Business Intelligence Management Studies
Vol. 12, No.45, Fall 2023; P 313-346
ims.atu.ac.ir
DOI: 10.22054/ims.2023.69377.2213

ATU PRESS

Original Research
Accepted: 17 October 2023
Revised: 17 May 2023
Received: 09 August 2022
EISSN: 2821-0816
ISSN: 2821-0964

A Model for Detecting Abnormal Claims in Agricultural Insurance Using Deep Learning

Yaqub Ahmadlou — Ph.D. Student of Information Technology Management, Science and Research Branch, Islamic Azad University, Tehran, Iran.

Alireza Pourebrahimi * — Assistant Professor of Department of Industrial Management, Karaj Branch, Islamic Azad University, Karaj, Iran.

Jafar Tanha — Associate Professor of Department of Information Technology Engineering, Faculty of Electrical and Computer Engineering, University of Tabriz, Tabriz, Iran.

Ali Rajabzade Ghatari — Professor of Department of Industrial Management, Faculty of Management and Economics, Tarbiat Modares University, Tehran, Iran.

Abstract

Fraud cases have increased in recent years, especially in important and sensitive financial and insurance fields. Therefore, to deal with such frauds, there is a need for different measures than traditional inspection methods. Agricultural insurance is also not exempted from this threat due to its nature and wide extent and every year a lot of money is spent on paying fake damages. This research was presented with the aim of providing a model to discover unrealistic damage claims in agricultural insurance by using data mining and machine learning techniques. It was used to build a deep learning model. The data used was obtained from the Agricultural Insurance Fund and related to wet and rainfed wheat insurance policies of Khuzestan province, for which compensation was paid in the 2018-2019 crop year. After preparing and preprocessing the data, using deep learning to discover unusual cases, the action and results were evaluated by the experts of the Agricultural Insurance Fund. After analyzing the results, it was found that 1% of the damages paid were related to unrealistic requests and more care

* Corresponding Author: Poorebrahimi@gmail.com

How to Cite: Ahmadlou Y., Pourebrahimi A., Tanha J., Rajabzade A. (2023). A Model for Detecting Abnormal Claims in Agricultural Insurance Using Deep Learning, *Journal of Business Intelligence Management Studies*, 12(45), 313-346.

should be taken in paying the damages. The accuracy of the model in detecting unusual cases for wet and dry wheat was 53.53 and 63.37 percent, respectively. In the review of the results, it was found that 5 categories of unusual behavior have led to the payment of unrealistic damages, and the behavior of not providing damage documentation was more frequent than the others.

1. Introduction

Insurance fraud refers to the immoral act of committing a crime with the intention of abusing an insurance policy to obtain illegal profit from an insurance company; In general, insurance is made to protect the assets and business of individuals or organizations against financial loss and may occur at any stage of the insurance process by anyone such as customers or fraudulent agents (Al -Hashedi & Magalingam, 2021). Insurance fraud not only reduces the profit of the insurance company and leads to major losses, but also affects the pricing strategy of the insurance company and its socio-economic benefits in the long term (Yaram, 2016). Every year, significant sums of money are defrauded from the insurance industry, but not all of them are discovered. According to the statistics published by the Insurance Anti-Fraud Coalition, an amount of about eighty billion dollars is added to customers' expenses in the United States through fraud, and they must compensate for the amount of fraud by paying higher insurance premiums in the following year (Fraud statistics, 2020). In Iran, there is no accurate estimate of the amount of compensations paid to unreal damage claims or any other fraud, and one of the goals of this research is to estimate the amount of fraud in wheat crop insurance using deep learning.

2. Research Question(s)

This research seeks to find answers to these questions: In rainfed and irrigated wheat crop insurance, what percentage of the paid compensations are related to unrealistic and fictitious damage claims, and what is the accuracy of deep learning detection for this purpose?

3. Literature Review

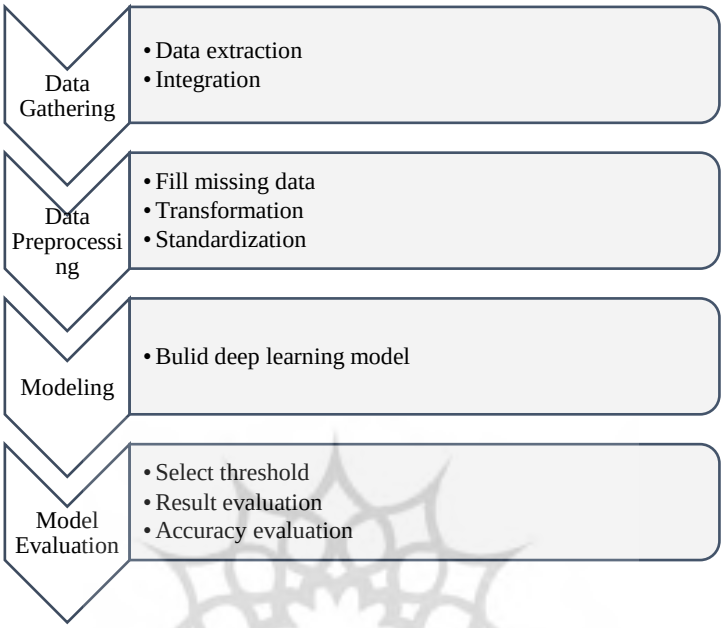
Ghahari et al. (2019) in their study investigated the use of deep learning in predicting agricultural performance in time and space with unstable weather conditions. They compared the performance of machine learning next to weather stations with conventional methods. Their findings showed that deep learning provides the highest

prediction accuracy compared to other approaches. It can also be inferred from this result that the use of deep learning can play a role in reducing agricultural insurance costs by knowing the exact measures of crop yield (Newlands et al., 2019). Gomez et al. (2021) presented a new deep learning method to gain pragmatic insight into the behavior of an insured individual using the unsupervised effective variable. Their proposed method can be used in the fields of pension insurance, investment and other broader areas of the insurance industry. Their proposed method enables auto encoder and variable auto encoder to be used in semi-supervised/unsupervised effective variable analysis to identify cheating agents (Gomes et al., 2021). Xia et al. (2022) in their study proposed a deep learning model to detect car insurance fraud by combining convolutional neural network, long-term and short-term memory, and deep neural network. In their proposed method, more abstract features were extracted and helped the experts in the complex process of feature extraction which is very critical in traditional machine learning algorithms. The results of the experiments showed that their method can effectively improve the accuracy of car insurance fraud detection.

4. Methodology

The current research method is practical from the point of view of the objective and is data-oriented from the point of view of its nature. For machine learning modeling, the standard CRISP process has been used, which includes the stages of data collection, data preparation and preprocessing, modeling and model evaluation, and obtaining results. Figure 1 shows the general process of anomaly detection and analysis.

Figure 1. Anomaly detection process framework



In this research, the data related to one agricultural year of wet and dry wheat crop were obtained from the Agricultural Insurance Fund. The national code of the insurers has been removed from the data set to maintain confidentiality. The extracted data is related to the crop insurance policies of wet and rainfed wheat for the crop year 2018-2019 of Khuzestan province. In this crop year, compensation has been paid for these insurance policies according to the claim of the damage they had, in other words, the data set includes those insurance policies of wet and dry wheat whose product is damage Seen and compensated for them. The data were obtained from the comprehensive system of the insurance fund in the form of a csv report. The obtained data set had 23 features.

5. Conclusion

The results of the research show that in wheat insurance, about 1% of the compensations paid are allocated to unrealistic claims, so they need to be further investigated by experts before payment. This amount of compensations paid to unrealistic claims was close to the prediction of insurance fund inspection experts who stated that about

1.5% of claims are unrealistic. Also, according to the results, 5 categories of behavior or methods were identified in the beneficiaries to receive compensation for unrealistic claims, which are mentioned below:

1. Lack of sufficient documentation to prove the damage: This means that the necessary documents that should be uploaded in the system according to the implementation methods are not available or some of them have not been uploaded. Payment of compensation without the existence of documents indicating the occurrence of damage can be caused by the negligence or collusion of the appraiser or broker with the insured.

2. The documents are not in accordance with the declared damage: the documents uploaded in the system according to the relevant instructions do not show the occurrence of the type of registered damage. For example, the speed of storm damage is mentioned as 50 km/h, but in meteorological documents it is 15 km/h.

3. The damage documentation is not true: for example, in some documents, the risk factor is mentioned in the expert form of drought, but the picture sent shows flood damage. In this case, it is probably due to negligence. In another possibility, it is also possible to send the image of damaged agricultural land instead of healthy agricultural land.

4. Non-observance of the damage notification period: According to the executive instructions of the insurance fund, the time limit for the declaration of damage until the time of payment of compensation is one month. Outside of that, it is against the instructions. Sometimes it was observed that the damage had been declared before the accident.

5. The date of damage does not match with the time of its announcement: according to the executive instructions of the insurance fund, in the case of damage to agriculture, the visit must be done one week after the occurrence of the damage; before removing the damage, the type and amount of the damage should be carefully checked. In some cases, it was observed that the announcement date was recorded one month after the damage occurred. It is clear that

after removing the effects of damage, the payment of compensation can seem suspicious because there may not have been any damage in the past.

Keywords: Anomaly Detection, Crop Insurance, Deep Learning, Auto Encoder.

مقاله پژوهشی تاریخ ارسال: ۱۴۰۱/۰۵/۱۸ تاریخ بازنگری: ۱۴۰۱/۰۲/۲۷ تاریخ پذیرش: ۱۴۰۲/۰۷/۲۵ ISSN: 2821-0964 eISSN: 2821-0816

مطالعات مدیریت کسب و کار هوشمند
سال دوازدهم، شماره ٤٥، پاییز ١٤٠٢، ص ٣١٣ تا ٣٤٦
ims.atu.ac.ir
DOI: 10.22054/ims.2023.69377.2213

مدلی برای تشخیص ادعاهای غیرعادی خسارت در بیمه کشاورزی با استفاده از یادگیری عمیق

یعقوب احمدلو دانشجوی دکتری، گروه مدیریت فناوری اطلاعات، دانشکده مدیریت و اقتصاد، واحد علوم و تحقیقات، دانشگاه آزاد اسلامی، تهران، ایران.

علیرضا پورابراهیمی * استادیار گروه مدیریت صنعتی، دانشکده مدیریت و حسابداری، واحد کرج، دانشگاه آزاد اسلامی، کرج، ایران.

جعفر تنها دانشیار گروه مهندسی فناوری اطلاعات، دانشکده مهندسی برق و کامپیوتر، دانشگاه تبریز، تبریز، ایران.

علی رجب‌زاده قطری استاد گروه مدیریت صنعتی، دانشکده مدیریت و اقتصاد، دانشگاه تربیت مدرس، تهران، ایران.

چکیده

موارد کلاهبرداری در سال‌های اخیر به‌ویژه در زمینه‌های مهم و حساس مالی و بیمه‌ای افزایش یافته است. ازاین‌رو، برای مقابله با این‌گونه کلاهبرداری‌ها نیاز به اقدامات متفاوتی نسبت به روش‌های بازرسی دستی و سنتی وجود دارد. بیمه کشاورزی نیز با توجه به ماهیت و گستردگی وسیع آن از این تهدید مستثنا نبوده و سالانه هزینه‌های زیادی صرف پرداخت به خسارت‌های ساختگی می‌شود. این پژوهش باهدف ارائه مدلی برای کشف ادعاهای خسارت غیرواقعی در بیمه کشاورزی با به‌کارگیری تکنیک‌های داده‌کاوی و یادگیری ماشین تدوین شد. برای ساخت مدل از الگوریتم خودرمزگذار که به‌عنوان یکی از تکنیک‌های یادگیری عمیق به‌حساب می‌آید استفاده شده است. داده‌های مورداستفاده برای آموزش مدل از صندوق بیمه کشاورزی اخذ شد و مربوط به بیمه‌نامه‌های گندم آبی و دیم استان خوزستان بود که در سال زراعی ۱۳۹۹-۱۳۹۸ برای آن‌ها غرامت پرداخت شده بود. بعد از آماده‌سازی و پیش‌پردازش داده‌ها، با استفاده از

مقاله حاضر برگرفته از رساله دکتری رشته مدیریت فناوری اطلاعات دانشگاه آزاد اسلامی واحد علوم و تحقیقات تهران است.

* نویسنده مسئول: Poorebrahimi@gmail.com

یادگیری عمیق نسبت به کشف موارد غیرعادی اقدام و نتایج توسط کارشناسان صندوق بیمه کشاورزی مورد ارزیابی قرار گرفت. بعد از تحلیل نتایج مشخص شد یک درصد از خسارت‌های پرداختی مربوط به درخواست‌های غیرواقعی بوده و در پرداخت خسارت بایستی دقت و بررسی بیشتری انجام شود. دقت مدل در تشخیص موارد غیرعادی برای گندم آبی و دیم به ترتیب برابر با ۵۳/۵۳ و ۶۳/۳۷ درصد به دست آمد. در بررسی نتایج مشخص شد ۵ دسته رفتار غیرعادی منجر به پرداخت خسارت غیرواقعی شده‌اند که رفتار عدم ارائه مستندات خسارت فراوانی بیشتری نسبت به بقیه داشت.

کلیدواژه‌ها: تشخیص ناهنجاری، بیمه کشاورزی، یادگیری عمیق، خودرمزگذار.

مقدمه

کلاهبرداری بیمه‌ای به عمل غیراخلاقی ارتکاب جرم با قصد سوءاستفاده از بیمه‌نامه برای کسب سود غیرقانونی از یک شرکت بیمه اطلاق می‌شود؛ به‌طورکلی، بیمه برای محافظت از دارایی‌ها و کسب‌وکار افراد یا سازمان‌ها در برابر ضرر مالی ساخته می‌شود و ممکن است در هر مرحله از فرآیند بیمه توسط هرکسی مانند مشتریان یا نمایندگان کلاهبرداری رخ دهد (Al-Hashedi & Magalingam, 2021). کلاهبرداری بیمه‌ای نه‌تنها سود شرکت بیمه را کاهش داده و منجر به زیان‌های اساسی می‌شود، بلکه بر استراتژی قیمت‌گذاری شرکت بیمه و منافع اقتصادی-اجتماعی آن در بلندمدت تأثیر می‌گذارد (Yaram, 2016).

طبق گزارش بخش کشاورزی آژانس مدیریت ریسک در ایالات‌متحده[1]، نرخ پرداخت برای تقاضاهای نادرست در بیمه محصولات کشاورزی برای سال مالی ۲۰۱۹ حدود ۲/۹۵ درصد بود که کمتر از میانگین نرخ برای همه برنامه‌های دولتی (۴/۶۷ درصد) است (Crop Insurance Statistics, 2022). طبق آماری که از طرف ائتلاف ضد تقلب بیمه منتشر شده است، مبلغی در حدود هشتاد میلیارد دلار در آمریکا از طریق تقلب به هزینه‌های مشتریان اضافه می‌شود و آن‌ها بایستی در سال بعد میزان کلاهبرداری را با پرداخت حق بیمه بیشتر جبران کنند (Fraud stats, 2020). فدراسیون بیمه و بیمه اتکایی اروپا، تخمین می‌زنند که کل مطالبات تقلبی در اروپا در سال ۲۰۱۷ بالغ بر ۱۳ میلیارد یورو بوده است (Debener et al., 2023). طبق مارزن[2] (۲۰۱۳) کلاهبرداری در بیمه کشاورزی مانند سایر زمینه‌های بیمه وجود دارد و طبق تخمین انجام‌شده در سال ۲۰۱۲، تقلب در بیمه محصولات زراعی حدود ۵ درصد از کل آن را شامل می‌شود.

تقلب در بیمه کشاورزی می‌تواند از طریق اعلام نادرست خسارت به‌صورت آگاهانه، ادعاهای خسارت غیرواقعی و ساختگی، اعلام اشتباه میزان برداشت نهایی محصول، تبانی بین عامل بیمه‌گر یا ارزیاب خسارت و بیمه‌گذار اتفاق بیفتد (GAO, 2006). به‌عنوان مثالی

1. U.S. Department of Agriculture, *Risk Management Agency*
2. Marzen C. G.

از روش کلاهبرداری بیمه کشاورزی، یک کشاورز از کارولینای شمالی برای سال‌ها بیش از ۹ میلیون دلار غرامت مشکوک دریافت کرده بود. یکی از موارد تقلب این بود که به کارگران دستور داده بود تا تکه‌های یخ و گلوله‌های نفتالین را در مزرعه گوجه‌فرنگی تنسی پراکنده کنند تا نشان دهد گیاهان در اثر طوفان تگرگ آسیب دیده‌اند (Marzen, 2013).

در ایران برآورد دقیقی از میزان غرامت‌های پرداختی به ادعاهای غیرواقعی خسارت یا هرگونه کلاهبرداری دیگری در دست نیست و یکی از اهداف این پژوهش برآورد میزان تقلب در بیمه محصول گندم آبی و دیم با استفاده از یادگیری عمیق می‌باشد. ازاین‌رو پژوهش حاضر دنبال پاسخ به این پرسش‌هاست؛ در بیمه محصول گندم آبی و دیم چند درصد از غرامت‌های پرداختی مربوط به ادعاهای غیرواقعی و ساختگی خسارت می‌باشد و دقت تشخیص یادگیری عمیق برای این‌منظور چقدر است؟ در ادامه ساختار مقاله به ترتیب مبانی نظری و پیشینه پژوهش، روش‌شناسی، ساخت و ارزیابی مدل، یافته‌های پژوهش و بحث و نتیجه‌گیری خواهد بود.

مبانی نظری

ازآنجاکه شناسایی دستی موارد کلاهبرداری پرهزینه و ناکارآمد است و باید تقلب قبل از پرداخت ادعا شناسایی شود، تکنیک‌های داده‌کاوی به‌طور گسترده به‌عنوان روشی مؤثر در مبارزه با تقلب شناخته می‌شود (Hilal et al., 2022). تکنیک‌های داده‌کاوی و یادگیری ماشینی پتانسیل شناسایی به‌موقع موارد مشکوک را دارند و به‌طور بالقوه هزینه‌ها را برای بیمه‌گران و بیمه‌گذاران کاهش می‌دهند (Nian et al., 2016).

تشخیص ناهنجاری یکی از رایج‌ترین کاربردهای داده‌کاوی می‌باشد که برای یافتن سوابق تقلب استفاده‌شده (Kirlidog & Asuk, 2012) و الگوهای متفاوت در داده که با رفتار عادی مطابقت ندارد را کشف کرده (Chandola et al., 2019) و معمولاً برای کشف ناهنجاری‌های بیمه‌ای، تشخیص نفوذ شبکه و کارت‌بانکی به‌کار می‌رود (Zhao et

al., 2019). در گزارش دفتر پاسخگویی دولت ایالات‌متحده[1] آمده است که آژانس مدیریت ریسک آمریکا، استفاده از تکنیک‌های داده‌کاوی برای تجزیه‌وتحلیل داده‌ها و ایجاد ارتباط با کشاورزان پرریسک را علت کاهش حداقل ۳۰۰ میلیون دلاری در پرداخت مطالبات مشکوک بین سال‌های ۲۰۰۱ تا ۲۰۰۴ دانسته است (GAO, 2006).

یکی از نیازهای رایج هنگام تجزیه‌وتحلیل مجموعه‌داده‌های دنیای واقعی، تعیین و پیدا کردن نمونه‌هایی است که با بقیه متفاوت هستند؛ چنین نمونه‌هایی به‌عنوان ناهنجاری شناخته می‌شوند و هدف از تشخیص ناهنجاری تعیین موارد غیرعادی به روش داده محور است (Chalapathy & Chawla, 2019). تشخیص ناهنجاری شاخه بسیار مهمی از یادگیری ماشینی است که کاربردهای عملی گسترده‌ای داشته و هدف آن شناسایی نقاط خاص و متفاوت در مجموعه‌داده‌ها است (Zhang et al., 2021)؛ بنابراین تشخیص ناهنجاری را می‌توان به‌عنوان یک شکل کلی از کشف تقلب در نظر گرفت، به‌عبارتی کشف تقلب یکی از کاربردهای تشخیص ناهنجاری است (Chandola et al., 2009).

اکین و همکاران[2] (۲۰۱۹) در پژوهش خود بیان کرده‌اند ازآنجاکه بیمه‌گران در عصری فعالیت می‌کنند که در آن ترکیب فناوری و شیوه‌های عملیات مجرمانه پویا است، نمی‌توان از یک مجموعه الگوریتم برای کشف کلاهبرداری بیمه استفاده کرد. دشواری یادگیری این ویژگی‌های رفتاری پویا ازنظر پیچیدگی به‌طور تصاعدی افزایش می‌یابد و یادگیری نظارت‌شده، در عمل نمی‌تواند چنین متغیرهای پنهان پیچیده در داده‌ها را یاد بگیرد؛ لذا این محدودیت‌ها مانع استفاده از مدل‌سازی نظارت‌شده در کشف تقلب بیمه‌شده و فرصتی برای یک الگوی مدل‌سازی جایگزین و منعطف که یادگیری با مدل‌سازی نظارت‌نشده است را فراهم می‌کند. با توجه به محدودیت‌های ذاتی مدل‌های نظارت‌شده در کشف تقلب بیمه، استفاده از آن‌ها در این زمینه تا حد زیادی غیرعملی است. به دلیل رفتار پویای کلاهبرداران مدل‌های یادگیری ماشین نظارت‌نشده می‌توانند نقش مؤثری نسبت به روش‌های یادگیری نظارت‌شده ایفا کرده (Gomes et al., 2021) و موارد پرت

1. U.S. Governance Accountability Office
2. Ekin et al.,

را تنها بر اساس ویژگی‌های ذاتی داده‌ها شناسایی می‌کند (Chalapathy & Chawla, 2019).

یادگیری ماشین و یادگیری عمیق دو موضوع پرکاربرد در هوش مصنوعی هستند که یادگیری عمیق یکی از انواع تخصصی یادگیری ماشین به‌حساب می‌آید (قباخلو و همکاران، ۱۴۰۱). مدل‌های یادگیری نظارت‌نشده مبتنی بر شبکه‌های عصبی با لایه‌های «پنهان» متعدد که به‌عنوان پرسپترون‌های چندلایه[1] نیز شناخته می‌شوند اغلب به‌عنوان یادگیری عمیق نظارت‌نشده شناخته می‌شوند (Gomes et al., 2021). شبکه‌های خودرمزگذار[2] نوعی شبکه عصبی بوده و به‌عنوان اصلی‌ترین معماری یادگیری عمیق نظارت‌نشده است که به‌طور گسترده در تشخیص ناهنجاری استفاده می‌شود (Chalapathy & Chawla, 2019). هدف اصلی شبکه‌های خودرمزگذار، بازتولید ورودی با حداقل خطای بازسازی است، درحالی‌که در پرسپترون چندلایه، شبکه سعی می‌کند خروجی موردنظر را با ورودی‌های مشخص پیش‌بینی کند (Zamini & Montazer, 2018).

طبق ژانگ و همکاران[3] (۲۰۲۱) خودرمزگذار یک شبکه عصبی چندلایه است که سیگنال ورودی را بازتولید می‌کند. برای بازتولید داده‌های ورودی، خودرمزگذار باید مهم‌ترین ویژگی‌ها را که می‌تواند داده‌های ورودی را نشان دهد، استخراج کند. هنگامی‌که تعداد گره‌های لایه پنهان میانی کمتر از تعداد گره‌های لایه ورودی باشد، تنها مهم‌ترین ویژگی‌های داده یاد گرفته می‌شود. یکی از مزایای خودرمزگذار استخراج ویژگی‌های مفید و حذف ویژگی‌های غیرمفید است (Liu et al., 2017). بعد از بازتولید ورودی، آن دسته از نمونه‌ها که دارای خطای بازسازی بالایی هستند به‌عنوان موارد ناهنجار یا غیرعادی شناسایی می‌شوند و موارد عادی دارای خطای بازسازی پایین خواهند بود.

1. Multi-layer perceptron (MPL)
2. Autoencoder
3. Zhang et al.,

پیشینه پژوهش

زمینی و منتظر[1] (۲۰۱۸) از روش یادگیری نظارت‌نشده خودرمزگذار برای تشخیص تقلب کارت‌های اعتباری استفاده کردند. آن‌ها یک خودرمزگذار با سه لایه پنهان ایجاد و خوشه‌بندی کا-میانگین[2] را برای ۲۸۴،۸۰۷ تراکنش بانک اروپایی آزمایش کردند. بر اساس نتایج به‌دست‌آمده، دقت روش ارائه‌شده ۹۸/۹ درصد بوده و دارای ۸۱ درصد حساسیت بود.

نیولند و همکاران[3] (۲۰۱۹) در مطالعه خود، استفاده از یادگیری عمیق در پیش‌بینی عملکرد کشاورزی در زمان و مکان با شرایط ناپایدار جوی را بررسی کردند. آن‌ها عملکرد یادگیری ماشین در کنار ایستگاه‌های هواشناسی را با روش‌های مرسوم مقایسه کردند. یافته‌های آن‌ها نشان داد که یادگیری عمیق نسبت به سایر رویکردها بالاترین دقت پیش‌بینی را ارائه می‌دهد.

رضاپور[4] (۲۰۱۹) در مطالعه خود برای کشف ناهنجاری در کارت‌های اعتباری، سه روش ماشین بردار پشتیبان تک کلاسه[5]، فاصله ماهالانوبیس[6] و خودرمزگذار را مورداستفاده قرار داد. ایشان با توجه به این‌که داده‌ها دارای برچسب بودند دو روش ماشین بردار پشتیبان تک کلاسه و خودرمزگذار را به‌صورت نظارت‌شده و فاصله ماهالانوبیس را به‌صورت نظارت‌نشده مدل‌سازی کردند. ارزیابی نتایج نشان داد خودرمزگذار عملکرد بهتری نسبت به بقیه تکنیک‌ها داشت و توانست ۹۱ درصد موارد را به‌درستی تشخیص دهد.

چالاپاتی و چاولا[7] (۲۰۱۹) در مطالعه جامع خود، ابتدا مروری ساختاریافته و جامع از روش‌های پژوهش در تشخیص ناهنجاری مبتنی بر یادگیری عمیق ارائه کرده سپس میزان

1. Zamini, M., & Montazer, G.
2. K-means
3. Newlands et al.,
4. Rezapour, M.
5. One-Class Support Vector Machine (OC-SVM)
6. Mahalanobis distance
7. Chalapathy, R., & Chawla, S.

کاربرد این روش‌ها برای کشف ناهنجاری در حوزه‌های کاربردی مختلف ازجمله بیمه بررسی کرده و اثربخشی آن‌ها را ارزیابی کردند. آن‌ها تکنیک‌های تشخیص ناهنجاری عمیق را در دسته‌های مختلف گروه‌بندی کردند. در هر دسته، تکنیک اصلی تشخیص ناهنجاری را همراه با انواع آن و فرضیات کلیدی ارائه کردند تا بین رفتار عادی و غیرعادی تمایز قائل شود. علاوه بر این، برای هر دسته، مزایا و محدودیت‌ها آن را نیز ارائه و پیچیدگی محاسباتی تکنیک‌ها را در حوزه‌های کاربردی واقعی موردبحث قرار دادند. درنهایت، مسائل باز در پژوهش‌ها و چالش‌هایی که در هنگام اتخاذ تکنیک‌های تشخیص ناهنجاری عمیق برای با آن‌ها مواجه است را ترسیم کردند.

گومز و همکاران[1] (۲۰۲۱) یک روش یادگیری عمیق جدید برای به دست آوردن بینش عمل‌گرایانه در مورد رفتار یک فرد بیمه‌شده با استفاده از متغیر مؤثر[2] نظارت‌نشده ارائه دادند. روش پیشنهادی آن‌ها می‌تواند در زمینه‌های بیمه بازنشستگی، سرمایه‌گذاری و سایر حوزه‌های گسترده‌تر صنعت بیمه به کار گرفته شود. روش پیشنهادی آن‌ها خودرمزگذار و خودرمزگذار متغیر[3] را قادر می‌سازد تا در تجزیه‌وتحلیل متغیر مؤثر نیمه‌نظارت‌شده/نظارت‌نشده برای شناسایی عوامل تقلب استفاده شوند.

طبق ژانگ و همکاران (۲۰۲۱) اگرچه روش‌های کاهش ابعاد و تخمین چگالی پیشرفت زیادی در سال‌های اخیر داشته‌اند، بااین‌حال حفظ اطلاعات کلیدی داده‌های اصلی در روش‌های کاهش بعد هنوز دارای مشکل هستند که روش خودرمزگذار عمیق می‌تواند بر این مشکل فائق آید؛ بنابراین در مطالعه خود به‌منظور بهبود عملکرد تشخیص ناهنجاری نظارت‌نشده، یک روش تشخیص ناهنجاری مبتنی بر خودرمزگذار عمیق برای استخراج ویژگی‌های مهم و خوشه‌بندی برای تشخیص ناهنجاری ارائه کردند. آن‌ها ابتدا مدل یادگیری عمیق را به‌منظور فشرده‌سازی ویژگی‌ها بکار گرفته سپس از تکنیک خوشه‌بندی برای تشخیص ناهنجاری استفاده کردند. طرح پیشنهادی آن‌ها توانست اطلاعات اضافی

1. Gomes et al.,
2. Variable importance
3. Variational AutoEncoder (VAE)

موجود در داده‌ها را حذف کند و عملکرد روش‌های خوشه‌بندی در شناسایی نمونه‌های غیرعادی را بهبود بخشد.

شیا و همکاران[1] (۲۰۲۲) در مطالعه خود یک مدل یادگیری عمیق برای شناسایی تقلب بیمه خودرو با ترکیب شبکه عصبی کانولوشن[2]، حافظه بلندمدت و کوتاه‌مدت[3] و شبکه عصبی عمیق[4] پیشنهاد دادند. در روش پیشنهادی آن‌ها ویژگی‌های انتزاعی بیشتری استخراج‌شده و به متخصصان در فرآیند پیچیده استخراج ویژگی که به‌شدت در الگوریتم‌های یادگیری ماشین سنتی حیاتی است کمک کرد. نتایج آزمایش‌ها نشان داد که روش آن‌ها می‌تواند به‌طور مؤثری دقت شناسایی تقلب در بیمه خودرو را بهبود بخشد.

روش‌شناسی پژوهش

روش پژوهش حاضر از منظر هدف، کاربردی بوده و از منظر ماهیت داده-محور می‌باشد. برای مدل‌سازی یادگیری ماشین از فرآیند استاندارد کریسپ[5] بهره گرفته شده است که شامل مراحل جمع‌آوری داده‌ها، آماده‌سازی و پیش‌پردازش داده‌ها، مدل‌سازی و ارزیابی مدل و اخذ نتایج می‌باشد. در شکل ۱ فرآیند کلی تحلیل و کشف ناهنجاری نشان داده شده است.

1. Xia et al.,
2. Convolutional Neural Network
3. Long short-term memory
4. Deep Neural Network
5. Cross Industrial Standard Process for Data Mining (CRISP-DM)

شکل ۱. فرآیند کلی تحلیل و کشف ناهنجاری (منبع: یافته‌های تحقیق حاضر)

در ادامه مجموعه‌داده مورداستفاده در پژوهش حاضر معرفی‌شده، نحوه جمع‌آوری و فرآیند پیش‌پردازش و آماده‌سازی آن ارائه‌شده و در انتها الگوریتم خودرمزگذار که برای ساخت مدل یادگیری عمیق مورداستفاده قرارگرفته توضیح داده شده است.

معرفی مجموعه‌داده

در این پژوهش داده‌های مربوط به یک‌سال زراعی محصول گندم آبی و دیم از صندوق بیمه کشاورزی اخذ شده است. داده‌های استخراج‌شده مربوط به بیمه‌نامه‌های محصول گندم آبی و دیم برای سال زراعی ۱۳۹۹-۱۳۹۸ استان خوزستان می‌باشد. این بیمه‌نامه‌ها در این سال زراعی طبق ادعای خسارتی که داشته‌اند برای آن‌ها غرامت پرداخت شده است، به‌عبارتی مجموعه‌داده شامل آن‌دسته از بیمه‌نامه‌های گندم آبی و دیم می‌شوند که محصول آن‌ها خسارت‌دیده و برای آن‌ها غرامت پرداخت شده است. داده‌ها از سامانه جامع صندوق

بیمه به‌صورت گزارش csv اخذ شدند. مجموعه‌داده به‌دست‌آمده دارای ۲۳ ویژگی یا فیلد بود. جزییات مجموعه‌داده استخراج‌شده در جدول ۱ لیست شده است.

جدول ۱. مشخصات مجموعه‌داده

ردیف	نام ویژگی	نوع داده	نوع مقیاس	توضیح
۱	ItemID	Integer	Nominal	شناسه مورد بیمه‌شده (به‌صورت یکتا برای هر قطعه زمین بیمه‌شده تولید می‌شود)
۲	InsuranceID	Integer	Nominal	شناسه بیمه‌نامه یکتا (برای هر بیمه‌نامه صادرشده تولید می‌شود)
۳	Practice	String	Nominal	نوع کشت محصول (آبی یا دیم)
۴	Province	String	Nominal	عنوان استان
۵	City	String	Nominal	نام شهر
۶	Branch	String	Nominal	نام شعبه بانک کشاورزی که در آن کارهای پرداخت غرامت انجام می‌شود
۷	Village	String	Nominal	دهستان
۸	Local	String	Nominal	محل کشت
۹	Max_Premium	Integer	Ratio	حداکثر تعهد بیمه‌گر
۱۰	Ins_type	String	Nominal	نوع بیمه (پایه یا تکمیلی)
۱۱	Indemnity	Integer	Ratio	میزان غرامت پرداخت‌شده
۱۳	Occur_date	Date	Interval	تاریخ وقوع خسارت
۱۴	Insurer	String	Nominal	بیمه‌گذار
۱۵	Issuer	String	Nominal	صادرکننده بیمه‌نامه
۱۶	Agent	String	Nominal	نام نماینده بیمه‌گر
۱۷	Insured_acres	Float	Ratio	میزان مساحت بیمه‌شده
۱۸	Damaged_acres	Integer	Ratio	میزان مساحت خسارت دیده‌شده
۱۹	Insurer_Primium	Integer	Ratio	حق بیمه سهم کشاورز
۲۰	Plan_Type	String	Nominal	نوع طرح (انفرادی، عمومی، پایه، تجمیعی، ...)
۲۱	Factor	String	Nominal	عامل خطر (طوفان، تگرگ، سیل، سرما و یخبندان، خشک‌سالی، باد گرم)
۲۲	Assessor	String	Nominal	نام ارزیاب
۲۳	Ave_Dmg_acres	Float	Ratio	فیلد مشتق شده از «میزان مساحت خسارت‌دیده» که نشان‌دهنده میانگین خسارت در آن محل است

پس از استخراج و تجمیع داده‌ها، تعداد کل نمونه‌های خسارت‌دیده برای یک سال زراعی در استان خوزستان که غرامت دریافت کرده بودند ۲۱،۰۴۳ نمونه شد. همان‌طور که مشاهده می‌شود مجموعه‌داده به‌دست‌آمده فاقد ویژگی برچسب می‌باشد؛ یعنی برای نمونه-ها نوع متقلبانه بودن یا نبودن غرامت مشخص نشده است. لذا با مسئله نظارت‌نشده روبه‌رو بوده و رویکردهای متناظر با این نوع مجموعه‌داده بایستی در پیش گرفته شوند.

آماده‌سازی و پیش‌پردازش داده‌ها

حدود ۹۰ درصد از زمان دانشمندان علوم داده صرف آماده‌سازی و پیش‌پردازش داده‌ها می‌شود؛ بنابراین این مرحله پرهزینه‌ترین مرحله در هر پروژه یادگیری ماشین می‌باشد. انتخاب ویژگی‌هایی مؤثر برای آموزش مدل‌های یادگیری تأثیر زیادی بر دقت و عملکرد مدل دارد، لذا ویژگی‌های نامربوط و تا حدی مرتبط می‌تواند بر عملکرد مدل تأثیر منفی داشته باشند (Brownlee, 2020)؛ همچنین اگر ویژگی‌ها به‌دقت فیلتر نشوند، مدل به‌دست‌آمده روی داده‌های جدید عملکرد خوبی نخواهند داشت (Kim et al., 2020)؛ بنابراین چند تکنیک آماده‌سازی و پیش‌پردازش برای بالابردن کیفیت مجموعه‌داده مورد استفاده قرار می‌گیرد.

ازآنجاکه ورود داده‌ها به‌صورت دستی انجام می‌شود برای بسیاری از ویژگی‌ها مقادیر ازدست‌رفته[1] وجود داشت. لذا ویژگی‌هایی که فاقد مقدار بودند یا به‌عبارتی دارای مقادیر ازدست‌رفته بودند تکمیل شدند. برای این کار که زمان‌بر بود، مقادیر متناسب با نوع ویژگی بر اساس مقادیر مشابه آن ویژگی در سایر نمونه‌ها انتخاب و با مقادیر ازدست‌رفته جایگزین شدند.

با بررسی مجموعه‌داده مشخص شد تعداد ۳،۰۲۳ نمونه تکراری وجود دارد که در ادامه تمام آن‌ها حذف شدند. تعداد ۱۸،۰۲۰ نمونه در مجموعه‌داده نهایی بعد از حذف موارد تکراری باقی ماندند که به ترتیب ۱۱،۲۹۳ برای بیمه‌نامه‌های گندم آبی و ۶۷۲۷ نمونه

1. Missing Value

مربوط به بیمه‌نامه‌های گندم دیم بود.

انتخاب ویژگی می‌تواند با حذف ویژگی‌های نامربوط و اضافی، زیرمجموعه کوچکی از ویژگی‌های مرتبط را از ویژگی‌های اصلی انتخاب کند (Miao & Niu, 2016). روش انتخاب ویژگی برحسب نوع مجموعه‌داده که دارای برچسب است یا نه، ممکن است متفاوت باشد. یک رویکرد اساسی برای انتخاب ویژگی، آستانه واریانس است؛ یعنی همه ویژگی‌هایی را که واریانس آن‌ها از آستانه تجاوز نمی‌کند حذف می‌کند. به‌طور پیش‌فرض تمام ویژگی‌های واریانس صفر یا به عبارتی ویژگی‌های دارای مقادیر یکسان در تمام نمونه را حذف می‌کند. این الگوریتم فقط به ویژگی‌های ورودی نگاه می‌کند نه خروجی‌ها، بنابراین می‌تواند برای یادگیری نظارت‌نشده شده مورداستفاده قرار گیرد (Sarker, 2021). در این پژوهش که برچسب نمونه‌ها در دسترس نبود از روش واریانس استفاده شد و ویژگی‌هایی که دارای واریانس کم و صفر بودند حذف شدند، درنهایت از مجموع ۲۳ ویژگی اولیه، ۱۳ ویژگی به‌عنوان ویژگی‌های نهایی انتخاب شدند.

ازآنجاکه یادگیری عمیق و شبکه عصبی مقادیر عددی را به‌عنوان ورودی قبول می‌کنند بایستی ویژگی‌های اسمی (دسته‌ای) به‌واسطه تبدیل[1] به‌صورت عددی تبدیل شوند. برای این کار روش‌های مختلفی وجود دارد که بسته به کاربردشان می‌توان از آن‌ها استفاده کرد. در این پژوهش از روش کدگذاری باینری[2] استفاده شده است. برای تبدیل ویژگی تاریخ وقوع خسارت، بعد از تبدیل آن‌ها از حالت شمسی به میلادی، به فرمت یونیکس[3] تبدیل شدند. تاریخ در حالت یونیکس از مبدأ آن‌که معادل با اول ژانویه ۱۹۷۰ می‌باشد، صفر ثانیه در نظر گرفته می‌شود و اختلاف آن تا تاریخ موردنظر محاسبه‌شده و معادل عددی آن جایگزین تاریخ وقوع در جدول نهایی می‌شود.

تعداد ویژگی‌ها نهایی بعد از کدگذاری باینری ویژگی‌های اسمی و تبدیل تاریخ وقوع، ۴۰ عدد برای بیمه‌نامه‌های گندم آبی و ۳۷ عدد برای بیمه‌نامه‌های گندم دیم شد.

---

1. Transform
2. Binary Encoding
3. Epoch Unix Timestamp

اختلاف در تعداد نهایی ویژگی‌های کدگذاری شده ناشی از اختلاف در تعداد ویژگی‌های اسمی است که در دو نوع کشت وجود دارد.

مقیاس‌بندی ویژگی‌ها در یادگیری ماشینی یکی از حیاتی‌ترین مراحل در فرآیند پیش‌پردازش داده‌ها قبل از ایجاد یک مدل یادگیری ماشینی است؛ مقیاس‌بندی می‌تواند بین یک مدل یادگیری ماشین ضعیف و یک مدل بهتر تفاوت ایجاد کند (Bisong, 2019).

با استفاده از استانداردسازی، مقادیر ویژگی دارای میانگین صفر و انحراف معیار یک خواهند شد تا ویژگی‌ها دارای شکل توزیع نرمال باشند؛ این کار باعث می‌شود یادگیری وزن‌ها در شبکه عصبی آسان‌تر شود. علاوه بر این، استانداردسازی اطلاعات مفیدی را در مورد مقادیر پرت حفظ می‌کند و الگوریتم را در مقایسه با نرمال‌سازی حداقل–حداکثر که داده‌ها را در بازه محدودی از مقادیر [۱–۰] مقیاس‌بندی می‌کند، نسبت به مقادیر پرت حساس‌تر می‌کند (Raschka & Mirjalili, 2019).

از رابطه (۱) می‌توان برای مقیاس‌بندی استاندارد ویژگی‌ها استفاده کرد:

$$y = \frac{x - mean(x)}{sqrt(var(x))} \quad (۱)$$

که میانگین نمونه‌های (x) از رابطه (۲) به‌دست می‌آید:

$$mean(x) = \frac{1}{N}\sum_{i=1}^{N} x_i \quad (۲)$$

خودرمزگذار

یکی از روش‌های کارآمد نظارت‌نشده برای تشخیص ناهنجاری، معماری مبتنی بر خودرمزگذار است (Finke et al., 2021). خودرمزگذار یک شبکه عصبی است که از دو بخش رمزگذار و رمزگشا تشکیل‌شده و مقادیر ورودی را بازسازی می‌کند (Zhang et al., 2021). به‌عبارت‌دیگر نوعی شبکه عصبی است که برای کپی کردن ورودی خود به خروجی آموزش داده شده است. بااین‌حال درصورتی‌که یک خودرمزگذار طوری آموزش داده شود که بتواند تمام ورودی‌ها را بدون خطا بازسازی کند کاربردی نخواهد

بود؛ بنابراین، خودرمزگذارها طوری طراحی شده‌اند که قادر به یادگیری کپی کردن کامل نباشند. ازآنجاکه مدل مجبور است اولویت‌بندی کند که کدام ویژگی ورودی باید کپی شود، اغلب ویژگی‌های مفید داده‌ها را یاد می‌گیرد (Goodfellow et al., 2016).

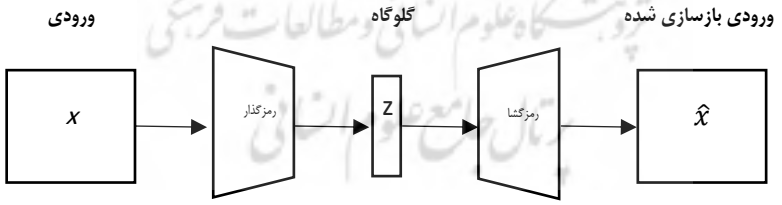
طبق گومز و همکاران (۲۰۲۱) خودرمزگذار نوعی از شبکه‌های عصبی است که ورودی را به خروجی با کمترین خطا تبدیل می‌کند. یک خودرمزگذار دارای یک تابع برای رمزگذاری به نام f و یک تابع رمزگشا بنام g می‌باشد. هدف تابع f فشرده‌سازی لایه ورودی برای نمایش کمتر آن‌ها با استفاده از ویژگی‌های مهم داده‌های ورودی می‌باشد. به همین ترتیب تابع رمزگشای g ورودی را از روی داده‌های فشرده‌شده بازسازی می‌کند؛ بنابراین عملکرد یک خودرمزگذار را می‌توان به‌صورت رابطه (۳) ارائه کرد:

$$g\big(f(x)\big) = r(x) \qquad (۳)$$

که $r(x)$ ورودی بازسازی‌شده را ارائه می‌دهد.

دینامیک یک خودرمزگذار شبیه به یک شبکه عصبی معمولی است. لایه رمزگذار/رمزگشا متشکل از واحدهای مخفی متقارن در هر لایه است که از معماری متقارن پیروی می‌کند که البته تقارن آن «ضروری» نمی‌باشد. معماری خودرمزگذار مورداستفاده در این پژوهش به‌صورت کلی در شکل ۲ نشان داده شده است.

شکل ۲. معماری کلی شبکه خودرمزگذار (منبع: یافته‌های تحقیق حاضر)



داده‌های رمزگذاری شده/رمزگشایی‌شده در لایه پنهان j، ماتریس وزن متعلق به لایه پنهان j و ماتریس بایاس در لایه j به ترتیب به‌صورت $z^{(j)}$ ،W_j و b_j نشان داده می‌شوند. عملیات در طول لایه پنهان j را می‌توان به‌صورت رابطه (۴) نشان داد:

$$z^{(j)} = \tilde{f}^{j}(W_j z^{(j-1)} + b_j) \qquad (۴)$$

که $\tilde{f}^j$ نشان‌دهنده تابع فعال‌سازی برای لایه j ام است. بهتر است برای لایه رمزگذار و رمزگشا از یک تابع فعال‌سازی و برای فعال‌سازی خروجی از یک تابع دیگر استفاده شود؛ همچنین به‌تازگی نشان داده شده است که عملکرد سیگموید به‌عنوان تابع فعال‌سازی خروجی نتایج بهتری ارائه می‌دهد (Brownlee, 2021). برخی از توابع فعال‌ساز که به‌طور گسترده در خودرمزگذار مورداستفاده قرار می‌گیرند عبارت‌اند از: تانژانت هیپربولیک، یک‌سوکننده[1] و سیگموید، با رابطه‌های (۵)، (۶) و (۷) نشان داده شده است:

$$\tanh(x) = \frac{2}{1+e^{-2x}} - 1 \quad (۵)$$

$$ReLu(x) = \begin{cases} 0 \; for \; x < 0 \\ x \; for \; x \geq 0 \end{cases} \quad (۶)$$

$$Sigmoid(x) = \frac{1}{1+e^{-x}} \quad (۷)$$

خطای خروجی با استفاده از یک تابع ضرر محاسبه می‌شود که نشان می‌دهد خروجی شبکه چقدر به مقدار هدف نزدیک است. به‌طور سنتی، از تابع میانگین مربعات خطا[2] به‌عنوان تابع ضرر استفاده می‌شود. برای مشخص کردن موارد مشکوک یا ناهنجار نمونه‌هایی که خطای بازسازی آن‌ها که از طریق محاسبه میانگین مربعات خطا به دست می‌آید، از آستانه مشخص‌شده بیشتر باشد جزو موارد مشکوک طبقه‌بندی می‌شوند و بقیه موارد جزو دسته عادی طبقه‌بندی می‌شوند. میانگین مربعات خطا از رابطه (۸) محاسبه می‌شود که به تابع ضرر معروف است.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{x}_i)^2 \quad (۸)$$

ساخت و ارزیابی مدل

همان‌طور که در قسمت قبل مشاهده گردید ماتریس فشرده را می‌توان برای هر لایه j ام از طریق رابطه (۴) به‌دست آورد. در این پژوهش از ۵ لایه با تعداد گره‌های ۵۰،۲۵،۱۰،۲۵،۵۰

1. ReLu
2. Mean Square Error (MSE)

در هر لایه برای ساخت مدل استفاده گردید. لایه ورودی برای بیمه‌نامه‌های آبی دارای ۳۹ ویژگی و برای بیمه‌نامه‌های دیم ۳۶ ویژگی دارند (ویژگی شناسه بیمه‌نامه حذف می‌شود)؛ بنابراین برای ساخت مدل خودرمزگذار لایه‌ها به‌صورت رابطه (۹) ایجاد می‌شوند. x بردار ورودی و $\hat{x}$ بردار خروجی مدل می‌باشد. در این پژوهش از تابع یک‌سوکننده به‌عنوان تابع فعال‌ساز لایه‌های پنهان رمزگذار و رمزگشا و از تابع سیگموید به‌عنوان تابع فعال‌سازی لایه خروجی استفاده شده است. پارامترهای بهینه‌ساز آدام[1] و تابع ضرر میانگین مربعات خطا برای کامپایل مدل بهره گرفته شد. با توجه به اینکه هیچ‌گونه برچسبی در دسترس نبود لذا جهت اطمینان از نتایج، برای به دست آوردن مقادیر لازم برای تنظیم پارامترها، از پروژه-های مشابه که نتایج قابل قبولی اخذ کرده بودند استفاده شد.

$$
\begin{aligned}
z^{(1)} &= ReLu(W_1 x + b_1),\\
z^{(2)} &= ReLu\left(W_2 z^{(1)} + b_2\right),\\
z^{(3)} &= ReLu\left(W_3 z^{(2)} + b_3\right),\\
z^{(4)} &= ReLu\left(W_4 z^{(3)} + b_4\right),\\
z^{(5)} &= ReLu\left(W_5 z^{(4)} + b_5\right),\\
\hat{x} &= \sigma\left(\beta z^{(5)} + b_6\right).
\end{aligned}
\qquad (۹)
$$

پارامترهای تنظیمی مدل یادگیری به شرح جدول ۲ می‌باشد.

جدول ۲. مقادیر پارامترهای خودرمزگذار (منبع: یافته‌های تحقیق حاضر)

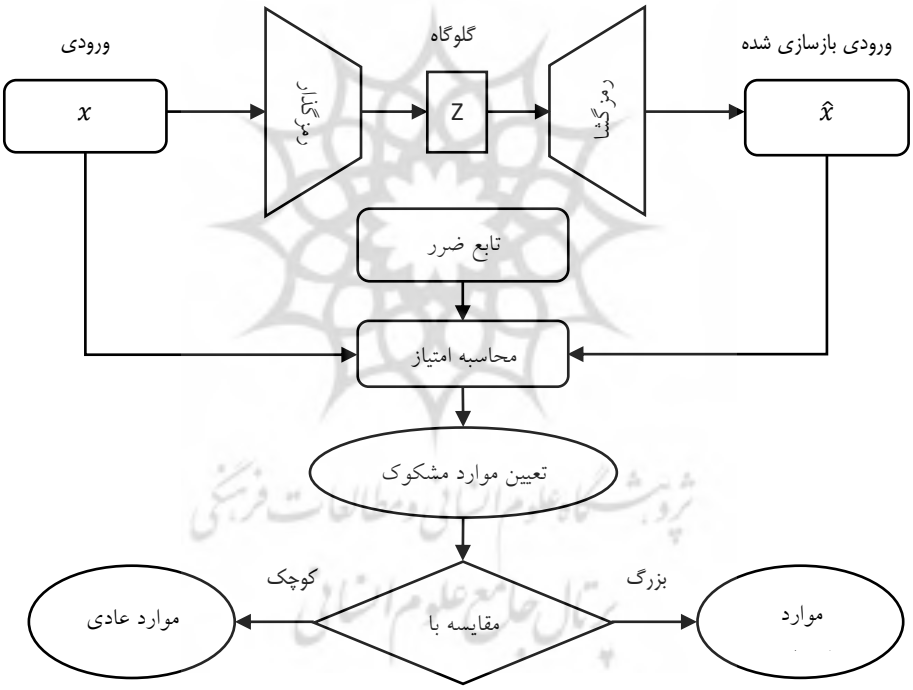
پارامتر	مقدار
تابع فعال‌ساز کدگذار و گدگشا	یک‌سوساز (ReLu)
تابع فعال‌ساز خروجی	سیگموید
تعداد ویژگی‌های ورودی	۳۹ عدد برای گندم آبی، ۳۶ عدد برای گندم دیم
تابع ضرر	میانگین مربعات خطا
بهینه‌ساز	آدام
نرخ یادگیری بهینه‌ساز	$۱۰^{-۳}$
تعداد لایه‌ها	۵
تعداد گره در هر لایه (به ترتیب)	۵۰-۲۵-۱۰-۲۵-۵۰
تعداد گره در لایه گلوگاه	۱۰

1. Adam

پارامتر	مقدار
نرخ ترک یادگیری	۰/۲
تعداد تکرار	۱۰۰

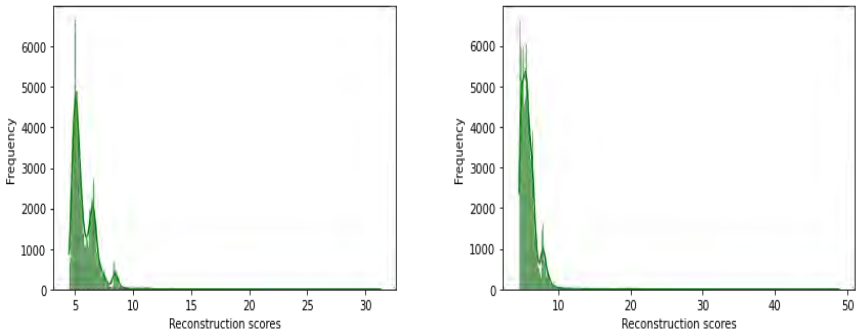
در شکل ۳ شمای کلی مدل خودرمزگذار برای کشف ناهنجاری‌ها نشان داده شده است. همان‌طور که مشاهده می‌شود موارد عادی و غیرعادی با مقایسه آستانه و امتیاز ناهنجاری به دست می‌آید. امتیاز ناهنجاری برابر با خطای بازسازی است که توسط رابطه (۸) یا تابع ضرر محاسبه می‌شود.

شکل ۱. شمای کلی مدل خودرمزگذار (منبع: یافته‌های تحقیق حاضر)



ازآنجاکه بردار ورودی استانداردسازی شده ولی نرمال‌سازی نشده است، لذا بازه آن اعدادی بین ۰ و ۱ نمی‌باشد. خروجی تابع سیگموید بین ۰ و ۱ خواهد بود که با محاسبه میانگین مربعات خطا بین x و $\hat{x}$امتیاز خطای بازسازی به دست می‌آید. در شکل ۴ نمودار هیستوگرام خطای بازسازی برای بیمه‌نامه گندم آبی و دیم نشان داده شده است.

شکل ۲. نمودار هیستوگرام خطای بازسازی بیمه‌نامه گندم آبی (راست) و گندم دیم (چپ)



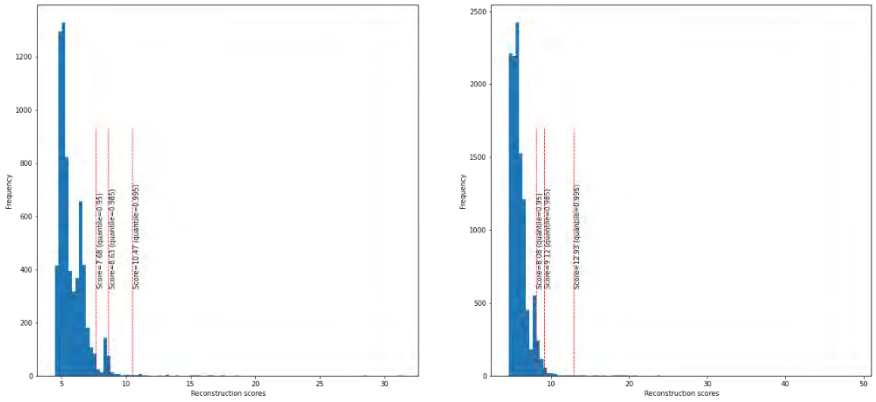
در مرحله بعد بایستی مقدار آستانه مشخص گردد. برای این منظور ازنظر کارشناسان حوزه بازرسی بیمه کشاورزی بهره‌برداری شد. لذا در ادامه نمودار هیستوگرام خطای بازسازی به همراه آستانه‌های پیش‌فرض مشخص‌شده طبق شکل ۵ برای تمام نمونه‌های آبی و دیم در اختیار کارشناسان حوزه بازرسی قرار گرفت و از آن‌ها درخواست شد تا آستانه مناسب انتخاب شود. همچنین جدول ۳ شامل مقدار آستانه به همراه تعداد ناهنجاری نیز در اختیارشان قرار گرفت تا در انتخاب آستانه بهتر عمل نمایند.

جدول 2. مقادیر آستانه و میزان ناهنجاری

چارک	مقدار آستانه بیمه-نامه گندم آبی	تعداد ناهنجاری بیمه-نامه گندم آبی	مقدار آستانه بیمه-نامه گندم دیم	تعداد ناهنجاری بیمه-نامه گندم دیم	درصد آلودگی
۰/۹۵	۸/۰۸	۵۶۵	۷/۶۸	۳۷۳	۵/۰
۰/۹۸۵	۹/۱۲	۱۷۰	۸/۶۳	۱۰۱	۱/۵
۰/۹۹۵	۱۲/۹۳	۵۷	۱۰/۴۸	۳۴	۰/۵

نمودار هیستوگرام خطای بازسازی در ادغام با آستانه‌ها در شکل ۵ به ترتیب برای گندم آبی و دیم نشان داده شده است.

شکل ۳. نمودار هیستوگرام خطای بازسازی به همراه آستانه‌ها (راست: گندم آبی، چپ: گندم دیم)



طبق بررسی انجام‌شده توسط کارشناسان، چارک ۰/۹۸۵ برای تعیین ناهنجاری نمونه‌ها به‌عنوان آستانه انتخاب شد. طبق این آستانه میزان آلودگی مجموعه‌داده ۱/۵ درصد در نظر گرفته می‌شود. در این آستانه تعداد ۱۷۰ مورد مشکوک برای بیمه‌نامه گندم آبی و ۱۰۱ مورد مشکوک برای بیمه‌نامه گندم دیم استخراج شد. در این آستانه می‌توان نمودار امتیاز ناهنجاری یا خطای بازسازی را برای تمام نمونه‌ها استخراج کرد. این نمودار در شکل ۶ برای هر دو نوع بیمه‌نامه نشان داده شده است.

شکل ٤. امتیاز ناهنجاری‌ها و آستانه مشخص‌شده (راست: گندم آبی، چپ: گندم دیم)

یافته‌ها

با توجه به اینکه مجموعه‌داده بدون برچسب بود، بنابراین نمونه‌های مشکوک در اختیار کارشناسان صندوق بیمه قرار گرفت و مشخص گردید که ناهنجاری‌های شناسایی‌شده در ۵ دسته قرار می‌گیرند. دسته‌بندی ناهنجاری‌ها در جدول ۴ لیست شده است.

جدول 3. دسته‌بندی ناهنجاری‌ها (منبع: یافته‌های تحقیق حاضر)

عنوان دسته‌بندی	شرح ناهنجاری
D1	عدم وجود مستندات کافی برای اثبات خسارت
D2	مستندات مطابق با خسارت اعلامی نیست
D3	مستندات خسارت با واقعیت مطابق نیست
T1	رعایت نشدن بازه زمانی اعلام خسارت
T2	عدم مطابقت تاریخ وقوع خسارت با زمان اعلامی آن

در جدول ۵ تعداد ناهنجاری‌های قرارگرفته در هر یک از دسته‌ها ارائه‌شده است. به دلیل نوع کشت متفاوت، همچنین تفاوت در نوع خسارت‌های وارده به دو محصول گندم آبی و دیم، تعداد ناهنجاری در آن‌ها متفاوت است. به‌عنوان‌مثال در کشت آبی طوفان و تگرگ و در کشت دیم خشک‌سالی از عوامل مهم دریافت خسارت محسوب می‌شود.

همان‌طور که در جدول ۵ مشاهده می‌شود از مجموع ۱۱،۲۹۳ نمونه بیمه‌نامه گندم آبی که برای آن‌ها غرامت پرداخت‌شده ۱۷۰ نمونه به‌عنوان غیرعادی تشخیص داده شده است که ۷۹ نمونه از آن‌ها مربوط به ادعاهای خسارت عادی بوده‌اند که غیرعادی تشخیص داده شده‌اند. همچنین از مجموع ۶،۷۲۷ نمونه بیمه‌نامه گندم دیم که برای آن‌ها غرامت پرداخت‌شده، ۱۰۱ مورد به‌عنوان غیرعادی تشخیص داده شده که ۳۷ مورد جزو ادعاهای خسارت عادی بوده‌اند که به‌عنوان غیرعادی تشخیص داده شده‌اند.

جدول ٤. نتایج بررسی موارد ناهنجار توسط کارشناسان (منبع: یافته‌های تحقیق حاضر)

نوع بیمه‌نامه	تعداد کل نمونه‌ها	تعداد کل موارد ناهنجاری شناسایی‌شده	نوع ناهنجاری و تعداد آن (TP)					تعداد تشخیص اشتباه (FP)	میزان دقت
			D1	D2	D3	T1	T2		
گندم آبی	۱۱،۲۹۳	۱۷۰	۴۱	۴	۱۶	۱۲	۱۸	۷۹	۵۳/۵۳
گندم دیم	۶،۷۲۷	۱۰۱	۳۷	۳	۹	۸	۱۳	۳۷	۶۳/۳۷

به دلیل عدم دسترسی به متغیر برچسب، تنها معیار ارزیابی در اختیار، معیار دقت[1] می‌باشد که می‌توان از رابطه (۱۰) آن را محاسبه نمود:

$$Precision = \frac{TP}{TP + FP} \qquad (۱۰)$$

در این رابطه TP نشان‌دهنده تعداد موارد مثبتی هستند که به‌درستی مثبت تشخیص داده شده و برابر با ۹۱ می‌باشد و FP تعداد موارد منفی هستند که به‌اشتباه مثبت تشخیص داده شده و برابر با ۶۴ می‌باشد؛ بنابراین دقت تشخیص مدل برای شناسایی ادعاهای غیرواقعی گندم آبی ۵۳/۵۳ درصد و برای گندم دیم ۶۳/۳۷ درصد برآورد می‌شود.

بحث و نتیجه‌گیری

نتایج حاصل از پژوهش نشان می‌دهد در بیمه گندم حدود ۱ درصد از غرامت‌های پرداخت‌شده، به ادعاهای غیرواقعی تخصیص یافته است، بنابراین قبل از پرداخت غرامت نیاز به بررسی بیشتر توسط کارشناسان دارند. این میزان از غرامت‌های پرداخت‌شده به ادعاهای غیرواقعی نزدیک به میزان پیش‌بینی‌شده کارشناسان حوزه بازرسی بود که اظهار داشتند حدود ۱/۵ درصد هستند.

همچنین طبق نتایج حاصله، ۵ دسته رفتار یا روش در ذی‌نفعان برای دریافت غرامت بابت ادعاهای غیرواقعی شناسایی شدند که در ادامه به آن‌ها اشاره می‌شود:

۱. **عدم وجود مستندات کافی در سامانه:** بدین معنی که مستندات لازم که می‌بایست

1. Precision

طبق روش‌های اجرایی در سامانه بارگذاری شود موجود نبوده یا برخی از آن‌ها بارگذاری نشده‌اند. پرداخت غرامت بدون وجود مستندات دال بر وقوع خسارت می‌توان ناشی از سهل‌انگاری یا تبانی ارزیاب یا کارگزار با بیمه‌گذار باشد.

۲. **مستندات گویای وقوع خسارت اعلامی نیست:** مستندات بارگذاری شده در سامانه طبق دستورالعمل مربوطه گویای بروز نوع خسارت ثبت شده نیستند. به‌عنوان‌مثال، سرعت در خسارت طوفان ۵۰ کیلومتر بر ساعت ذکرشده ولی در مدارک هواشناسی ۱۵ کیلومتر بر ساعت می‌باشد.

۳. **عدم مطابقت مستندات بارگذاری‌شده با واقعیت:** به‌عنوان‌مثال در برخی از مستندات عامل خطر در فرم کارشناسی خشک‌سالی ذکرشده ولی تصویر ارسالی گویای خسارت سیل می‌باشد. به‌احتمال‌زیاد ارسال تصویر زمین زراعی آسیب‌دیده به‌جای زمین زراعی سالم متصور است.

٤. **رعایت نشدن بازه زمانی اعلام خسارت:** طبق دستورالعمل اجرایی صندوق بیمه، مهلت زمان اعلام خسارت تا زمان واریز غرامت به مدت یک ماه می‌باشد و خارج از آن مغایر با دستورالعمل می‌باشد. بعضاً مشاهده شد قبل از وقوع حادثه اعلام خسارت شده بود.

۵. **عدم تطابق تاریخ وقوع خسارت با زمان اعلام خسارت:** طبق دستورالعمل‌های اجرایی صندوق بیمه، در خسارت زراعت می‌بایست یک هفته پس از بروز خسارت بازدید انجام شود؛ تا قبل از حذف آثار خسارت نوع و میزان آن به‌صورت دقیق بررسی شود. در بعضی از موارد مشاهده شد تاریخ اعلام یک‌ماه بعد از رخ دادن خسارت ثبت‌شده است. واضح است بعد از حذف آثار خسارت پرداخت غرامت می‌تواند مشکوک به نظر برسد چراکه ممکن است در گذشته اصلاً خسارتی وارد نشده باشد.

در انجام این پژوهش محدودیت‌هایی وجود داشت، ازجمله می‌توان به فقدان مجموعه‌داده استاندارد برای بیمه کشاورزی و فقدان ویژگی مختصات جغرافیایی در

مجموعه‌داده برای مشخص شدن محل دقیق کشت محصول برای جلوگیری از تکراری شدن نمونه‌ها اشاره کرد.

می‌توان بر اساس نتایج به‌دست‌آمده پیشنهادهای زیر را در جهت افزایش دقت تشخیص ادعاهای مشکوک ارائه داد:

۱. همان‌طور که ملاحظه شد، اغلب ادعاهای غیرعادی در دسته اول قرار گرفته‌اند. در این دسته به ادعاهای خسارتی غرامت پرداخت‌شده که مستندات کافی مبنی بر وارد شدن خسارت در پرونده آن‌ها وجود نداشته است؛ این موضوع ممکن است درنتیجه تبانی بین بیمه‌گذار با نماینده یا ارزیاب اتفاق افتاده باشد که نیازمند بررسی میدانی توسط کارشناسان بیمه می‌باشد.

۲. با ادغام مدل یادگیری ارائه‌شده در این پژوهش با سامانه جامع صندوق بیمه می‌توان ادعاهای غیرعادی خسارت را به‌موقع شناسایی کرده و از پرداخت غرامت جلوگیری نمود.

۳. با برچسب‌گذاری تعدادی از داده‌ها می‌توان از روش‌های یادگیری نیمه نظارت‌شده بهره‌برداری کرد، نتایج این روش دارای دقت بیشتری نسبت به یادگیری نظارت‌نشده خواهد بود.

۴. در بررسی ادبیات پژوهش، این موضوع نیز مشخص شد که کشاورزان در خارج از ایران در هر زمانی نمی‌توانند نسبت به کشت محصول اقدام نمایند و درصورتی‌که در خارج از بازه اعلامی از طرف شرکت بیمه اقدام به کشت نمایند پرداخت غرامت شامل آن‌ها نخواهد شد؛ این موضوع در بیمه کشاورزی ایران مطرح نیست و بایستی در شرایط عقد قرارداد بیمه‌نامه لحاظ گردد؛ چراکه کاشت در زمان نامناسب می‌تواند در بازده محصول نقش داشته و منجر به درخواست غرامت گردد.

با در نظر گرفتن محدودیت‌های پژوهش حاضر، برای رسیدن به‌دقت بالا، توصیه می‌شود در پژوهش‌های آتی سایر ویژگی‌های کلیدی از قبیل مختصات جغرافیایی، داده‌های جوی

و داده‌های سنسورهای اینترنت اشیاء[1] نیز استفاده شود تا نتایج دقیق‌تری به دست آید.

تعارض منافع

تعارض منافع وجود ندارد.

سپاسگزاری

از صندوق بیمه کشاورزی ایران بابت همکاری‌ها و حمایت‌های شایسته‌ای که انجام دادند تا این پژوهش به سرانجام برسد، سپاسگزاری می‌شود.

ORCID

Yaqub Ahmadlou **https://orcid.org/0000-0002-3778-094X**
Alireza Pourebrahimi **https://orcid.org/0000-0001-5741-0260**
Jafar Tanha **https://orcid.org/0000-0002-0779-6027**
Ali Rajabzade Ghatari **https://orcid.org/0000-0002-8470-3568**

1. Internet of Things (Iot)

منابع

۱. قباخلو، م.؛ رجب‌زاده قطری، ع.؛ طلوعی اشلقی، ع؛ و البرزی، م. (۱۴۰۱). طراحی سیستم پیشنهاد بانکی فردی با استفاده از تجزیه‌وتحلیل احساسات در رسانه‌های اجتماعی. *مطالعات مدیریت کسب‌وکار هوشمند*، ۱۰(۳۹)، ۲۸۹-۲۵۷. https://doi.org /10.22054/ims.2021.59775.1932

References

2. Al-Hashedi, K. G., & Magalingam, P. (2021). Financial fraud detection applying data mining techniques: A comprehensive review from 2009 to 2019. *Computer Science Review*, 40, 100402. https://doi.org/10.1016/j.cosrev.2021.100402
3. Bisong, E. (2019). Optimization for Machine Learning: Gradient Descent. In E. Bisong (Ed.), *Building Machine Learning and Deep Learning Models on Google Cloud Platform: A Comprehensive Guide for Beginners* (pp. 203-207). Apress. https://doi.org/10.1007/978-1-4842-4470-8_16
4. Brownlee, J. (2020). Data preparation for machine learning: data cleaning, feature selection, and data transforms in Python. Machine Learning Mastery. https://www.google.com/books/edition/Data_Preparation_for_Machine_Learning/uAPuDwAAQBAJ?hl=en&gbpv=1&dq=Data%20Preparation%20for%20Machine%20Learning&pg=PP1&printsec=frontcover
5. Brownlee, J. (2021). How to choose an activation function for deep learning. Machine Learning Mastery. https://machinelearningmastery.com/choose-an-activation-function-for-deep-learning/
6. Chalapathy, R., & Chawla, S. (2019). Deep learning for anomaly detection: A survey. *arXiv preprint arXiv:1901.03407*. https://doi.org/10.48550/arXiv.1901.03407
7. Chandola, V ,.Banerjee, A., & Kumar, V. (2009). Anomaly detection: A survey. *ACM computing surveys (CSUR)*, 41(3), 1-58. https://doi.org/10.1145/1541880.1541882
8. Crop Insurance Statistics. (2022). Cropinsurance.org. Retrieved July 22, 2022, from https://cropinsurance.org/wp-content/uploads/2021/02/2020-Crop-Insurance-Myths-v-Facts-Improper-Payment-Rate.pdf
9. Debener, J., Heinke, V., & Kriebel, J. (2023). Detecting insurance fraud using supervised and unsupervised machine learning. *Journal of Risk and Insurance*. https://doi.org/10.1111/jori.12427
10. Ekin, T., Lakomski, G., & Musal, R. M. (2019). An unsupervised Bayesian hierarchical method for medical fraud assessment. Statistical Analysis and Data Mining. *The ASA Data Science Journal*, 12(2), 116-124. https://doi.org/10.1002/sam.11408

11. Finke, T., Krämer, M., Morandini, A., Mück, A., & Oleksiyuk, I. (2021). Autoencoders for unsupervised anomaly detection in high energy physics. *Journal of High Energy Physics*, 2021(6), 1-32. https://doi.org/10.1007/JHEP06(2021)161
12. Fraud stats. (2020). Retrieved from https://insurancefraud.org/fraud-stats/
13. GAO. (2006). Crop insurance: More needs to be done to reduce program's vulnerability to fraud, waste, and abuse. Retrieved from https://www.gao.gov/assets/gao-06-878t.pdf
14. Gomes, C., Jin, Z., & Yang, H. (2021). Insurance fraud detection with unsupervised deep learning. *Journal of Risk and Insurance*, 88(3), 591-624. https://doi.org/10.1111/jori.12359
15. Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep learning. MIT press. https://www.google.com/books/edition/Deep_Learning/omivDQAAQBAJ?hl=en&gbpv=1&dq=deep+learning+goodfellow&pg=PR5&printsec=frontcover
16. Hilal, W., Gadsden, S. A., & Yawney, J. (2022). Financial Fraud: A Review of Anomaly Detection Techniques and Recent Advances. *Expert Systems with Applications*, *193*, 116429. https://doi.org/https://doi.org/10.1016/j.eswa.2021.116429
17. 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), 7285. https://doi.org/10.3390/s20247285
18. Kirlidog, M., & Asuk, C. (2012). A fraud detection approach with data mining in health insurance. *Procedia-Social and Behavioral Sciences*, *62*, 989-994. https://doi.org/https://doi.org/10.1016/j.sbspro.2012.09.168
19. Liu, W., Wang, Z., Liu, X., Zeng, N., Liu, Y., & Alsaadi, F. E. (2017). A survey of deep neural network architectures and their applications. *Neurocomputing*, *234*, 11-26. https://doi.org/https://doi.org/10.1016/j.neucom.2016.12.038
20. Marzen, C. G. (2013). Crop Insurance Fraud and Misrepresentations: Contemporary Issues and Potential Remedies. *SSRN Electronic Journal*, 675-707.
21. Miao, J., & Niu, L. (2016). A Survey on Feature Selection. *Procedia Computer Science*, 91, 919-926. https://doi.org/10.1016/j.procs.2016.07.111
22. Newlands, N., Ghahari, A., Gel, Y. R., Lyubchich, V., & Mahdi, T. (2019). Deep learning for improved agricultural risk management. https://scholarspace.manoa.hawaii.edu/bitstream/10125/59543/1/0103.pdf
23. Nian, K., Zhang, H., Tayal, A., Coleman, T., & Li, Y. (2016). Auto insurance fraud detection using unsupervised spectral ranking for anomaly. *The Journal of Finance and Data Science*, *2*(1), 58-75. https://doi.org/https://doi.org/10.1016/j.jfds.2016.03.001

24. Raschka, S., & Mirjalili, V. (2019). *Python machine learning: Machine learning and deep learning with Python, scikit-learn, and TensorFlow 2*. Packt Publishing Ltd. https://www.igi-global.com/pdf.aspx?tid%3D267132%26ptid%3D254262%26ctid%3D17%26t%3Dpython+machine+learning%3A+machine+learning+and+deep+learning+with+python%2C+scikit-learn%2C+and+tensorflow+2%2C+third+edition%26isxn%3D
25. Rezapour, M. (2019). Anomaly detection using unsupervised methods: credit card fraud case study. *International Journal of Advanced Computer Science and Applications*, *10*(11).
26. Sarker, I. H. (2021). Machine Learning: Algorithms, Real-World Applications and Research Directions. *SN Computer Science*, *2*(3), 160. https://doi.org/10.1007/s42979-021-00592-x
27. Yaram, S. (2016, 23-25 Aug. 2016). Machine learning algorithms for document clustering and fraud detection. Paper presented at the 2016 International Conference on Data Science and Engineering (ICDSE). https://doi.org/10.1109/ICDSE.2016.7823950
28. Zamini, M., & Montazer, G. (2018). Credit Card Fraud Detection using autoencoder based clustering. Paper presented at the 2018 9th International Symposium on Telecommunications (IST). https://doi.org/10.1109/ISTEL.2018.8661129
29. Zhang, C., Liu, J., Chen, W., Shi, J., Yao, M., Yan, X.,... Chen, D. (2021). Unsupervised Anomaly Detection Based on Deep Autoencoding and Clustering. Security and Communication Networks, 2021, 7389943. https://doi.org/10.1155/2021/7389943
30. Zhao, Y., Nasrullah, Z., & Li, Z. (2019). Pyod: A python toolbox for scalable outlier detection. *arXiv preprint arXiv:1901.01588.*

References [In Persian]

1. Ghobakhloo, M., Rajabzadeh, A., & Toloie, A., Alborzi, M. (2022) Designing a Banking Personalized Recommender System Using Sentiment Analysis in Social Media. *Journal of Business Intelligence Management Studies*, *10*(39), 257-289. https://doi.org/10.22054/ims.2021.59775.1932 [In Persian]

استناد به این مقاله: احمدلو، یعقوب. پورابراهیمی، علیرضا. تنها، جعفر. رجب زاده قطری، علی. (۱۴۰۲). مدلی برای تشخیص ادعاهای غیرعادی خسارت بیمه کشاورزی با استفاده از یادگیری عمیق ، *مطالعات مدیریت کسب وکار هوشمند*، ۱۲(۴۵)، ۳۱۳-۳۴۶.

DOI: doi.org/10.22054/ims.2023.69377.2213