تصادفی یک ویژگی  $q$  از میان تمامی ویژگی های



می باشد. معادله درجه دوم را می توان با استفاده از بهینه ساز متوالی کمینه حل کرد. تابع تصمیم به صورت زیر خواهد بود:

$$g(x) = \text{sign}((w, \rho(x)) - \rho) \quad (6)$$

که مشخص می کند نقطه  $x$  در داخل (مثبت) یا خارج (منفی) از مجموعه تخمین زده شده قرار گرفته است. هسته گاوسی تنها هسته ای است که با موفقیت در ماشین بردار پشتیبان تک کلاسه اعمال شده (Schölkopf et al., 2001) و به صورت زیر می باشد:

$$k(x, y) = \exp\left(-\frac{\|x - y\|^2}{2\sigma^2}\right) \quad (7)$$

به این هسته، هسته تابع پایه شعاعی نیز گفته می شود.

#### الگوریتم تشخیص ناهنجاری ترکیبی

همان طور که در بخش های قبلی به آن اشاره شد، برای جبران ماهیت ناپایدار الگوریتم های بدون نظارت، متخصصان حوزه های پرمخاطره مانند مالی، سلامت و امنیت، ترجیح می دهند گروهی از مدل ها را برای ترکیب و تحلیل بسازند (Zhao et al., 2021). در این مطالعه از الگوریتم تشخیص ناهنجاری «ترکیب انتخابی محلی در مجموعه های پرت موازی» (Locally Selective Combination in Parallel Outlier Ensembles) استفاده شده است. این الگوریتم به عنوان یک چارچوب کلی، با انواع مختلف الگوریتم های تشخیص دهنده پایه سازگار می باشد (Zhao et al., 2019). یکی از دلایل اصلی کاهش عملکرد الگوریتم های تشخیص ناهنجاری بدون نظارت، وجود ساختار داده محلی می باشد. همچنین فقدان مقادیر برچسب در رویکرد بدون نظارت، ترکیب تشخیص دهنده های پایه را با چالش جدی مواجه کرده است (Zhao et al., 2019). این الگوریتم سعی می کند مشکل را با شناسایی مناطق محلی به دست آمده از نزدیکترین همسایه خود و ایجاد تشخیص دهنده رقابتی برای هر منطقه محلی حل کرده و نتایج قوی تری را ارائه دهد.

همان طور که در شکل ۲ نشان داده شده است، این الگوریتم از چهار مرحله اصلی تشکیل شده است. در مرحله اول، برچسب های شبه هدف از مجموعه داده آموزشی تولید می شود. فرض کنید  $X_{train}$  داده های آموزشی باشد و  $C = \{C_1, C_2, \dots, C_n\}$  مجموعه ای از تشخیص دهنده های پایه با ابرپارامترهای متفاوت باشد. همچنین، فرض کنید  $O(X_{train}) = [C_1(X_{train}), \dots, C_n(X_{train})]$  ماتریس امتیاز ناهنجاری باشد. سپس، شبه هدف که با  $target$  نشان داده می شود، با تجمیع (Aggregate) امتیاز ناهنجاری تشخیص دهنده های پایه  $C$  به یکی از روش های میانگین گرفتن یا حداکثر کردن  $O(X_{train})$ ، به دست می آید:

$$target = \phi(O(X_{train})) \in R^{n \times I} \quad (8)$$

برچسب شبه هدف در این الگوریتم با استفاده از داده های

ماشین بردار پشتیبان تک کلاسه

(Schölkopf et al., 2001) و (Muller et al., 2018)

یک روش ماشین بردار پشتیبانی ارائه کرده اند که به عنوان طبقه بندی تک کلاسه شناخته می شود. طبقه بندی تک کلاسه یک مسئله یادگیری بدون نظارت است (Muller et al., 2018) که طبقه بندی کننده مبتنی بر هسته با حداکثر حاشیه از مبدأ در فضای ویژگی پیدا می کند (Schölkopf et al., 2001) و (Muller et al., 2018) و می تواند برای یک مشکل تشخیص ناهنجاری یا پرت استفاده شود (Muller et al., 2018). ماشین بردار پشتیبان تک کلاسه در تشخیص ناهنجاری مجموعه داده های چندوجهی، غیرخطی و غیرمحدب (Barbado et al., 2022) و نامتوازن (Ertekin et al., 2007) مناسب است.

طبق گفته (Schölkopf et al., 1999) مراحل کشف ناهنجاری

در ماشین بردار پشتیبان تک کلاسه به صورت زیر می باشد:

- ۱- نگاشت نقاط به فضای با ابعاد بالاتر؛
- ۲- جداسازی تمام نقاط داده از مبدأ در فضای ویژگی با استفاده از ابرصفحه؛
- ۳- استفاده از پارامتر  $v$  به عنوان کسری از نقاط ناهنجار در داده ها؛
- ۴- به حداکثر رساندن فاصله بین ابرصفحه و مبدأ؛
- ۵- شناسایی نقاط زیر ابرصفحه که نزدیک به مبدأ هستند به عنوان نقاط پرت.

ماشین بردار پشتیبان تک کلاسه همان طور که در (Schölkopf

et al., 2001) فرموله شده، مجموعه ای را تخمین می زند که بیشترین نمونه داده شده از  $m$  نقطه  $\{x_i\}_{i=1}^m, x_i \in \mathbb{R}^d$  را شامل شود. هر نقطه  $x_i$  توسط یک نگاشت  $\varphi: \mathbb{R}^d \rightarrow \mathbb{R}^d$  از فضای ویژگی  $\mathbb{R}^d$  به فضای هسته با ابعاد بالای  $K$  که توسط هسته  $k(x, y)$  تولید می شود، تبدیل می شود. هسته از طریق  $k(x, y) = (\varphi(x), \varphi(y))$  به یک نمونه داخلی در فضای هسته مرتبط می شود. ماشین بردار پشتیبان تک کلاسه یک ابرصفحه را در فضای هسته پیدا می کند که داده ها را از مبدأ با حداکثر حاشیه جدا می کند. اگر چنین ابرصفحه ای وجود نداشته باشد، متغیرهای سست  $\xi_i$  (Slack) اجازه می دهند که برخی از نقاط در حاشیه (پرت) باشند و مقدار آن توسط پارامتر  $v \in [0, 1]$  کنترل می شود. به طور کلی،  $v$  یک کران بالا در کسری از نقاط پرت مشخص می کند. ابرصفحه در فضای هسته یک سطح غیرخطی در فضای ویژگی ایجاد می کند. بطور دقیق تر اگر بیان شود، ماشین بردار پشتیبان تک کلاسه، معادله درجه دوم زیر را حل می کند:

$$\min_{w, \xi, \rho} \frac{1}{2} \|w\|^2 + \frac{1}{v} \sum_{i=1}^m \xi_i - \rho \quad (9)$$

$$s. t. (w, \phi(x_i)) \geq \rho - \xi_i, \xi_i \geq 0, i = 1, 2, \dots, m$$

که در آن  $w$  بردار نرمال به ابرصفحه بوده و  $\rho$  به عنوان حاشیه

بین هر تشخیص‌دهنده پایه و شبه هدف در منطقه محلی قابل محاسبه می‌باشد. همبستگی پیرسون بین  $O(\Psi_j)$  و  $target^{\Psi_j}$  اعمال می‌شود.

در مرحله آخر، یک هیستوگرام از امتیاز همبستگی پیرسون برای هر یک از تشخیص‌دهنده‌ها ساخته شده و با فواصل مساوی  $b$  درج می‌شود. سپس مجموعه‌ای از تشخیص‌دهنده‌های مربوط به پرتکرارترین فواصل برای ترکیب در مرحله بعدی نگهداری می‌شود. چارچوب کلی الگوریتم در شکل ۲ نشان داده شده است.

### نتایج و بحث

در این بخش به شرح مدل سازی کشف ناهنجاری و همچنین استخراج نتایج از آن پرداخته می‌شود. برای ایجاد یک مدل ترکیبی تشخیص ناهنجاری از دو الگوریتم جنگل ایزوله و ماشین بردار پشتیبان تک کلاسه به عنوان مدل‌های پایه استفاده شده است. برای افزایش قدرت تشخیص‌دهنده ترکیبی نیاز است تا انواع تشخیص‌دهنده‌های پایه مختلف با پارامترهای متفاوت تنظیم گردد. بنابراین، هر کدام از این الگوریتم‌های پایه تشخیص با ۶ ابرپارامتر مختلف تنظیم شده و در مجموع ۱۲ تشخیص‌دهنده به‌عنوان تشخیص‌دهنده‌های پایه به عنوان ورودی به الگوریتم داده می‌شود. برای تعیین ابرپارامتر تعداد درخت در جنگل ایزوله، ۶ عدد به‌صورت تصادفی در بازه ۱۰ الی ۲۰۰ لحاظ شده است. هسته ماشین بردار پشتیبان از نوع تابع پایه شعاعی انتخاب شد. پارامتر نو (V) که هم یک کران پایین برای تعداد نمونه‌هایی است که بردارهای پشتیبان هستند و هم یک کران بالا برای تعداد نمونه‌هایی که در سمت اشتباه ابرصفحه قرار دارند و معمولاً معادل با میزان آلودگی مجموعه در نظر گرفته می‌شود. با مشورت کارشناسان صندوق بیمه ۶ عدد برای ابرپارامتر میزان آلودگی یا ناهنجاری احتمالی در مجموعه داده در بازه ۱ الی ۲ لحاظ گردید. پارامتر گاما به عنوان ضریب هسته در این الگوریتم، برای تنظیم میزان تأثیر نمونه آموزشی مورد استفاده قرار می‌گیرد. هرچه این پارامتر بزرگتر باشد، نقاط نزدیک به خط جداکننده انتخاب می‌شوند و هرچه کوچکتر باشد، نقاط دورتر انتخاب می‌شوند. مقدار پیش فرض برای این پارامتر در الگوریتم

آموزشی تولید می‌شود و صرفاً برای انتخاب تشخیص‌دهنده استفاده می‌شود.

در مرحله دوم، منطقه محلی ساخته می‌شود. با توجه به یک نمونه‌داده برای آزمایش  $X_j$  منطقه محلی  $\Psi_j$  به عنوان ک- نزدیکترین همسایه آن به‌صورت زیر تعریف می‌شود:

$$\Psi_j = \{X_i | X_i \in X_{train}, X_i \in L_{K_{ENS}}(X_j)\} \quad (9)$$

برای تعریف منطقه محلی  $L_{K_{ENS}}(X_j)$ ،  $t$  گروه از ویژگی‌های  $[\frac{d}{2}, d]$  به‌صورت تصادفی برای ساخت فضای ویژگی جدید انتخاب می‌شود. سپس ک- نزدیک‌ترین همسایه داده شناسایی شده و وقتی  $X_j$  بیش از  $\frac{t}{2}$  بار در همسایگی ظاهر شود به  $L_{K_{ENS}}(X_j)$  اضافه شده و منطقه محلی تعریف می‌شود. اندازه منطقه محلی مشخص نیست، زیرا به تعداد داده‌های آموزشی که معیارهای انتخاب را دارند بستگی دارد. مقدار تعداد همسایه در مسائل بدون نظارت که برچسب ندارند و در آن نمی‌توان از اعتبارسنجی متقابل استفاده کرد، در بازه [۱۰۰، ۳۰۰] پیشنهاد شده است.

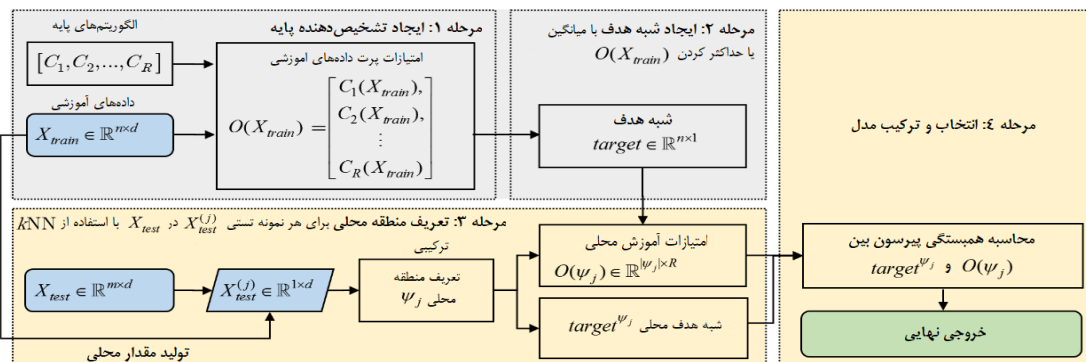
در مرحله سوم، انتخاب مدل و ترکیب انجام می‌شود. برای هر نمونه آموزشی، شبه هدف محلی  $Z$  را می‌توان با بازیابی مقادیر مرتبط با منطقه محلی  $Z$  از هدف به دست آورد:

$$target^{\Psi_j} = \{target_{x_i} | x_i \in \Psi_j\} \in \mathbb{R}^{|\Psi_j| \times 1} \quad (10)$$

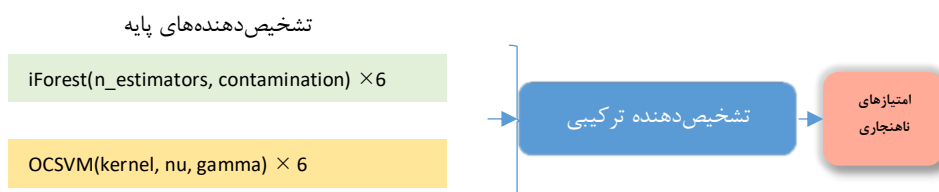
به‌طور مشابه، امتیاز ناهنجاری آموزش محلی  $O(\Psi_j)$  را می‌توان از ماتریس امتیاز ناهنجاری آموزشی از پیش محاسبه شده  $O(X_{train})$  به دست آورد:

$$O(\Psi_j) = [C1(\Psi_j), \dots, Cr(\Psi_j)] \in \mathbb{R}^{|\Psi_j| \times R} \quad (11)$$

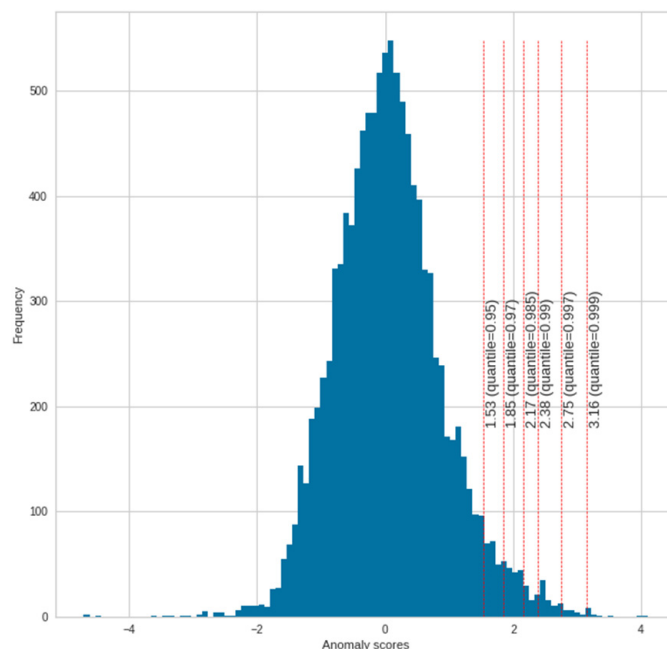
همچنین، اگر  $O(\Psi_j)$  ماتریس امتیاز آموزشی بوده و از ماتریس امتیاز ناهنجاری به دست آمده باشد، همبستگی



شکل ۲: چارچوب کلی الگوریتم «ترکیب انتخابی محلی در مجموعه‌های پرت موازی» (Zhao et al., 2019)



شکل ۳: دیاگرام کلی تشخیص ناهنجاری ترکیبی



شکل ۴: هیستوگرام امتیازات ناهنجاری گندم آبی

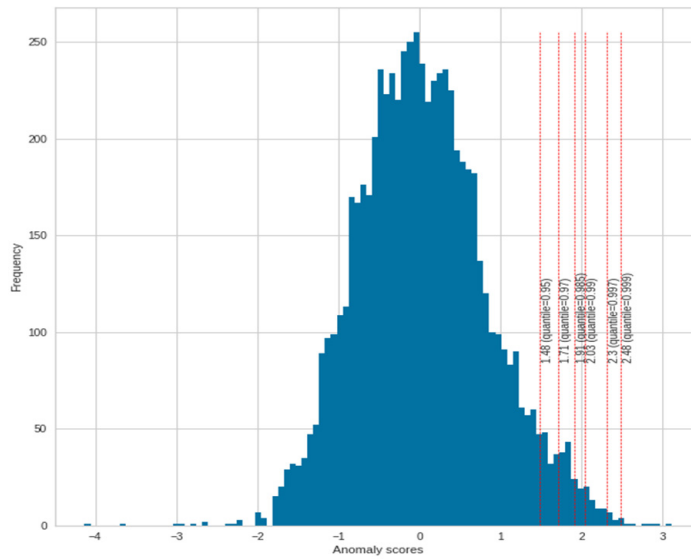
جدول ۲: تعداد ناهنجاری ها در آستانه های مختلف (گندم آبی)

تعداد ناهنجاری	مقدار آستانه
۵۶۵	۰/۹۵
۳۳۹	۰/۹۷
۱۷۰	۰/۹۸۵
۱۱۳	۰/۹۹
۳۴	۰/۹۹۷
۱۲	۰/۹۹۹

تهیه شده است. از امکانات گوگل برای تحقیقات علوم داده (Colab) استفاده شد. در این محیط منابع قابل دسترس شامل ۱۳ گیگابایت حافظه و پردازنده دو هسته ای ۲/۲ گیگاهرتزی بود. شکل ۳ شمای کلی مدل پیاده سازی شده برای کشف ناهنجاری را نشان می دهد. نمودار هیستوگرام امتیازهای ناهنجاری برای بیمه نامه های گندم آبی استان خوزستان حاصل از اجرای الگوریتم «ترکیب انتخابی محلی در مجموعه های پرت موازی» در شکل ۴ نشان داده شده است. خطوط نقطه چین عمودی قرمز رنگ، آستانه های ناهنجاری

<sup>1</sup>تنظیم شده است. برای این پارامتر نیز ۶ مقدار تصادفی بین تعداد ویژگی ها ۰/۰۱ الی ۱۰۰ در نظر گرفته شد.

مراحل آماده سازی داده ها در محیط های اکسل، سیستم مدیریت پایگاه داده میکروسافت (MS-SQL) و پایتون نسخه ۳/۱۰ انجام گرفت. از جعبه ابزار PyOD که توسط Zhao et al. (2019) معرفی شده، برای پیاده سازی مدل استفاده گردید. این جعبه ابزار شامل الگوریتم های کشف ناهنجاری است که برای محیط پایتون



شکل ۵: هیستوگرام امتیازات ناهنجاری گندم دیم

جدول ۳: تعداد ناهنجاری ها در آستانه های مختلف (گندم دیم)

تعداد موارد ناهنجار	مقدار آستانه
۳۳۷	۰/۹۵
۲۰۲	۰/۹۷
۹۷	۰/۹۸۵
۶۸	۰/۹۹
۲۱	۰/۹۹۷
۷	۰/۹۹۹

بررسی صورت گرفته شده، مشخص گردید که ناهنجاری های شناسایی شده توسط مدل در ۵ دسته قرار می گیرند که لیست آنها در **جدول ۴** ارائه شده است.

در **جدول ۵** نتایج حاصل از بررسی هر یک از بیمه نامه ها ارائه شده است. همان طور که مشاهده می شود، از مجموع ۱۱،۲۹۳ بیمه نامه گندم آبی که منجر به دریافت خسارت شده، تعداد ۱۷۰ مورد به عنوان ناهنجاری تشخیص داده شده است. در بررسی صورت گرفته مشخص گردید از این تعداد ناهنجاری تعداد ۴۸ مورد ادعاهای خسارت نرمال بوده اند که به عنوان غیرنرمال تشخیص داده شده اند. همچنین از مجموع ۶،۷۲۷ بیمه نامه گندم دیم که منجر به دریافت خسارت شده، تعداد ۹۷ مورد به عنوان ناهنجاری تشخیص داده شده است که در بررسی صورت گرفته مشخص گردید از این تعداد ناهنجاری تعداد ۳۱ مورد ادعاهای خسارت نرمال بوده اند که به عنوان غیرنرمال تشخیص داده است.

همان طور که مشاهده می شود، فراوانی ناهنجاری ها در بیمه نامه گندم آبی و دیم متفاوت از همدیگر هستند. این تفاوت ناشی از نوع کشت و تفاوت در خسارت های وارده می باشد. برای مثال، برای کشت دیم عامل خشکسالی از مهمترین عوامل دریافت خسارت است ولی در کشت آبی، طوفان و تگرگ مهم ترین عامل دریافت خسارت می باشد.

را با مقادیر صدک متفاوت نشان می دهد. در **جدول ۲** تعداد موارد ناهنجار در صدک های متفاوت آورده شده است. همان طور که مشاهده می شود، در صدک ۰/۹۸۵ تعداد موارد ناهنجار برابر ۱۷۰ مورد می باشد که ۱/۵ درصد از مجموع داده های بیمه نامه گندم آبی است و با مقدار آلودگی که کارشناسان پیش بینی می کردند مطابق دارد. بنابراین، این آستانه برای ناهنجاری انتخاب می شود.

برای داده های گندم دیم نیز نمودار هیستوگرام امتیازهای ناهنجاری در **شکل ۵** و اطلاعات ناهنجاری و آستانه ها در **جدول ۳** نشان داده شده است:

همان طور که مشاهده می شود، در **جدول ۳** برای آستانه ۰/۹۸۵ تعداد ۹۷ ناهنجاری تشخیص داده شده که حدود ۱/۵ درصد از مجموع داده های بیمه گندم دیم که ۶،۷۲۷ مورد بود، می باشد.

چالش اصلی در یادگیری بدون نظارت، ارزیابی نتایج آن است، زیرا معمولاً روی داده هایی اعمال می شوند که دارای برچسب نیستند؛ لذا برای پوشش این مسئله از کارشناسان خبره آن حوزه در ارزیابی نتایج بهره گرفته می شود. بنابراین، در قدم بعدی بیمه نامه هایی که به عنوان موارد غیرنرمال یا ناهنجار شناسایی شده بودند توسط گروهی از کارشناسان حوزه بازرسی صندوق بیمه کشاورزی مورد بررسی قرار گرفت تا صحت هر کدام مورد سنجش قرار گیرد. در

جدول ۴: دسته بندی ناهنجاری های تشخیص داده شده

عنوان دسته بندی	شرح ناهنجاری	توضیح
D <sup>۱</sup>	عدم وجود مستندات کافی	بدین معنی است که مستندات لازم که می بایست طبق روش های اجرایی در سامانه بارگذاری شود، موجود نمی باشد و یا برخی از آنها بارگذاری نشده است.
D <sup>۲</sup>	مستندات گویای وقوع خسارت اعلامی نمی باشد	مستندات بارگذاری شده در سامانه طبق دستورالعمل مربوطه گویای بروز نوع خسارت ثبت شده نمی باشد. به عنوان مثال، سرعت در خسارت طوفان ۶۰ کیلومتر بر ساعت ذکر شده، ولی در مدارک هواشناسی ۲۰ کیلومتر بر ساعت می باشد.
D <sup>۳</sup>	عدم مطابقت مستندات با واقعیت	در برخی از مستندات عامل خسارت در فرم کارشناسی خسارت خشکسالی عنوان گردیده، ولی عکس خسارت گویای خسارت سیل می باشد.
T <sup>۱</sup>	عدم رعایت بازه زمانی خسارت	طبق دستورالعمل اجرایی، مهلت زمان اعلام خسارت تا زمان واریز غرامت به مدت یک ماه می باشد. خارج از آن مغایر با دستورالعمل می باشد.
T <sup>۲</sup>	عدم مطابقت تاریخ وقوع خسارت با زمان اعلام خسارت	طبق دستورالعمل های اجرایی صندوق بیمه، در خسارت زراعت می بایست یک هفته پس از بروز خسارت بازدید انجام شود؛ تا قبل از حذف آثار خسارت نوع و میزان آن به صورت دقیق بررسی شود.

جدول ۵: نتایج بررسی موارد ناهنجار توسط کارشناسان

نوع بیمه نامه	تعداد کل نمونه ها	تعداد موارد ناهنجاری شناسایی شده	نوع ناهنجاری و تعداد آن					تعداد تشخیص اشتباه	میزان دقت
			D <sup>۱</sup>	D <sup>۲</sup>	D <sup>۳</sup>	T <sup>۱</sup>	T <sup>۲</sup>		
گندم آبی	۱۱،۲۹۳	۱۷۰	۷۶	۱۰	۸	۱۲	۱۶	۴۸	۰/۷۲
گندم دیم	۶،۷۲۷	۹۷	۲۶	۳	۴	۲۸	۵	۳۱	۰/۶۸

### جمع بندی و پیشنهادها

با توجه به هدف پژوهش که استفاده از تکنیک های یادگیری ماشین بدون نظارت برای تشخیص ادعاهای دریافت خسارت غیرواقعی بود؛ نتایج به دست آمده حاکی از وجود ۵ دسته رفتار غیرنرمال در پرداخت خسارت می باشد. بنابراین، با اتکا به این نتایج می توان پاسخ سؤالات پژوهش را این گونه داد.

در پاسخ به سؤال اول که در فرایند پرداخت خسارت بیمه محصول گندم (دیم یا آبی) چه میزان ادعاهای مشکوک وجود دارد؟ می توان گفت که حدود ۱/۵ درصد از مجموع داده ها را موارد مشکوک یا غیرنرمال یا به عبارت دیگر ادعاهای خسارت غیرواقعی تشکیل می دهند. موارد مشکوک یافت شده برای محصول گندم آبی ۱۷۰ مورد و برای گندم دیم ۹۷ مورد بود. این موارد می تواند ناشی از رفتارهای نادرست از طرف ارزیاب، نماینده بیمه گر یا از طرف بیمه گذار باشد. روش های کلاهبرداری که در نتیجه بررسی موارد غیرنرمال توسط کارشناسان به دست آمد را می توان به صورت زیر در ۵ نوع دسته بندی کرد:

- ۱- عدم وجود مستندات کافی مبنی بر وقوع خسارت؛
  - ۲- مستندات گویای وقوع خسارت اعلامی نمی باشد؛
  - ۳- عدم مطابقت مستندات با واقعیت؛
  - ۴- عدم رعایت بازه زمانی خسارت؛
  - ۵- عدم مطابقت تاریخ وقوع خسارت با زمان اعلام خسارت.
- در پاسخ به سؤال دوم که چگونه می توان از تکنیک های یادگیری ماشین برای تشخیص ادعاهای مشکوک خسارت در بیمه

گندم استفاده کرد؟ می توان این گونه پاسخ داد که با استفاده از الگوریتم های یادگیری ماشین بدون نظارت که به صورت ترکیبی مورد استفاده قرار گرفتند، مدل ساخته شده قادر به تشخیص ۷۲ درصد موارد مشکوک بیمه نامه های گندم آبی و ۶۸ درصد موارد مشکوک بیمه نامه های گندم دیم می باشد. البته ذکر این نکته حائز اهمیت است که وقتی داده های مورد مطالعه دارای برچسب نیستند، ساخت مدل و ارزیابی دقت آن با چالش جدی مواجه است، لذا برای رفع این مسئله از کارشناسان و خبرگان بهره گرفته می شود. در این تحقیق نیز با توجه به اینکه داده ها دارای برچسب نبودند، ارزیابی نتایج با همکاری کارشناسان حوزه بازرسی صندوق بیمه کشاورزی، نتایج به صورت دقیق مورد بررسی قرار گرفتند. با توجه به حساسیت الگوریتم های یادگیری ماشین به تنظیمات ابرپارامترها، استفاده از تخمین گرهای مختلف با تنظیمات متفاوت ابرپارامترها و ترکیب نتایج آنها مدل های باثبات تری را ارائه می دهند. در ترکیب نتایج از روش های مختلفی از قبیل رأی اکثریت، آدابوست، بسته بندی می توان استفاده کرد. همچنین کمک گرفتن از کارشناسان خبره آن حوزه نیز در مواردی که داده ها برچسب ندارند، می تواند تأثیر بسزایی در ارزیابی نتایج به همراه داشته باشد.

پیشنهادهای زیر براساس تحقیق حاضر می تواند در بهبود روند تشخیص ادعاهای مشکوک خسارت مؤثر باشد:

- ۱- با وجود محدودیت های ذکر شده، رویکرد بدون نظارت ارائه شده در این تحقیق می تواند با سامانه های موجود صندوق بیمه کشاورزی ادغام شود و برای شناسایی ادعاهای مشکوک به صورت

### مشارکت نویسندگان

یعقوب احمدلو: جمع آوری مطالعات مرتبط و تدوین مدل، علیرضا پورابراهیمی: کنترل چهارچوب تدوین و استانداردهای پژوهشی، جعفر تنها: ارزیابی مدل ها و نتیجه گیری، یعقوب احمدلو: مروری بر ادبیات پژوهش، علی رجب زاده قطری: روش پژوهش و متدولوژی.

### تشکر و قدردانی

این تحقیق با حمایت مالی صندوق بیمه کشاورزی ایران انجام گرفته و بدین وسیله بابت این حمایت ها و همکاری های انجام شده در ارائه داده ها و ارزیابی نتایج توسط کارشناسان حوزه بازرسی، تقدیر و تشکر می گردد.

### تعارض منافع

نویسندگان اعلام می دارند که در مورد انتشار این مقاله تضاد منافع وجود ندارد. علاوه بر این، موضوعات اخلاقی شامل سرقت ادبی، رضایت آگاهانه، سوءرفتار، جعل داده ها، انتشار و ارسال مجدد و مکرر توسط نویسندگان رعایت شده است.

### دسترسی آزاد

کپی رایت نویسنده (ها) ©2023: این مقاله تحت مجوز بین المللی Creative Commons Attribution 4.0 اجازه استفاده، اشتراک گذاری، اقتباس، توزیع و تکثیر را در هر رسانه یا قالبی مشروط به درج نحوه دقیق دسترسی به مجوز CC منوط به ذکر تغییرات احتمالی بر روی مقاله می باشد. لذا به استناد مجوز مذکور، درج هرگونه تغییرات در تصاویر، منابع و ارجاعات یا سایر مطالب از اشخاص ثالث در این مقاله باید در این مجوز گنجانده شود، مگر اینکه در راستای اعتبار مقاله به اشکال دیگری مشخص شده باشد. در صورت عدم درج مطالب مذکور و یا استفاده فراتر از مجوز فوق، نویسنده ملزم به دریافت مجوز حق نسخه برداری از شخص ثالث می باشد.

به منظور مشاهده مجوز بین المللی Creative Commons Attribution 4.0 به آدرس زیر مراجعه گردد:  
<https://creativecommons.org/licenses/by/4.0>

### منابع

- Aggarwal, C.C., (2013). Outlier ensembles: Position paper. ACM SIGKDD explorations newsletter., 14(2): 49-58 (10 Pages).
- Aggarwal, C.C.; Sahte, S., (2017). Outlier ensembles: An introduction. Springer.
- The World Bank (2022). Agriculture & rural development.
- Barbado, A.; Corcho, Ó.; Benjamins, R., (2022). Rule extraction in unsupervised anomaly detection for model explainability: Application to OneClass SVM. Expert Syst. Appl., 189(1): 1-20 (20 Pages).
- Bauder, R.; Da Rosa, R.; Khoshgoftaar, T., (2018). Identifying medicare provider fraud with unsupervised machine learning. In 2018 IEEE international conference on information reuse and integration (IRI), 285-292 (8 Pages).
- Bose, I.; Mahapatra, R.K., (2001). Business data mining —A machine learning perspective. Inf. Manage., 39(3): 211-225

غیربرخط مورد استفاده قرار گیرد.

۲- پیشنهاد می شود به منظور بهبود نتایج تجزیه و تحلیل، از دانشمندان داده برای انجام آماده سازی و پیش پردازش دقیق تر داده ها بهره گرفته شود.

۳- با ایجاد داده های برچسب دار برای تعدادی از داده ها در کنار انبوهی از داده های بدون برچسب، از روش های یادگیری ماشین نیمه نظارت شده بهره برداری شود، چراکه در بهبود نتیجه تأثیر مثبتی خواهد گذاشت.

۴- از آنجا که درصد زیادی از موارد غیرنرمال را دسته اول، یعنی عدم ارائه مستندات کافی تشکیل می دهد؛ لذا این موضوع می تواند ناشی از ارائه بیمه نامه های صوری و دریافت خسارت بدون ارائه مستندات و با تبانی بیمه گذار، ارزیاب و نماینده بیمه گر صورت گرفته باشد. لذا ضروری است اصالت زمین های مزروعی که بیمه می شوند توسط کارشناسان صندوق حین صدور بیمه نامه مورد تأیید قرار بگیرند.

۵- با مشخص شدن حدود اربعه زمین های زراعی، مسئله تکراری بودن داده ها مرتفع می گردد؛ لذا پیشنهاد می گردد داده های مربوط به حدود اربعه زمین ها نیز در اختیار محققین قرار گیرد تا تجزیه و تحلیل دقیق تری روی داده ها صورت پذیرد.


انجام تحقیق حاضر با محدودیت هایی روبه رو بود. اول اینکه داده های موقعیت مکانی در دسترس نبود و در نتیجه زمین های زراعی بیمه شده که دارای مشخصات یکسانی بوده ولی در موقعیت های جغرافیایی متفاوتی بودند، به عنوان داده های تکراری مطرح شدند و از آنجا که داده های تکراری در آموزش مدل هیچ تأثیری ندارند، از مجموعه داده حذف شدند. همچنین عدم شناخت کافی صندوق بیمه کشاورزی از میزان دقیق موارد کلاهبرداری (ادعاهای ساختگی یا سایر فرایندهای پرداخت) موجب شد تا میزان ناهنجاری های به دست آمده متفاوت از مقدار واقعی آن باشد؛ لذا برای رسیدن به بالاترین میزان دقت تشخیص موارد غیرنرمال تحقیقات آینده، بایستی با لحاظ نمودن سایر اطلاعات از قبیل تعداد دفعات دریافت خسارت توسط بیمه گذار، تضمین اصالت و مختصات جغرافیایی زمین های زراعی و سایر ویژگی های کلیدی انجام شود.

### (15 Pages).

- Bouazza, I.; Ameer, E.B.; Ameer, F., (2019). Datamining for fraud detecting, state of the art. Adv. Intell. Syst. Sustainable Dev., 205-219 (15 Pages).
- Brownlee, J., (2020). Data preparation for machine learning.
- Chandola, V.; Banerjee, A.; Kumar, V., (2009). Anomaly detection: A survey. ACM Comput. Surv., 41(3): 1-58 (58 Pages).
- Clayton, C., (2022). FBI touts work on crop insurance fraud case involving kentucky tobacco farmers. Insur. Agents.
- Cole, S.A.; Xiong, W., (2017). Agricultural insurance and economic development. Annu. Rev. Econ., 9(1): 235-262 (28 Pages).
- Bertoni, D., (2006). Crop insurance more needs to be done to reduce program's vulnerability to fraud, waste, and abuse. U.S. Gov. Account. Office., 1-18 (18 Pages).
- Das, S.; Wong, W.K.; Dietterich, T.; Fern, A.; Emmott, A., (2016).

- Incorporating expert feedback into active anomaly discovery. In 2016 IEEE 16th international conference on data mining (ICDM), 853-858 **(6 Pages)**.
- Ertekin, S.; Huang, J.; Bottou, L.; Giles, L., (2007). Learning on the border: Active learning in imbalanced data classification. In Proceedings of the 16th ACM conference on conference on information and knowledge management., 127-136 **(10 Pages)**.
- Gill, K.M.; Woolley, A.; Gill, M., (2005). Insurance fraud: The business as a victim?. *Crime.*, 73-82 **(10 Pages)**.
- Gill, W., (2009). Fighting fraud with advanced analytics. *Can. Underwriter. Bus. Inf. Group.*, 76(9): 28-32 **(5 Pages)**.
- Goodarzi, A.; Janatbabaei, S., (2017). Evaluation of Three Data Mining Algorithms (Decision Tree, Naive Bayes, Logistic Regression) in Auto Insurance Fraud Detection. *Insur. Res.*, 1(2): 61-80 **(19 Pages)**. [In Persian]
- Gomes, C.; Jin, Z.; Yang, H., (2021). Insurance fraud detection with unsupervised deep learning. *J. Risk. Insur.*, 88(3): 591-624 **(34 Pages)**.
- Han, J.; Kamber, M.; Pei, J., (2011). *Data mining: Concepts and techniques*. Morgan Kaufmann publishers Inc., 1-740 **(740 Pages)**.
- Hassan, A.K.I.; Abraham, A., (2016). Modeling insurance fraud detection using ensemble combining classification. *Int. J. Comput. Inf. Syst. Indust. Manage. App.*, 8(1): 257-265 **(9 Pages)**.
- Hosseini, A.; Rezaei, A., (2019). Fraud detection and management strategies in the insurance organizations using data mining techniques. *Social. Secur. Res. Inst.*, 14(4): 111-136 **(26 Pages)**. [In Persian]
- Kim, D.; Lee, S.; Lee, J., (2020). An ensemble-based approach to anomaly detection in marine engine sensor streams for efficient condition monitoring and analysis. *Sensors.*, 20(24): 1-16 **(16 Pages)**.
- Kirlidog, M.; Asuk, C., (2012). A fraud detection approach with data mining in health insurance. *Procedia. Social. Behav. Sci.*, 62(1): 989-994 **(6 Pages)**.
- Li, J.; Huang, K.Y.; Jin, J.; Shi, J., (2008). A survey on statistical methods for health care fraud detection. *Health care. Manage. Sci.*, 11(1): 275-287 **(13 Pages)**.
- Liu, F.T.; Ting, K.M.; Zhou, Z.H., (2008). Isolation forest. In 2008 8th IEEE international conference on data mining., 413-422 **(10 Pages)**.
- Marzen, C.G., (2013). Crop insurance fraud and misrepresentations: Contemporary issues and potential remedies. *SSRN. Electron. J.*, 675-707 **(33 Pages)**.
- Mouret, F.; Albughdadi, M.; Duthoit, S.; Kouamé, D.; Rieu, G.; Tourneret, J.Y., (2021). Outlier detection at the parcel-level in wheat and rapeseed crops using multispectral and sar time series. *J. Remote Sens.*, 13(5): 1-25 **(25 Pages)**.
- Müller, K.R.; Mika, S.; Tsuda, K.; Schölkopf, K., (2018). An introduction to kernel-based learning algorithms. In *handbook of neural network signal processing.*, 1-4 **(4 Pages)**.
- Nassif, A.B.; Talib, M.A.; Nasir, Q.; Dakalbab, F.M., (2021). Machine learning for anomaly detection: A systematic review. *IEEE Access.*, 9(1): 78658-78700 **(43 Pages)**.
- Ngai, E.W.T.; Hu, Y.; Wong, Y.H.; Chen, Y.; Sun, X., (2011). The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature. *Decis. Support Syst.*, 50(3): 559-569 **(11 Pages)**.
- Randhawa, K.; Loo, C.K.; Seera, M.; Lim, C.P.; Nandi, A.K., (2018). Credit card fraud detection using adaboost and majority voting. *IEEE Access.*, 6(1): 14277-14284 **(8 Pages)**.
- Rejesus, R.M.; Little, B.B.; Lovell, A.C., (2004). Using data mining to detect crop insurance fraud: Is there a role for social scientists?. *J. Financ. Crime.*, 12(1): 24-32 **(9 Pages)**.
- Roholamin, M.; Paygozar, H.; Khalili Dermani, M.K., (2019). An overview of fraud detection methods in the insurance industry. Paper presented at the 3rd national conference on electrical and computer engineering., 1-10 **(10 Pages)**. [In Persian]
- Rostami, H.; MonazamiTabar, S., (2021). Insurance fraud in the compulsory insurance law (2016) (by looking at the american legal system). *Iran. J. Insur. Res.*, 10(3): 155-186 **(32 Pages)**. [In Persian]
- Ruff, L.; Kauffmann, J.R.; Vandermeulen, R.A.; Montavon, G.; Samek, W.; Kloft, M.; Dietterich, D.G.; Müller, K.R., (2021). A unifying review of deep and shallow anomaly detection. *Proc. IEEE.*, 109(5): 756-795 **(40 Pages)**.
- Sahni, S.; Mittal, A.; Kidwai, F.; Tiwari, A.; Khandelwal, K., (2020). Insurance fraud identification using computer vision and IOT: A study of field fires. *Procedia. Comput. Sci.*, 173(1): 56-63 **(8 Pages)**.
- Schölkopf, B.; Platt, J.C.; Shawe-Taylor, J.; Smola, A.J.; Williamson, R.C., (2001). Estimating the support of a high-dimensional distribution. *Neural Comput.*, 13(7): 1443-1471 **(29 Pages)**.
- Schölkopf, B.; Williamson, R.C.; Smola, A.; Shawe-Taylor, J.; Platt, J., (1999). Support vector method for novelty detection. *Adv. Neural. Inf. Process. Syst.*, 582-588 **(7 Pages)**.
- Soltani Halvaeie, N.; Akbari, M.K., (2014). A novel model for credit card fraud detection using artificial immune systems. *Appl. Soft Comput.*, 24(1): 40-49 **(10 Pages)**.
- Verma, A.; Taneja, A.; Arora, A., (2017). Fraud detection and frequent pattern matching in insurance claims using data mining techniques. Paper presented at the 2017 10th international conference on contemporary computing (IC3), 1-7 **(7 Pages)**.
- Zhao, Y.; Hu, X.; Cheng, C.; Wang, C.; Wan, C.; Wang, W.; Yang, J.; Bai, H.; Li, Z.; Liao, C.; Wang, Y.; Qiao, Z.; Sun, J.; Akoglu, L., (2021). Suod: accelerating large-scale unsupervised heterogeneous outlier detection. *Proc. Mach. Learn. Syst.*, 463-478 **(16 Pages)**.
- Zhao, Y.; Nasrullah, Z.; Hryniewicki, M.K.; Li, Z., (2019). LSCP: Locally selective combination in parallel outlier ensembles. *Proceedings of the 2019 SIAM international conference on data mining (SDM).*, 585-593 **(9 Pages)**.
- Zhao, Y.; Nasrullah, Z.; Li, Z., (2019). Pyod: A Python toolbox for scalable outlier detection. *J. Mach. Learn. Res.*, 1-7 **(7 Pages)**.

AUTHOR(S) BIOSKETCHES	معرفی نویسندگان
<ul style="list-style-type: none"><li>Email: <a href="mailto:Y.ahmadlou@srbiau.ac.ir">Y.ahmadlou@srbiau.ac.ir</a></li><li>ORCID: 0000-0002-3778-094X</li><li>Homepage: <a href="https://srb.iau.ir/management/fa">https://srb.iau.ir/management/fa</a></li></ul>	<p>یعقوب احمدلو، دانشجوی دکتری مدیریت فناوری اطلاعات، گروه مدیریت فناوری اطلاعات، دانشکده مدیریت و اقتصاد، واحد علوم و تحقیقات، دانشگاه آزاد اسلامی، تهران، ایران</p>
<ul style="list-style-type: none"><li>Email: <a href="mailto:A.pourebahimi@kiaou.ac.ir">A.pourebahimi@kiaou.ac.ir</a></li><li>ORCID: 0000-0001-5741-0260</li><li>Homepage: <a href="http://www.alirezapourebahimi.ir">http://www.alirezapourebahimi.ir</a></li></ul>	<p>علیرضا پوراابراهیمی، استادیار گروه مدیریت صنعتی، دانشکده مدیریت و حسابداری، واحد کرج، دانشگاه آزاد اسلامی، کرج، ایران</p>
<ul style="list-style-type: none"><li>Email: <a href="mailto:Tanha@tabrizu.ac.ir">Tanha@tabrizu.ac.ir</a></li><li>ORCID: 0000-0002-0779-6027</li><li>Homepage: <a href="https://asatid.tabrizu.ac.ir/fa/pages/default.aspx?tanha">https://asatid.tabrizu.ac.ir/fa/pages/default.aspx?tanha</a></li></ul>	<p>جعفر تنها، دانشیار گروه مهندسی فناوری اطلاعات، دانشکده مهندسی برق و کامپیوتر، دانشگاه تبریز، تبریز، ایران</p>
<ul style="list-style-type: none"><li>Email: <a href="mailto:Alirajabzadeh@modares.ac.ir">Alirajabzadeh@modares.ac.ir</a></li><li>ORCID: 0000-0002-8470-3568</li><li>Homepage: <a href="https://www.modares.ac.ir/pro/academic_staff/alirajabzadeh">https://www.modares.ac.ir/pro/academic_staff/alirajabzadeh</a></li></ul>	<p>علی رجبزاده، استاد گروه مدیریت صنعتی، دانشکده مدیریت و اقتصاد، دانشگاه تربیت مدرس، تهران، ایران</p>

<p><b>HOW TO CITE THIS ARTICLE</b></p> <p>Ahmadlou, Y.; Pourebahimi, A.; Tanha, J.; Rajabzadeh, A., (2023). An ensemble based model for detecting suspicious claims in crop insuranc. <i>Iran. J. Insur. Res.</i>, 12(1): 63-78.</p> <p>DOI: <a href="https://doi.org/10.22056/ijir.2023.01.06">10.22056/ijir.2023.01.06</a></p> <p>URL: <a href="https://ijir.irc.ac.ir/article_155353.html?lang=en">https://ijir.irc.ac.ir/article_155353.html?lang=en</a></p>	
--	--