Ind. Manag. Persp. 2026, 16(1), 133-169; doi: 10.48308/JIMP.16.1.133

Industrial Management Perspective

Journal hompage: https://jimp.sbu.ac.ir/?lang=en

Original Article

Designing a health insurance fraud detection system using artificial intelligence algorithms

Mojtaba Farrokh*
Sirous Sharifi**
Nasrin Hozarmoghadam***
Abbas Raad****
Alireza Norouzi*****

Extended Abstract

Introduction: With the rapid expansion of healthcare services, fraud in health insurance systems has become a serious challenge. This study aims to design and develop an intelligent and modular framework for fraud detection in health insurance. The framework is designed to identify abusive and fraudulent behaviors regardless of the type of service or actor involved, and to adapt effectively to dynamic and complex environments. The primary objective is to provide a flexible solution that enhances the accuracy of fraud detection while reducing human error in the decision-making process.

Methods: The proposed framework consists of four key modules. First, a knowledge-based module leverages insights from insurance and medical experts to build a simulation framework for fraud detection, enabling the medical-insurance team to describe and visualize abnormal behaviors based on the actions of different actors. Second, a two-stage data warehouse is designed to efficiently process large volumes of insurance data. In the first-stage warehouse, the ETL (extract–transform–load) process ingests claims data, cleanses data quality issues, and removes inconsistencies and errors to prepare the data for feature extraction required for fraud detection. In the second-stage warehouse, in collaboration with insurance and medical experts, relevant features for fraud detection are extracted and selected. To this end, a framework for simulating the fraud-detection process is built to enable the medical-insurance team to describe, analyze, and visualize abnormal behaviors based on the actions of different actors. Accordingly, a list of twenty key features for fraud detection was extracted and documented, covering information about actors, productss/services, and related features for each type of fraud. Third, the fraud detection engine is based on a proposed algorithm called K-IF, which first clusters data using Isolation Forest (IF) and then identifies suspicious samples using K-Means. Fourth, visualization tools and a dynamic management dashboard are developed to support interactive analysis and real-time updates by users.

Received: Aug. 31, 2025; Revised: Sep. 17, 2025; Accepted: Oct. 03, 2025; Published Online: Nov. 01, 2025.

*Assistant Professor, Department of Operations Management and Information Technology, Faculty of Management, Kharazmi University, Tehran, Iran.
Corresponding Author: Farrokh@khu.ac.ir

**Assistant Professor, Department of Banking, Insurance and Customs, Faculty of Management, Kharazmi University, Tehran, Iran.

***Assistant Professor, Insurance Research Institute, Tehran, Iran.

****Assistant Professor, Department of Industrial Management and Information Technology, Faculty of Management and Accounting, Shahid Beheshti University, Tehran, Iran.

*****Master's Student, Department of Operations Management and Information Technology, Faculty of Management, Kharazmi University, Tehran, Iran.

Ind. Manag. Persp. 2026, 16(1), 133-169; doi: 10.48308/JIMP.16.1.133

Industrial Management Perspective

Journal hompage: https://jimp.sbu.ac.ir/?lang=en

Original Article

Results and discussion: Experimental results on labeled datasets demonstrate that the proposed algorithm, by leveraging the discriminative power of IF and the clustering precision of K-Means, achieves better performance across multiple metrics and computational times than common algorithms such as LOF, OCSVM, EE, DBSCAN, AE, and K-Means. Furthermore, results from applying the proposed algorithm to real data from a health insurance company indicate that this approach, with reduced dependence on contamination rate and improved accuracy in detecting edge cases, demonstrates strong anomaly-detection capabilities. Ultimately, the framework has been developed as a software package for private insurance companies, offering advanced analytical tools that significantly enhance decision-making and reduce the need for human intervention.

Conclusion: This study highlights that success in detecting insurance fraud is directly tied to the quality and precision of features extracted from healthcare transaction data. The synergy between demographic, financial, and service-related data plays a crucial role in increasing the sensitivity of machine learning models to anomalous behaviors. However, the lack of accurate and structured data remains a major challenge in developing effective fraud detection software. The developed framework, designed as a software package for managing health insurance claims, integrates machine learning models, a modular architecture, and a modern user interface to deliver high scalability and rapid responsiveness to organizational needs. It is recommended that insurance company managers adopt this solution as part of their digital strategy for claims management. By integrating with existing systems and utilizing secure databases and interactive dashboards, they can achieve improved efficiency, greater transparency, and reduced fraud-related costs.

Keywords: Fraud Detection; Health Insurance; Unsupervised Anomaly Detection Algorithms; Isolation Forest; Software.

How to Cite: Farrokh, Mojtaba; Sharifi, Sirous; Hozarmoghadam, Nasrin; Raad, Abbas; Norouzi, Alireza (2026). Designing a health insurance fraud detection system using artificial intelligence algorithms. *Ind. Manag. Persp.*, 16(1), 133-169 *(In Persian)*.

چشم‌انداز مدیریت صنعتی، ۱۴۰۵، دوره ۱۶، شماره ۱، صص ۱۳۳ - ۱۶۹، doi : 10.48308/JIMP.16.1.133

چشم‌انداز مدیریت صنعتی

Journal hompage : https://jimp.sbu.ac.ir/?lang=en

نوع مقاله: پژوهشی

طراحی سیستم کشف تقلب بیمه درمان به کمک الگوریتم‌های هوش مصنوعی

مجتبی فرخ*
سیروس شریفی**
نسرین حضارمقدم***
عباس راد****
علیرضا نوروزی*****

چکیده گسترده

مقدمه و اهداف: با گسترش روزافزون خدمات درمان، تقلب در نظام‌های بیمه درمانی به یک چالش جدی تبدیل شده است. پژوهش حاضر با هدف طراحی و توسعه یک چارچوب هوشمند و ماژولار برای کشف تقلب در بیمه درمانی انجام شده است. این چارچوب به‌گونه‌ای طراحی شده که مستقل از نوع خدمت یا بازیگر، توانایی شناسایی رفتارهای سوءاستفاده‌گرانه و تقلبی را داشته باشد و بتواند با محیط‌های پیچیده و پویا سازگار شود. هدف اصلی، ارائه راهکاری منعطف برای ارتقای دقت در تشخیص تقلب و کاهش خطاهای انسانی در فرآیند کشف تقلب بیمه درمان است.

روش‌ها: چارچوب پیشنهادی شامل چهار ماژول کلیدی است: نخست، ماژول دانش‌محور که با بهره‌گیری از دیدگاه‌های کارشناسان بیمه و پزشکی، یک فریم‌ورک برای شبیه‌سازی فرایند تشخیص تقلب ایجاد می‌شود تا تیم پزشکی-بیمه بتواند رفتارهای غیرعادی را بر اساس رفتار بازیگران مختلف توصیف و نمایش دهد. دوم، یک انبار داده دو مرحله‌ای برای پردازش کارآمد داده‌های حجیم بیمه طراحی شده است؛ در انبار داده مرحله اول فرایند استخراج، تبدیل و بارگذاری داده‌های ادعاهای بیمه‌ای انجام می‌شود، نواقص داده‌ای اصلاح و ناسازگاری‌ها و خطاها از بین می‌روند تا داده‌ها برای استخراج ویژگی‌ها لازم برای کشف انواع تقلب مناسب شوند. در انبار داده مرحله دوم ویژگی‌های مرتبط با تقلب با همکاری متخصصان استخراج و انتخاب می‌گردند. فهرستی از بیست ویژگی مؤثر برای تشخیص تقلب استخراج و مستندسازی شد که برای هر نوع تقلب، اطلاعات مربوط به بازیگران، کالاها و ویژگی‌های مرتبط را دربرمی‌گیرد. سوم، موتور کشف تقلب براساس یک الگوریتم پیشنهادی موسوم به K-IF است که ابتدا با استفاده از الگوریتم جنگل ایزوله (IF) داده‌ها را خوشه‌بندی کرده و سپس با الگوریتم K-Means نمونه‌های مشکوک را شناسایی می‌کند. چهارم، ابزارهای تجسم و داشبورد مدیریتی برای تحلیل تعاملی و به‌روزرسانی پویا توسط کاربران طراحی و ارائه می‌شود.

تاریخ دریافت: ۱۴۰۴/۰۶/۰۹، تاریخ بازنگری: ۱۴۰۴/۰۶/۲۶، تاریخ پذیرش: ۱۴۰۴/۰۷/۱۱، تاریخ اولین انتشار: ۱۴۰۴/۰۸/۱۰.

*استادیار، گروه مدیریت عملیات و فناوری اطلاعات، دانشکده مدیریت، دانشگاه خوارزمی، تهران، ایران.

نویسنده مسئول : Farrokh@khu.ac.ir

**استادیار، گروه بانک، بیمه و گمرک، دانشکده مدیریت، دانشگاه خوارزمی، تهران، ایران.

***استادیار، پژوهشکده بیمه، تهران، ایران.

****استادیار، گروه مدیریت صنعتی و فناوری اطلاعات، دانشکده مدیریت و حسابداری، دانشگاه شهید بهشتی، تهران، ایران.

*****دانشجوی کارشناسی ارشد، گروه مدیریت عملیات و فناوری اطلاعات، دانشکده مدیریت، دانشگاه خوارزمی، تهران، ایران.

چشم‌انداز مدیریت صنعتي، ۱۴۰۵، دوره ۱۶، شماره ۱، صص ۱۳۳ - ۱۶۹، doi : 10.48308/JIMP.16.1.133

Industrial Management Perspective

چشم‌انداز مدیریت صنعتی

Journal hompage : https://jimp.sbu.ac.ir/?lang=en

نوع مقاله: پژوهشی

یافته‌ها: نتایج آزمایش‌های انجام‌شده بر روی مجموعه داده‌های برچسب‌دار نشان می‌دهد که الگوریتم پیشنهادی با بهره‌گیری از قدرت تفکیک IF و دقت خوشه‌بندی K-Means، عملکرد بهتری از نظر شاخص‌های عملکردی و زمان محاسباتی نسبت به الگوریتم‌های رایج مانند LOF، OCSVM، EE، DBSCAN، AE و K-Means داشته است. همچنین، اجرای این الگوریتم بر روی داده‌های واقعی شرکت بیمه دی نشان داد که وابستگی به نرخ آلودگی کاهش یافته و دقت در شناسایی نقاط لبه‌ای افزایش یافته است. در نهایت، این چارچوب به‌صورت یک بسته نرم‌افزاری برای شرکت‌های بیمه خصوصی توسعه یافته و با ارائه ابزارهای تحلیلی پیشرفته، نقش مؤثری در ارتقای تصمیم‌گیری و کاهش نیاز به مداخله انسانی ایفا می‌کند.

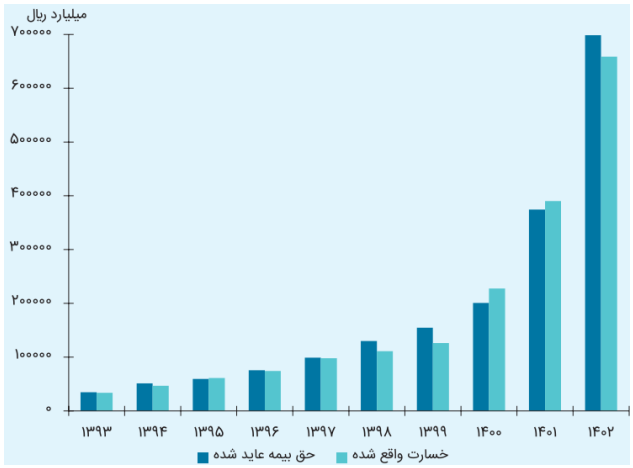
نتیجه‌گیری: پژوهش حاضر نشان می‌دهد که موفقیت در کشف تقلب‌های بیمه‌ای به‌طور مستقیم به کیفیت و دقت ویژگی‌های استخراج‌شده از داده‌های تراکنش‌های درمانی وابسته است. هم‌افزایی میان داده‌های جمعیتی، مالی و خدماتی نقش مهمی در افزایش حساسیت مدل‌های یادگیری ماشین نسبت به رفتارهای ناهنجار ایفا می‌کند، در حالی که کمبود داده‌های دقیق و ساختارمند یکی از چالش‌های اساسی در توسعه نرم‌افزارهای تشخیص تقلب محسوب می‌شود. چارچوب توسعه‌یافته به‌صورت بسته نرم‌افزاری برای مدیریت ادعاهای بیمه درمانی طراحی شده و با ترکیب مدل‌های یادگیری ماشین، معماری ماژولار و رابط کاربری مدرن، قابلیت گسترش‌پذیری بالا و پاسخگویی سریع به نیازهای سازمانی را فراهم می‌آورد و پیشنهاد می‌شود مدیران شرکت‌های بیمه این راهکار را به‌عنوان بخشی از استراتژی دیجیتال‌سازی مدیریت ادعاها به کار بگیرند تا با ادغام با سامانه‌های موجود و استفاده از پایگاه داده امن و داشبوردهای تعاملی، بهبود کارایی، شفافیت و کاهش هزینه‌های تقلب را تحقق بخشند.

واژه‌های کلیدی: کشف تقلب؛ بیمه درمان؛ الگوریتم‌های کشف ناهنجاری بدون ناظر؛ جنگل ایزوله؛ نرم‌افزار.

استناددهی: فرخ، مجتبی؛ شریفی، سیروس؛ حضارمقدم، نسرین؛ راد، عباس؛ نوروزی، علیرضا (۱۴۰۵). طراحی سیستم کشف تقلب بیمه درمان به‌کمک الگوریتم‌های هوش مصنوعی. چشم‌انداز مدیریت صنعتی، ۱۶(۱)، ۱۳۳-۱۶۹.

۱. مقدمه

افزایش مستمر امید به زندگی به دلیل پیشرفت‌های علوم بهداشتی، بهبود استانداردهای زندگی و آگاهی فزاینده از سبک‌های زندگی سالم، به‌طور قابل توجهی سهم خدمات درمان را در اقتصاد جهانی گسترش داده است [۴۲]. با این حال، با گسترش روزافزون خدمات درمان، تقلب در بیمه درمان به یک چالش جدی تبدیل شده است. تقلب در بیمه درمان نه‌تنها به هدر رفت منابع مالی بیمارستان‌ها و شرکت‌های بیمه منجر می‌شود، بلکه به کیفیت خدمات درمانی نیز آسیب وارد می‌کند [۶]. افزایش تعداد خسارت‌های بیمه درمان در سطح جهانی و پیچیدگی خسارت‌های بیمه در صنعت بهداشت و درمان در سال‌های اخیر، استفاده از سیستم‌های هوشمند و مؤثر مدیریت خسارت را ضروری می‌سازد [۲۶]. براساس شکل ۱، حق بیمه عاید شده در بازه زمانی ۱۳۹۳ تا ۱۴۰۲، از ۳۴۰۰۰ میلیارد ریال در سال ۱۳۹۳ به ۶۹۸۰۰۰ میلیارد ریال در سال ۱۴۰۲ رسیده است. همچنین، در سال ۱۴۰۲، برآورد می‌شود که حجم خسارت‌های ناشی از تقلب در این سال بین ۶۹۵۰ تا ۱۰۴۲۰ میلیارد تومان باشد. این آمار اهمیت توجه ویژه به میزان خسارت‌ها و نقش تقلب در این صنعت را نشان می‌دهد (سالنامه آماری بیمه مرکزی ج.ا.ایران، ۱۴۰۲).



شکل ۱. میزان خسارت واقع شده و حق بیمه عاید شده در بیمه درمان کشور

امروزه استفاده از فناوری‌های پیشرفته که از تجزیه و تحلیل داده‌ها، یادگیری ماشین و هوش مصنوعی برای تسریع در پردازش خسارت‌ها، اتوماسیون عملیات تکراری و تصمیم‌گیری دقیق‌تر در پرداخت خسارت‌ها استفاده می‌کنند، به طور چشمگیری افزایش یافته است. شرکت‌های بیمه می‌توانند خسارت‌ها را سریع‌تر ارزیابی کنند، فعالیت‌های تقلبی را شناسایی کنند و منابع را به طور مؤثرتر تخصیص دهند [۷]. در حوزه بیمه درمان، تقلب به هر نوع اقدام نادرست و عمدی اطلاق می‌شود که با هدف فریب و سوءاستفاده از سیستم بیمه انجام می‌شود، و منجر به دریافت غیرموجه مزایا، افزایش هزینه‌ها یا سوءاستفاده از منابع می‌شود. تقلب‌های بیمه همچنین می‌تواند شامل فعالیت‌های تقلبی گروهی باشد که با همکاری پزشک، داروخانه و بیمه‌گذار در حال فریب سیستم هستند تا مبالغ غیرمجاز دریافت کنند. تقلب در این حوزه می‌تواند به صورت فردی یا گروهی رخ دهد و انواع مختلفی دارد که می‌توان آنها را براساس نوع عامل مرتکب و شکل رفتار تقلبی، دسته‌بندی کرد. با توجه به اینکه مسئله تقلب دارای ابعاد پیچیده و متنوعی است، شناسایی و کشف الگوهای تقلب به یک نیاز ضروری برای کاهش هزینه های شرکت‌های بیمه تبدیل شده است. در این حوزه، الگوریتم‌های هوش مصنوعی[1] می‌توانند الگوهای نادر و مغایرت‌های غیرعادی، مانند تکرار ادعاهای مشابه با مشخصات ناسازگار، کدگذاری نادرست یا دستکاری شده، و اختلاف‌های هزینه‌ای بی‌مورد را شناسایی کنند. تحلیل این داده‌ها با کمک الگوریتم‌های هوش مصنوعی، نقش کلیدی در کشف موارد تقلب و جلوگیری از هدررفت منابع بیمه دارد [۸، ۲۷]. با این حال، با وجود پیشرفت‌های قابل توجه در این حوزه، هنوز چالش‌هایی در زمینه پیاده‌سازی و به کارگیری هوش مصنوعی در کشف تقلب بیمه درمان وجود

1. Artificial Intelligence (AI)

دارد. از جمله این چالش‌ها می‌توان به کمبود داده‌های با کیفیت، پیچیدگی در تنظیم الگوریتم‌های هوش مصنوعی و ناهمگونی داده‌های بیمه درمان اشاره کرد [۴۰]. از سوی دیگر، تقلب‌هایی که در بیمه درمان به صورت گروهی و با همکاری افراد مختلف مانند بیمار و پزشک انجام می‌شود، نیازمند تعریف ویژگی‌هایی از روی داده‌های موجود در پایگاه داده شرکت‌های بیمه بکمک کارشناسان بیمه و پزشکان به عنوان ورودی‌های الگوریتم‌های هوش مصنوعی برای سنجش و کشف آن‌ها دارد [۱۹، ۳۱]. در فرآیند کشف تقلب، یکی از چالش‌های اصلی این است که ابتدا باید ویژگی‌های مناسب و قابل اندازه‌گیری برای تشخیص تقلب تعریف شود، به طوری که پس از آموزش الگوریتم، بتوان موارد مشکوک را به سرعت شناسایی کرد. در پژوهش حاضر، نوع داده‌های جمع‌آوری شده عمدتاً از نوع داده‌های ساختاریافته هستند، زیرا تمامی اطلاعات نسخه‌های پزشکی در قالب فرمت‌های منظم و قابل دسته‌بندی در پایگاه داده شرکت‌های بیمه ثبت شده است. این نوع داده‌ها کارایی بالایی در تحلیل‌های داده‌محور دارند و امکان استفاده از الگوریتم‌های آماری و یادگیری ماشین برای کشف تقلب‌های بیمه درمانی را فراهم می‌آورد. از سوی دیگر، در حوزه شناسایی تقلب در بیمه های درمان، بسیاری از پروندهای موجود در پایگاه داده شرکت‌های بیمه فاقد برچسب تقلب هستند. در پژوهش جاری، چون نسبت نمونه‌های تقلب در داده‌های بیمه درمان کم است و داده‌ها فاقد برچسب نیز هستند، بسیاری از محققان برای داده‌هایی با این مشخصات استفاده از روش‌های کشف ناهنجاری[1] بر مبنای روش‌هایی با توانایی تشخیص ناهنجاری‌های نادر را پیشنهاد داده‌اند تا با دقت بالا بتوان تقلب‌های واقعی را شناسایی کرد، بدون اینکه منجر به تعداد زیادی هشدار نادرست یا موارد از دست رفته گردد [۶، ۲۷]. روش‌های کشف ناهنجاری مبتنی بر الگوریتم‌های بدون ناظر مانند الگوریتم شناسایی ناپایدارهای محلی[2] (LOF)، جنگل ایزوله[3] (IF)، خوشه‌بندی فضایی مبتنی بر چگالی برنامه‌ها با نویز[4] (DBSCAN) و خودرمزگذارها[5] (AE) و روش‌های دیگر هستند که در کشف تقلب بیمه درمان استفاده شده‌اند [۱۴، ۳۲، ۳۹].

این تحقیق با تعریف ویژگی‌های کشف تقلب و توسعه یک الگوریتم بدون ناظر مناسب در کشف تقلب در سیستم‌های بیمه درمان، به دنبال ارائه یک چارچوب ماژولار و منعطف مبتنی بر الگوریتم‌های هوش مصنوعی برای کشف تقلب در بیمه درمان است تا با شناسایی و تحلیل موارد تقلب، به مدیران و سیاست‌گذاران در اتخاذ تصمیم‌های مناسب برای کاهش هزینه‌های مرتبط با تقلب کمک کند. از سوی دیگر، نبود یک نرم‌افزار هوشمند برای کشف تقلب یکی دیگر چالش‌های کشف تقلب در شرکت‌های بیمه است که تحقیقات محدودی به آن پرداخته‌اند. به طور کلی، در این پژوهش تلاش می‌شود با توسعه یک چارچوب منعطف و ماژولار برای کشف هوشمند تقلب بیمه درمان، ویژگی‌هایی که منجر به بروز تقلب می‌شود را استخراج و در نهایت یک نرم‌افزار مبتنی بر بهترین الگوریتم‌های مختلف هوش مصنوعی، برای کشف تقلب ایجاد شود تا با دقت مناسب به شرکت‌های بیمه در تشخیص خودکار تقلب در بیمه درمان کمک کند. با توجه به بدون برچسب بودن داده‌ها، از الگوریتم‌های کشف ناهنجاری و از شاخص‌های اعتبارسنجی مناسب برای ارزیابی این الگوریتم‌ها استفاده خواهد شد. نوآوری این پژوهش طراحی یک چارچوب ماژولار کشف تقلب هوشمند و یک نرم‌افزار مبتنی بر آن برای ارائه خدمات کشف تقلب بیمه درمان به شرکت‌های بیمه‌ای خواهد بود، به طوری که در یک حالت تعاملی و ماژولار همواره بتوان این سیستم را توسعه داد.

۲. مبانی نظری و پیشینه پژوهش

۱.۲. کشف تقلب در بیمه درمان

در پژوهش جاری، بر روی تقلب و سوءاستفاده[6] تمرکز خواهیم شد، که به صورت رفتار عمدی توسط بیماران و/یا ارائه‌دهندگان پزشکی برای ایجاد مزایای غیرموجه برای خود یا افراد مرتبط تعریف شده است [۲۲]. تقلب و سوءاستفاده در خدمات درمانی تنها از سوی ارائه‌دهندگان پزشکی انجام نمی‌شود. بیماران بیمه‌شده، تأییدکنندگان خدمات و سایر بازیگران حوزه درمان نیز در اقدامات تقلبی و سوءاستفاده می‌توانند دخیل باشند

1. Anomaly Detection
2. Local Outlier Factor (LOF)
3. Isolation Forest (IF)
4. Density-Based Spatial Clustering of Applications with Noise (DBSCAN)
5. Autoencoders (AE)
6. Fraud and Abuse

[۳]. علاوه‌براین، تقلب و سوءاستفاده غالباً براساس همکاری یا حداقل همدستی بین بازیگران (به‌عنوان مثال: پزشک و داروساز، ارائه‌دهنده و بیمه‌شده و غیره) انجام می‌شوند. با این حال، بسیاری از مطالعات درباره تقلب و سوءاستفاده و روش‌های تحلیلی مرتبط با شناسایی آن‌ها، به دلیل پتانسیل بالای صرفه‌جویی و مبنای داده‌ای وسیع‌تر بر روی ارائه‌دهندگان پزشکی تمرکز دارند [۳].

متقلبین بیمه‌های درمان را می‌توان به سه گروه بزرگ شامل بیماران، ارائه دهندگان و شرکت‌های بیمه تقسیم کرد. ارائه‌دهندگان خدمات درمانی شامل پزشک، بیمارستان، داروخانه‌ها، آمبولانس و آزمایشگاه‌ها؛ بیمه‌گذار شامل بیماران و کارفرمایان بیماران [۱۴]. در این میان انواع مختلف تقلب توسط هر یک از این گروه‌ها ممکن است صورت گیرد. در میان انواع رفتارهای متقلبانه، ارتقا کد[1]، به معنی طبقه‌بندی بیمار در یک گروه تشخیص مربوطه[2] (DRG) با کدهای بالاتر است که بازپرداخت‌های بالاتری را به همراه دارد. به عبارت دیگر، جایی که ارائه‌دهندگان برای خدماتی که هزینه بیشتری دارند، صورتحساب صادر می‌کنند در حالی‌که واقعاً انجام نشده‌اند، نشانه‌ای از تقلب است [۲]. این عمل یکی از معروف‌ترین نوع تقلب کدگذاری است که بیماری بیماران را به بالاترین گروه درمانی ممکن به منظور مطالبات بازپرداخت بیشتر تبدیل می‌کند [۲۳]. صورتحساب تکراری[3]، که به معنای صدور صورتحساب برای همان رویه‌ها به‌صورت دوباره با برخی تغییرات کوچک است، نیز نشانه‌ای از تقلب است [۴۳]. از سوی دیگر، تفکیک صورتحساب[4] به معنی جدا کردن یا تفکیک خدماتی است که باید به صورت یکجا ارائه شوند، اما فرد ارائه‌دهنده آن‌ها را به صورت جداگانه فاکتور می‌کند تا درآمد بیشتری به دست آورد. صورت‌حساب جداگانه برای خدمات گروهی نوعی کلاهبرداری بیمه درمان به معنای استفاده غیرقانونی و نادرست از خدمات بیمه‌ای و ارسال صورت‌حساب‌های جداگانه برای خدماتی که باید یکجا محاسبه شوند به منظور دریافت مزایای بیشتری انجام می‌شود [۱۵]. در بیمه درمان، این نوع تقلب ممکن است شامل ثبت ادعاهای بی‌ربط و غیرواقعی، جعلی یا بیش از حد نیاز باشد که باعث خسارت‌های مالی به سیستم بیمه می‌شود. صدور صورتحساب برای خدمات غیرپوشش داده شده نیز نشانه‌ای از تقلب است [۴۸]. همچنین می‌توان به رفتارهای ارائه‌دهندگان برای شناسایی نشانه‌های تقلب نگاه کرد. نسبت‌های غیرمعمول ملاقات‌های بیمار و عدم تطابق بین تشخیص‌ها و برنامه‌های درمانی معمولاً نشانه‌هایی از کلاهبرداری به شمار می‌روند. یک برنامه درمانی ناکافی که تعداد بیماران آن بیشتر از حدی باشد که یک ارائه‌دهنده توانایی مدیریت آن را دارد، گاهی نشان دهنده‌ی احتمال تقلب است [۱۵].

شرکت‌های بیمه خصوصی و عمومی به‌عنوان تأمین‌کنندگان اصلی تأمین مالی خدمات درمانی، انگیزه قوی برای جلوگیری از خسارات ناشی از تقلب و سوءاستفاده دارند. با این حال، اغلب هیچ رویکرد سیستماتیکی برای رسیدگی به موضوع کشف تقلب در این شرکت‌ها وجود ندارد. در بسیاری از شرکت‌های بیمه، تمرکز اصلی معمولاً بر رفتار تاریخی ارائه‌دهندگان و همچنین تشخیص‌ها و رویه‌هایی است که معمولاً به موارد تقلب و سوءاستفاده مرتبط هستند، در حالی که به رفتار اعضا و تعامل بازیکنان مختلف اهمیت کمتری می‌دهند. در این شرایط، الگوهای جدید و پیچیده تنها به‌طور تصادفی شناسایی می‌شوند [۱۸]. از سوی دیگر، تقلب و رفتارهای سوءاستفاده معمولاً به مجموعه‌ای از ادعاها تقسیم‌بندی می‌شوند و فقط پس از تجزیه و تحلیل می‌توان آن‌ها را شناسایی کرد. به عبارت دیگر، تصمیم‌گیری در مورد اینکه آیا یک تراکنش خاص تقلبی یا سوءاستفاده‌ای است، بسیار دشوار است. به‌جای آن، تراکنش‌های قبلی توسط همان بازیگران باید در تجزیه‌وتحلیل مورد توجه قرار گیرد [۵]. یک تراکنش که به‌نظر بسیار عادی می‌رسد، در واقع می‌تواند بخشی از مجموعه‌ای از تراکنش‌های تقلبی در طی یک دوره زمانی باشد. بنابراین، برای یک فرآیند مؤثر شناسایی رفتارهای تقلبی و سوءاستفاده، تحلیل دقیقی مورد نیاز است [۱۵]. به‌عنوان یک مثال، ممکن است برای دو تراکنش با محتوای دقیقاً یکسان از نظر مجموعه داده‌های الکترونیکی ادعا، یکی از آن‌ها بخشی از تقلب باشد و دیگری نباشد، بستگی به تراکنش‌های قبلی بازیگران درگیر در هر تراکنش دارد.

به طورکلی، شناسایی رفتارهای تقلب و سوءاستفاده کار آسانی نیست و نیاز به رویکردهای حل هوشمندانه‌تری دارد که برای کاربران شهودی

1. Uncoding
2. Diagnosis Related Group (DRG)
3. Duplicate billing
4. Unbounding

و مبتنی بر روش‌های شهودی است زیرا ذات این مسئله بسیار غیرخطی و پیچیده است [۲۰]. از آن‌جایی که بررسی دستی ادعاهای مشکوک یک روش بسیار پرهزینه برای شناسایی رفتارهای تقلب و سوءاستفاده است و عملکرد آن به‌شدت سوال‌برانگیز می‌باشد، شرکت‌های بیمه در تلاشند تا یک سیستم هوشمند شناسایی تقلب و سوءاستفاده توسعه دهند که از داده‌های واقعی جمع‌آوری‌شده از طریق سیستم‌های مدیریت ادعا[1] استفاده کند. در سال‌های اخیر، محققان علاقه‌مندی فزاینده‌ای به تحقیقات شناسایی تقلب و سوءاستفاده، به‌ویژه آن دسته که از تکنیک‌های مبتنی بر روش‌های هوش مصنوعی استفاده می‌کنند، نشان داده‌اند [۳۳].

۲.۲. الگوریتم‌های کشف ناهنجاری

شناسایی تقلب‌های گروهی نیازمند بررسی روابط میان بازیگران مختلف است، برای این منظور نمی‌توان داده‌های برچسب‌دار را از قبل تهیه کرد، بنابراین مدل‌های یادگیری بدون نظارت[2] مناسب هستند. در مقایسه با یادگیری نظارت‌شده[3]، یادگیری غیرنظارت‌شده ممکن است کمتر قابل تفسیر باشد، زیرا الگوها و روابط کشف‌شده ممکن است ساده و شهودی نباشند و ارزیابی و اعتبارسنجی نتایج تحلیل دشوار باشد، زیرا هیچ برچسبی برای مقایسه وجود ندارد [۲۹]. در این زمینه، از یادگیری غیرنظارت‌شده تحت عنوان کشف ناهنجاری یاد می‌شود [۲۱، ۲۷]. شناسایی ناهنجاری بدون نظارت به فرایندی اطلاق می‌شود که در آن تلاش می‌شود داده‌هایی که به‌طور غیرعادی یا منحصربه‌فرد هستند، شناسایی شوند بدون اینکه نیاز به داده‌های برچسب‌گذاری شده یا اطلاعات قبلی خاصی در مورد ناهنجاری‌ها داشته باشد [۳۴، ۴۹]. برای شناسایی ناهنجاری‌ها می‌توان از الگوریتم‌های مختلف یادگیری بدون نظارت استفاده کرد، که شامل:

تحلیل مؤلفه‌های اصلی (PCA)[4]: تکنیکی برای کاهش ابعاد که داده‌های با ابعاد بالاتر را به فضایی با ابعاد پایین‌تر تبدیل می‌کند و در عین حال تا حد امکان اطلاعات اصلی را حفظ می‌کند [۴۰].

جنگل ایزوله (IF). یک الگوریتم شناسایی نقاط ناهنجاری بدون نظارت است که بر مبنای مفهوم جداسازی ناهنجاری‌ها کار می‌کند. ناهنجاری‌ها معمولاً از سایر نقاط عادی فاصله بیشتری دارند و می‌توانند به راحتی شناسایی شوند [۲۲]. این ویژگی به الگوریتم اجازه می‌دهد تا ناهنجاری‌ها را به‌سرعت شناسایی کند. امتیاز ناهنجاری یک نمونه براساس میانگین طول مسیر در درخت ایزوله محاسبه می‌شود. هرچه میانگین طول مسیر کمتر باشد، احتمال ناهنجاری بودن نمونه بیشتر است. همچنین این الگوریتم نیازمند اطلاعاتی درباره نرخ آلودگی[5] (نسبت تقلب) به عنوان پارامتر اصلی در داده‌های آموزش است که در سناریوهای داده‌های بزرگ ارائه آن مشکل‌ساز است. این نرخ باید در دامنه (۰.۵، ۰) قرار گیرد، که مقدار پیش‌فرض آن در الگوریتم IF برای آموزش مدل ۰.۱ تعیین شده است [۱]. این الگوریتم می‌تواند به‌راحتی با حجم بالای داده‌ها کار کند و نسبت به تغییرات در داده‌ها مقاوم است [۲۲].

K-Means. یک الگوریتم خوشه‌بندی مبتنی بر مرکز خوشه است. الگوریتم K-means تعداد خوشه‌ها و مجموعه داده‌ها را به عنوان ورودی می‌گیرد و نتایج را به صورت مجموعه‌ای از خوشه‌ها ارائه می‌دهد. در این روش، مقدار میانگین نمونه‌های درون هر خوشه به عنوان مرکز آن خوشه تعریف می‌شود. سپس، هر نمونه به نزدیک‌ترین خوشه اختصاص داده می‌شود، که براساس فاصله اقلیدسی بین میانگین خوشه و نمونه انجام می‌شود. پس از آن، در یکسری تکرار موقعیت مرکز هر خوشه بهینه‌سازی می‌شود. برای هر خوشه، با استفاده از نمونه‌های اختصاص یافته در تکرار قبلی، میانگین جدید به عنوان مرکز تازه تعیین می‌شود. سپس، تمامی نمونه‌های هر خوشه به مراکز اصلاح‌شده تخصیص داده می‌شوند [۱۶، ۲۵، ۳۶]. این تکرار ادامه می‌یابد تا دیگر نیاز به جابه‌جایی مراکز تمامی خوشه‌ها نباشد. در این الگوریتم، تعداد خوشه‌ها (K) می‌تواند نتایج متفاوتی را به همراه داشته باشد. بنابراین، یافتن مقدار بهینه K اهمیت زیادی دارد. نمره سیلوئت[6] برای ارزیابی عملکرد این الگوریتم و تعیین تعداد بهینه خوشه‌ها استفاده می‌شود و میزان تفکیک و انسجام خوشه‌ها را می‌سنجد [۱۶].

خوشه‌بندی فضایی مبتنی بر چگالی برنامه‌ها با نویز (DBSCAN). الگوریتم خوشه‌بندی مبتنی بر چگالی است که داده‌ها را براساس چگالی

1. Claim management system
2. Unsupervised learning
3. Supervised learning
4. Principal Component Analysis (PCA)
5. Contamination ratio
6. Silhouette Score

نقاط در فضا گروه‌بندی می‌کند. این الگوریتم به‌جای تعیین تعداد خوشه‌ها، با استفاده از دو پارامتر اصلی: حداقل نقاط (MinPts) و حداکثر فاصله (Eps)، مناطق با چگالی بالا را شناسایی می‌کند. نقاطی که دارای حداقل تعداد مشخصی از همسایه‌ها در محدوده فاصله Eps هستند، به‌عنوان هسته خوشه تلقی می‌شوند. سایر نقاط اطراف این نقاط، به‌عنوان همسایه در نظر گرفته می‌شوند. این کار ادامه می‌یابد تا زمانی که کل منطقه متصل شناخته شود. برای شناسایی ناهنجاری، نقاطی که در هیچ خوشه‌ای قرار نمی‌گیرند به‌عنوان ناهنجاری درنظر گرفته می‌شوند. این الگوریتم‌ها برای شناسایی خوشه‌های متراکم از داده‌ها و تشخیص نقاط پرت به‌خصوص در فضاهای با ابعاد زیاد مؤثر هستند [۳۲].

الگوریتم شناسایی ناپایدارهای محلی (LOF). یک روش مبتنی بر داده‌های محلی است که برای شناسایی نقاط خارج از توزیع در مجموعه داده‌ها طراحی شده است. این الگوریتم با مقایسه چگالی محلی هر نقطه با چگالی نقاط همسایه آن، میزان تفاوت و ناهنجاری آن را ارزیابی می‌کند. به عبارت دیگر، LOF در واقع ناهنجاری‌ها را براساس تفاوت در چگالی‌های محلی نسبت به همسایگان محاسبه می‌کند، به‌طوری که نقاط خارج از توزیع که در مناطق کم‌چگال‌تر قرار دارند، امتیاز LOF بالایی دریافت می‌کنند و به عنوان ناهنجار شناخته می‌شوند. این روش به‌ویژه در محیط‌هایی با توزیع داده‌های ناهمگن و ساختارهای پیچیده کاربرد دارد، زیرا می‌تواند ناهنجاری‌های محلی و نه تنها کلی را شناسایی کند و قابلیت تشخیص ناهنجاری‌های پنهان در بسترهای مختلف را دارا می‌باشد [۱۴، ۳۹].

الگوریتم SVM تک کلاس (OCSVM): این الگوریتم با استفاده از داده‌های نرمال، یک مرز تصمیم‌گیری در فضای ویژگی‌ها ایجاد می‌کند که بیشترین داده‌ها را در بر گیرد و نقاطی که خارج از این مرز قرار می‌گیرند به‌عنوان ناهنجار یا مشکوک شناسایی می‌شوند. این الگوریتم بر پایه نگاشت داده‌ها به فضای ویژگی‌های با ابعاد بالا و یافتن یک ابرصفحه (یا تابع تصمیم) عمل می‌کند که داده‌های نرمال را از مابقی فضای داده جدا کند. این روش به‌ویژه در کاربردهایی مانند تشخیص تقلب، شناسایی نفوذ در شبکه، و پایش سلامت سیستم‌ها بسیار مؤثر است، زیرا بدون نیاز به داده‌های برچسب‌خورده ناهنجار، می‌تواند رفتارهای غیرعادی را شناسایی کند [۶].

الگوریتم Elliptic Envelope (EE). این الگوریتم یک مدل احتمالی است که برای داده‌های چندمتغیره استفاده می‌شود تا نقاط ناهنجاری را با تخمین پارامتر کواریانس استوار شناسایی کند. این مدل فرض می‌کند داده‌ها از توزیع بیضی‌شکل پیروی می‌کنند و مناطق درون‌گردان[1] و خارج‌گردان[2] را براساس شکل داده‌ها برآورد می‌کند. نقاط داده که خارج از محدوده بیضی تخمین‌زده‌شده قرار می‌گیرند، به‌عنوان ناهنجاری‌ها طبقه‌بندی می‌شوند [۴]. این الگوریتم برای داده‌های نسبتاً کوچک تا متوسط و با توزیع تقریبی نرمال مناسب است و مزیت آن را می‌توان در تعیین محدوده نرمال داده و تشخیص سریع ناهنجاری‌ها دید، اما حساسیت بالایی به وجود داده‌های پرت شدید و تغییرات قابل توجه خوشه‌ها دارد و در چنین شرایطی ممکن است بیضی فرضی به درستی نمایشی از ناحیه عادی ارائه ندهد [۴۷].

خودرمزگذارها (AE): نوعی از شبکه‌های عصبی که برای یادگیری غیرنظارت‌شده به‌کار می‌روند و برای بازسازی داده‌های ورودی آموزش داده می‌شوند. آن‌ها برای وظایفی مانند حذف نویز از تصاویر و شناسایی ناهنجاری‌ها استفاده می‌شوند [۴۴].

۳.۲. پیشینه پژوهش

تشخیص تقلب در داده‌های بیمه درمان بسیار دشوار است، زیرا انواع خدمات ارائه شده توسط ارائه‌دهندگان (پزشک، دارخانه‌ها، غیره) و الگوهای تعامل بازیگران بسیار متفاوت است، به همین دلیل داده‌های بیمه درمان ویژگی‌های مختلفی دارد و توزیع الگوهای آن بسیار نامنظم است [۱۸]. علاوه‌براین، برای کشف ناهنجاری مدل‌های مختلف یادگیری ماشین مبتنی بر خوشه‌بندی [۱۶، ۲۴] مورد استفاده قرار گرفته‌اند که می‌توان آن‌ها را به چند دسته تقسیم کرد: روش پارتیشن‌بندی (مانند K-Means)، روش مبتنی بر فاصله (مانند OCSVM)، یا روش مبتنی بر چگالی (مانند DBSCAN و LOF). الگوریتم‌های پیشرفته و بدون نیاز به برچسب، مانند LOF [۳۹] و DBSCAN [۳۲]، وجود دارند، اما در حال حاضر تعداد کمی از این روش‌ها در سناریوهای کاربردیِ کشف تقلب بیمه درمان، به‌صورت پایدار و کارآمد عمل می‌کنند و با چندین محدودیت روبرو هستند. به عنوان مثال، روش‌های پارتیشن‌بندی حساس به عرض خوشه هستند که نیازمند تکرار آزمایش‌ها برای انتخاب عرض بهینه است [۲۴]. تکرار آزمایش‌ها در مجموعه داده‌های با حجم بالا بسیار زمان‌بر می‌شود. علاوه بر این، برای سایر مدل‌های خوشه‌بندی که براساس

1. Inlying
2. Outlying

فاصله یا معیارهای چگالی هستند [۱۱]، هزینه محاسباتی اندازه‌گیری فاصله بسیار بالا است [۲۲]. برای مثال، معمولا OCSVM، حساسیت زیادی به پارامترهای هسته و تنظیمات دارد و در داده‌های بسیار بزرگ و با ابعاد بالا ممکن است زمان‌بر و ناپایدار باشد. همچنین، LOF، به رغم قابلیت تشخیص نمونه‌های نادر، نیازمند محاسبات همسایگی پیچیده و زمان‌بر است، و در مواجهه با حجم داده‌های زیاد، کارایی خود را از دست می‌دهد. از سوی دیگر، یکی از روش‌های پیشرفته برای کشف تقلب مشارکتی، روشهای مبتنی بر گراف و تحلیل شبکه است که الگوریتم‌هایی مانند HAN[1] یا GNN[2] از این دست هستند. این روش می‌تواند روابط پنهان و الگوهای غیرعادی بین افراد، پزشکان، مراکز درمانی و شرکت‌های بیمه را آشکار کند. بااین حال، تحقیقات نشان می‌دهد روشهای مبتنی بر گراف به‌تنهایی کافی نیست و باید با روش‌های دیگر یادگیری ماشین ترکیب شود تا بتوان به نتیجه جامع‌تر و مطمئن‌تری رسید. از سوی دیگر، بسیاری از مدل‌های مبتنی بر گراف، مدل‌های جعبه سیاه هستند و حل یک مشکل خاص در حوزه مثل کشف تقلب بیمه سلامت، اغلب نیازمند میزان معینی از قابلیت تفسیرپذیری است [۱۷، ۲۲، ۵۰].

بنابراین، نیاز مبرم به توسعه الگوریتم‌های کشف ناهنجاری حس می‌شود که بتوانند با داده‌های حجیم و با ابعاد بالا، بدون برچسب و در محیط‌های پویا و پیچیده و در حال تکامل تقلب بیمه درمان، قابلیت استقرار بهتری در سیستم‌های کشف تقلب دنیای واقعی داشته باشند. برای این منظور، لیو[3] و همکاران (۲۰۰۸) رویکرد کشف ناهنجاری به نام IF را پیشنهاد دادند. این الگوریتم، در مطالعاتی برای کشف تقلب بیمه درمان مورد استفاده قرار گرفته است و اثربخشی آن از طریق داده‌های واقعی اثبات شده است [۲۷، ۲۸]. علاوه براین، عملکرد این الگوریتم‌ها در برخی تحقیقات [۱۴، ۲۷، ۳۹] با الگوریتم‌های دیگر مقایسه شده است که نتایج نشان می‌دهد الگوریتم IF در مقایسه با الگوریتم‌های OClSVM و LOF با دقت بیشتری قادر به کشف تقلب است. بااین حال، در حوزه‌های مختلف مانند کشف تقلب بیمه درمان، ویژگی‌های ابعادی بالا و حجم عظیم داده‌های بیمه درمان محدودیت‌هایی در کارایی سیستم‌های کشف ناهنجاری ایجاد می‌کند. برای حل این مسئله، برخی محققین تغییراتی در این الگوریتم داده‌اند. دینگ و فی[4] (۲۰۱۳) استفاده از قاب پنجره‌ای لغزان و الگوریتم تشخیص ناهنجاری داده‌های زنده انطباق‌پذیر به نام iForestASD بر پایه الگوریتم IF را پیشنهاد دادند که قادر است به طور مؤثر ناهنجاری‌ها را شناسایی کند. این الگوریتم برای شناسایی داده‌های ترافیک شبکه در برنامه‌هایی مانند شبکه‌های کامپیوتری و شبکه‌های حسگر طراحی شده است. پوجینی و مک‌لئون[5] (۲۰۱۸)، با ارائه روش‌های کاهش ابعاد و انتخاب متغیر، کشف ناهنجاری مبتنی بر IF را برای حل این مشکل پیشنهاد دادند. لاسکار[6] و همکاران (۲۰۲۱) روشی نوین مبتنی بر ترکیب الگوریتم K-Means و IF برای حل مسئله تعیین نرخ آلودگی در شناسایی ناهنجاری در کلان داده ترافیک شبکه‌ای ارائه دادند. در این روش، ابتدا بکمک IF نمره ناهنجاری محاسبه و سپس از این نمره به عنوان ورودی الگوریتم K-means برای تعیین ناهنجاری‌ها استفاده می‌شود. نتایج نشان می‌داد که سیستم پیشنهادی در شناسایی ناهنجاری‌های کلان داده‌ها مؤثر عمل می‌کند.

در پژوهش حاضر، الگوریتم IF که برای داده‌های حجیم بهینه شده است، به عنوان پایه انتخاب شده است. این الگوریتم قادر است در داده‌های با حجم زیاد، دقت نسبتاً پایداری در تشخیص ناهنجاری داشته باشد و مزایای آن شامل زمان آموزش کوتاه و سرعت بالا در شناسایی است، بنابراین در بسیاری از سناریوهای کشف ناهنجاری با حجم زیاد داده، مناسب است. در مقایسه با مدل‌های مبتنی بر خوشه‌بندی و همچنین مدل‌های مبتنی بر یادگیری عمیق که برای کشف ناهنجاری استفاده می‌شوند، الگوریتم IF چندین مزیت دارد. در مقایسه با مدل‌های خوشه‌بندی مبتنی بر فاصله و چگالی، IF از هیچ معیار فاصله یا چگالی برای شناسایی ناهنجاری در داده‌های با ابعاد بالا استفاده نمی‌کند. بنابراین هزینه محاسباتی را به طور قابل‌توجهی نسبت به روش‌های مبتنی بر فاصله و چگالی حذف می‌کند. علاوه بر این، IF دارای پیچیدگی زمانی خطی، با ضریب ثابت کم و نیاز کم حافظه است [۲۲]. مهم‌تر اینکه، لیو و همکاران (۲۰۰۸) نشان دادند که الگوریتم IF قابلیت

1. Hierarchical Attention Network
2. Graph Neural Network
3. Liou
4. Ding & Fei
5. Puggini & McLoone
6. Laskar

مقیاس‌پذیری برای حل مسائل با ابعاد بالا و در داده‌های با حجم زیاد را دارد. این ویژگی، IF را گزینه مناسبی برای کشف ناهنجاری در کشف تقلب بیمه درمان می‌سازد.

در خصوص IF دو مسئله اساسی وجود دارد. IF با محاسبه امتیار ناهنجاری نمونه داده‌ها تعیین می‌کند که یک نمونه ناهنجار است یا خیر. با این حال، نرخ آلودگی در داده‌های آموزش تا حد زیادی بر محاسبه این امتیاز تأثیر می‌گذارد، بنابراین IF به شدت به تنظیم این هایپرپارامتر وابسته است. در حالی که ارائه این اطلاعات ممکن است نیازمند ساخت مجموعه داده آموزشی برچسب‌گذاری شده به صورت دستی باشد که فرآیندی بسیار زمان‌بر در داده‌های بیمه درمان است [۲۲]. در محیط‌های واقعی، تنها می‌توان براساس تجربه دستی این نسبت را تنظیم کرد [۴۶]، که این به معنای عدم تضمین اینکه IF بهترین عملکرد را ارائه دهد، است. تنظیمات نادرست پارامترها نیز می‌تواند منجر به کمبود دقت و نرخ بالای هشدارهای کاذب در IF شود [۱۳]. در روش IF-KMeans پیشنهادی ارائه شده توسط لاسکار و همکاران (۲۰۲۱) نیاز به تعیین نرخ آلودگی برای پیش‌بینی نیست، زیرا نمره‌های ناهنجاری بدست آمده از روش IF را با استفاده از الگوریتم K-Means به برچسب‌های پیش‌بینی شده تبدیل می‌کند. علاوه براین، در IF باید حد آستانه‌ای برای تبدیل امتیازهای ناهنجاری پیش‌بینی‌شده توسط مدل به برچسب‌های مختلف پیدا کرد [۱۳، ۲۲]، که فرآیندی وقت‌گیر است و در مجموعه‌های داده بزرگ، هزینه‌ی بیشتری دارد. به همین منظور، برای حل این محدودیت‌ها، در پژوهش جاری روشی نوین برای کشف ناهنجاری (تقلب) در داده‌های بیمه درمان ارائه می‌دهیم که شامل ترکیب الگوریتم K-Meansو الگوریتم IF است. مدل پیشنهادی را K-IF نامگذاری می‌کنیم و نشان می‌دهیم که این مدل که از پارامترهای کمتری نسبت به الگوریتم IF استفاده می‌کند، نه تنها در کشف تقلب بیمه درمان با حجم زیاد داده قابل استفاده است، بلکه در کشف ناهنجاری‌ها نیز مؤثرتر از مدل اصلی IF است. در الگوریتم پیشنهادی، پس از پیش‌پردازش مجموعه داده، ویژگی‌های استخراج شده در مجموعه داده به عنوان ورودی به IF برای آموزش داده می‌شوند. برای حل مسئله تعیین پارامتر نرخ آلودگی، ابتدا در الگوریتم IF این نسبت بالاتر از نسبت واقعی در نظر گرفته می‌شود، که باعث می‌شود تعداد بیشتری از نمونه داده‌ها در الگوریتم IF تحت عنوان نمونه‌های مشکوک به تقلب شناسایی شوند. سپس در مرحله دوم به کمک الگوریتم K-means برچسب نمونه‌های خوشه مشکوک به تقلب پیش‌بینی می‌شود.

۳. روش‌شناسی پژوهش

۱.۳. چارچوب هوشمند کشف تقلب بیمه درمان

در حالی که متقلبان روش‌های خود را در انجام تراکنش‌های کاذب به مرور زمان تکامل می‌دهند، نیاز به توسعه یک چارچوب تشخیص تقلب است که بتوان همواره آن را تکامل بخشید. برخی محقق طی یک دهه گذشته برحسب این نیاز اقدام به توسعه چنین چارچوب‌هایی کرده‌اند [۳۰، ۳۹]. تصویر ارائه‌شده در شکل ۱ چارچوبی پیشنهادی فرآیند سیستماتیک کشف تقلب در داده‌های بیمه درمان را نشان می‌دهد. این چارچوب، یک مرجع منسجم و چندمرحله‌ای برای پیاده‌سازی سیستم هوشمند کشف تقلب است. چارچوب توسعه یافته در این پژوهش برای کشف تقلب بیمه درمان یک رویکرد نوآورانه است که سعی در غلبه بر کمبودهای تحقیقات موجود دارد. ابتدا، چارچوب توسعه یافته مستقل از بازیگران و خدمات/کالاهای ارائه شده است. به عبارت دیگر، موارد تقلبی که بررسی می‌شوند به رفتار یک نوع خاص از بازیگر یا مجموعه‌ای از کالاها/خدمات محدود نمی‌شوند. دوم اینکه، نحوه‌ی انجام رفتارهای تقلب و سوءاستفاده از نظر انواع بازیگران و کالاها/خدمات در طول زمان تغییر می‌کند، به این معنی که اکوسیستم تقلب پویاست. از این رو، یک چارچوب ماژولار و انعطاف پذیر مبتنی بر رویکرد یادگیری ماشین تعاملی اتخاذ می‌شود که قادر به مدیریت چنین تغییراتی است. روش تعاملی به کارشناسان و/یا کاربران امکان می‌دهد تا با تعریف انواع تقلب و ویژگی‌های جدید به صورت دوره‌ای با فرآیند آموزش در الگوریتم‌های یادگیری ماشین تعامل کنند. سوم اینکه، ما از ترکیب الگوریتم‌های شناخته شده‌ای در چارچوب خود جهت بهبود دقت کشف تقلب بیمه درمان استفاده کرده‌ایم، مانند الگوریتم IF برای شناسایی مقدماتی نسخه‌های تقلب و سپس الگوریتم K-Means برای شناسایی نهایی نسخه‌های تقلب، نمره‌های z ویژگی‌های استخراج شده از تراکنش‌های تاریخی برای بی‌مقیاس‌سازی ویژگی‌ها، انبار داده دو مرحله‌ای برای بهبود عملکرد سیستم به منظور دستیابی به تحلیل پیشگیرانه، و فناوری مبتنی بر داشبورد به صورت یک

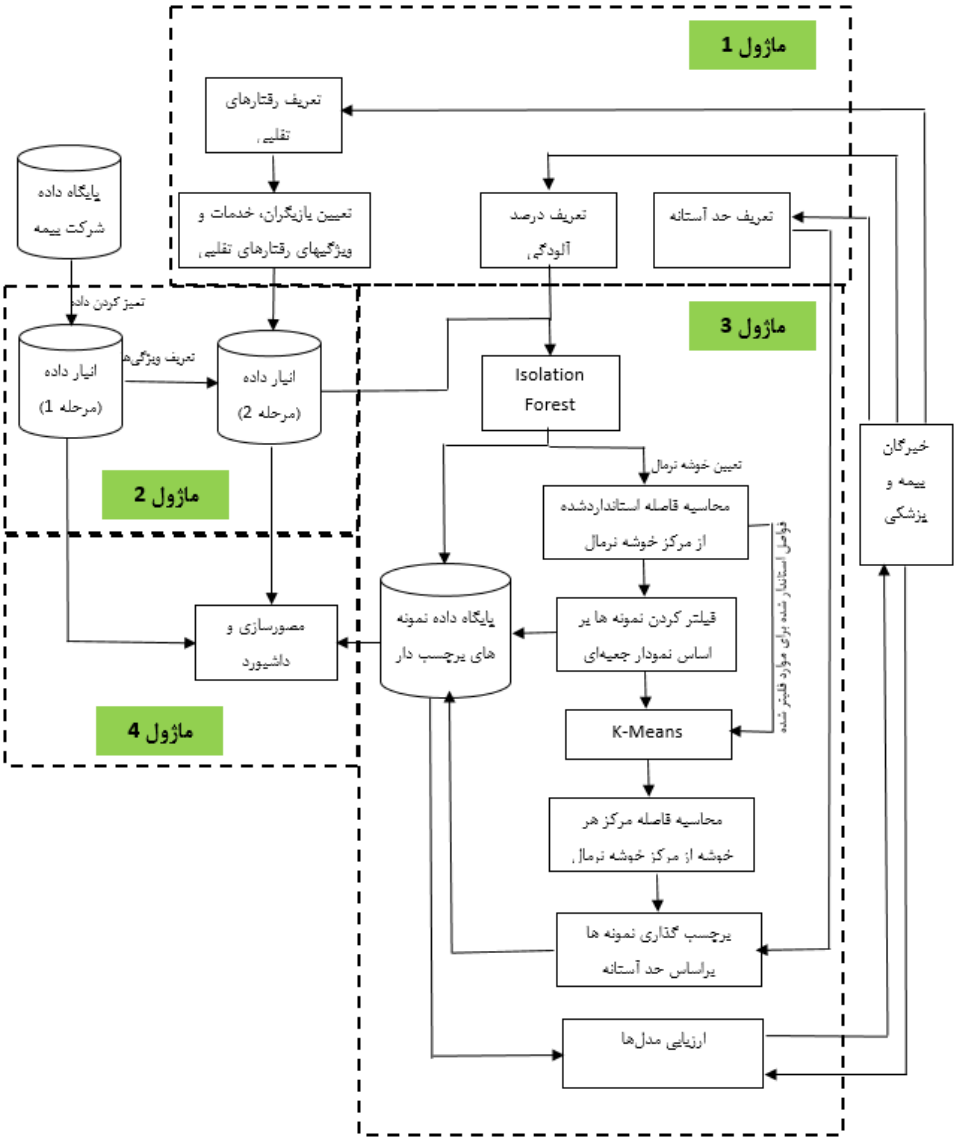
بسته نرم‌افزاری برای تجسم اطلاعات و کمک به مدل‌سازان جهت بازخور به سیستم استفاده می‌کنیم. چهارم اینکه، این معماری قادر است ادعاهای مجزا را در قالب یک سیستم کل‌نگر تحلیل کند. سامانه‌ای که با بهره‌گیری از داده‌های قبلی تراکنش‌ها و تاریخچه فعالیت‌های بازیگران طراحی شده است، توانایی تشخیص چنین موارد پیچیده و ظریف را نیز دارد. حتی در مواردی که دو تراکنش از نظر محتوای کاملاً یکسان هستند، بسته به سابقه تراکنش‌ها و نقش بازیگران در هر مورد، ممکن است یکی به عنوان تقلب شناسایی شده و دیگری نباشد. نهایتاً، چارچوب توسعه یافته یک رویکرد مبتنی بر نمره ناسازگاری در تعریف ویژگی‌ها است؛ اما بر خلاف سایر رویکردهای مبتنی بر نمره ناسازگاری که نمره ناسازگاری را به بازیگران یا کالاها به‌طور جداگانه اختصاص می‌دهند، نمره‌های ناسازگاری پیشنهادی برای ارتباط بین بازیگر و کالا به‌صورت کلی ارزیابی می‌شوند. بنابراین، هرچند یک بازیگر ممکن است از نظر کالاهای خاص دارای ناسازگاری باشد، فقط همان تراکنش‌هایی که شامل کالاها و بازیگران خاص هستند، به‌عنوان تقلبی یا غیرتقلبی برچسب‌گذاری می‌شوند. تعاملات کارشناسان با این چارچوب در طول مرحله آموزش و طبقه‌بندی در ادامه توضیح داده خواهد شد.

همانطوری که اشاره شد، چارچوب پیشنهادی به‌گونه‌ای طراحی شده است که ویژگی‌های «پویایی اکوسیستم تقلب»، «تعاملی بودن تشخیص تقلب» و «قابل گسترش بودن» را به‌صورت هم‌راستا با نیازمندی‌های دنیای واقعی پشتیبانی می‌کند. در این ساختار، با بهره‌گیری از معماری ماژولار و انعطاف‌پذیر، توسعه‌دهندگان و کارشناسان به‌راحتی قادر هستند تغییرات و تحولات در رفتارهای تقلبی و الگوهای فعالیت بازیگران مختلف را در سیستم وارد کنند و در فرآیند کشف تقلب اعمال نمایند. استفاده از ساختار ماژولار در برخی از مطالعات [۱۹، ۳۵، ۳۸] جهت انعطاف‌پذیری چارچوب کشف تقلب پیشنهاد شده است. درواقع، این سیستم قابلیت آن را دارد که با افزودن مدل‌های جدید و ویژگی‌های مرتبط، انواع تقلب‌ها را در زمان واقعی شناسایی و واکنش سریع نشان دهد. با بروزرسانی‌های دوره‌ای و تعامل مستمر با کارشناسان، ویژگی‌ها و الگوریتم‌ها به‌روزرسانی می‌شوند تا به بهترین شکل به تغییرات جدید پاسخ دهند و از تکرار خطاهای گذشته جلوگیری شود، بنابراین می‌توان با پویایی اکوسیستم تقلب به سازگاری رسید. در این چارچوب، نقش کارشناسان بسیار حیاتی است؛ آن‌ها می‌توانند انواع جدید تقلب، نواحی حساس و ویژگی‌های مهم را تعریف و مشخص کنند. این تعامل دوسویه میان سیستم و کارشناسان، باعث می‌شود مدل‌های تشخیص تقلب به‌روز و دقیق نگه داشته شوند و سیستم بتواند به‌شکل فعال و دینامیک به تغییرات واکنش نشان دهد. برخی چارچوب‌های ارائه شده برای کشف تقلب بیمه درمان و حذف خطاهای پرداخت مبتنی بر تعامل کارشناسان با الگوریتم به شکل یادگیری ماشین تعاملی توسعه یافته است [۲۰]. برای مثال، سان[1] و همکاران (۲۰۲۰) از انتخاب ویژگی تعاملی برای کشف تقلب و کوزه[2] و همکاران (۲۰۱۵) از یک یادگیری ماشین تعاملی برای تعریف انواع تقلب و ویژگی‌ها در بیمه درمان استفاده می‌کنند [۱۹، ۴۵].

علاوه‌بر این، از جهت قابل گسترش بودن، طراحی ماژولار این چارچوب به‌طور قابل توجهی فرآیند افزودن فناوری‌ها، الگوریتم‌ها و مدل‌های تحلیلی نوین را تسهیل می‌بخشد. مثلا، در آینده می‌توان الگوریتم‌های پیشرفته‌تری مانند یادگیری عمیق یا تحلیل شبکه‌های عصبی را به سیستم افزود و بدون نیاز به تغییر ساختار کلی، آن‌ها را بهره‌برداری کرد. این قابلیت گسترش، تضمین می‌کند که سیستم همواره آماده پاسخگویی به تحولاتی مانند شیوه‌های جدید تقلب و نیازهای متفاوت صنعت بیمه باشد و تضمین می‌کند که سیستم در طول زمان همچنان قابلیت ارتقاء و توسعه داشته باشد. در نهایت، ساختار این چارچوب به‌گونه‌ای است که می‌تواند تراکنش‌های پیچیده و ظریف را در سطح وسیع تحلیل کند. حتی در مواردی که دو ادعا مشابهت ظاهری دارند، براساس سابقه تراکنش‌ها و نقش بازیگران مختلف، قضاوت می‌کند که کدام‌یک تقلب است و کدام نیست.

قبل از پرداختن به بحث درباره جزییات فرآیند، در این قسمت ماژول‌های چارچوب توسعه یافته ارائه می‌شود. شکل ۲ چارچوب توسعه یافته را نشان می‌دهد که شامل چهار ماژول اصلی است. در ادامه به ماژول‌های مختلف این چارچوب پرداخته می‌شود.

1. Sun
2. Kose



شکل ۲. ساختار و ماژول‌های سیستم هوشمند کشف تقلب بیمه درمان

۱.۱.۳. فریم ورک تقلب و بخش دانشی چارچوب کشف تقلب (ماژول ۱)

چارچوب توسعه‌یافته در این پژوهش، رویکردی مبتنی بر مدل‌سازی فعالیت‌های کارشناسانه و با هدف ارزیابی ریسک ادعاهای بیمه درمان طراحی شده است. در این رویکرد، یک فریم‌ورک جهت شبیه‌سازی فرآیند تشخیص تقلب ساخته می‌شود که تیمی از متخصصان حوزه پزشکی و بیمه را قادر می‌سازد تا نحوه بروز رفتارهای غیرعادی شناخته‌شده را توصیف و تحلیل کنند. این فریم‌ورک، بازیگران، خدمات/کالاهای مرتبط و همچنین روابط بین این عوامل را مشخص می‌سازد، به گونه‌ای که ارتباط میان بازیگران و خدمات/کالاها، در اثر رفتارهای خاص و غیرعادی، مقادیر متفاوت و هم‌راستا با فعالیت‌های مشکوک را تولید می‌کند. این مقادیر بکمک تعریف ویژگی‌های هدفمند جهت کشف تقلب بیمه درمان در الگوریتم‌های یادگیری ماشین محقق می‌شود. ویژگی‌ها[1] معیارهایی هستند که نشان‌دهنده رابطه بین بازیگران و خدمات/کالاها هستند.

1. Attributes

براساس این چارچوب، اطلاعات مربوط به بازیگران، خدمات/کالاها و ویژگی‌های مربوطه برای هر نوع تقلب استخراج می‌شود. برای تحلیل رفتارهای غیرعادی، از سه نوع ویژگی اصلی بهره گرفته می‌شود. ویژگی‌های دوره‌ای[1] نشان‌دهنده مقادیر ماهانه است، و برای شناسایی نوسانات کوتاه‌مدت و الگوهای تغییر میان‌مدت کاربرد دارد. ویژگی‌های تجمعی[2] مقادیر سالانه را نمایش می‌دهند و برای ارزیابی روند کلی فعالیت‌ها در بازه‌های زمانی بلندمدت مفید هستند. ویژگی‌های تفاضلی[3] نرخ تغییر نسبت به ماه قبل را نشان می‌دهند و امکان تحلیل نوسانات سریع و احتمالا مشکوک در رفتارها را فراهم می‌آورند. این رویکرد، امکان تحلیل چندبعدی و جامع رفتارهای مشکوک در داده‌های ادعای بیمه درمان را فراهم می‌کند، و به تیم‌های کارشناسی کمک می‌کند تا با دقت و اطمینان بیشتری، نقاط ضعف و رفتارهای احتمالی تقلب را شناسایی و ارزیابی کنند. برای استخراج ویژگی‌های مؤثر در کشف تقلب‌های بیمه درمان، ابتدا چندین جلسه با یک تیم حرفه‌ای متشکل از کارشناسان، متخصصین و پزشکان بیمه درمان برگزار شد. در این جلسات، از اعضای تیم خواسته می‌شود مهم‌ترین فریم ورک‌های تقلب بیمه‌ای که بیشترین تأثیر را در فرآیندهای پرداخت و ارزیابی دارند، تعیین و اولویت‌بندی کنند. پس از مشخص شدن انواع تقلب، از اعضای تیم درخواست می‌شود تا برای هر نوع تقلب، ویژگی‌های کلیدی و قابل استخراج از داده‌های انبار مرحله یک برای کشف آن‌ها شناسایی و تعریف کنند. در این فرآیند، راهنمایی لازم در قالب پیشنهادهای ویژه و پرسش‌های هدفمند از سوی اعضای این تیم و مدلساز ارائه می‌شود تا ویژگی‌هایی که می‌توان از روی داده‌ها استخراج کرد، شناسایی و تعریف شوند. این ویژگی‌ها برای تسهیل محاسبات ناسازگاری نسخه‌های پزشکی برای شناسایی تقلب‌های بیمه‌ای، آماده می‌شوند.

۲.۱.۳. انبار داده دو مرحله‌ای (ماژول ۲)

مجموعه داده‌های تهیه شده مربوط به بخشی از پایگاه داده شرکت بیمه دی است که حاوی اطلاعات مربوط به نسخه‌های پزشکی، بیماران، خدمات درمانی و مشخصات ارائه‌دهندگان مراقبت‌های درمانی (مانند پزشک، مرکز درمانی، داروخانه و غیره) می‌باشد. هر رکورد در داده‌های تهیه شده نمایانگر یک نسخه پزشکی و یک ادعای خسارت است، و در مجموع شامل ۷۹۳،۵۸۹ رکورد است که بازه زمانی بین اسفند ۱٤۰۲ تا اسفند ۱۴۰۳ را در بر می‌گیرد. در این مجموعه داده، برخی از این متغیرها عبارتند از: کد شناسایی بیمار (ID)، تاریخ تولد، جنسیت، تاریخ پذیرش نسخه، تاریخ حواله، نوع پوشش بیمه، نوع خدمات دریافت شده (مانند ویزیت‌ها، عمل جراحی، خدمات اورژانس، دارو و غیره)، نام و نوع مرجع درمان، نام و تخصص پزشک یا ارائه‌دهنده خدمات، استان محل ارائه خدمات، کل مبلغ خسارت، و مبلغ تایید شده خسارت.

یکی از عوامل محدودکننده در توسعه راه‌حل‌های پیشگیرانه در حوزه کشف تقلب‌های بیمه‌ای، نیاز به توان محاسباتی بالا برای پردازش حجم عظیم داده‌ها است. به منظور غلبه بر این محدودیت، سیستم پیشنهادی از یک انبار داده دو مرحله‌ای بهره می‌برد. این طراحی تضمین می‌کند که عملیات محاسباتی در سطح بالایی از کارایی انجام شود و امکان بهره‌برداری موثر از رویکردهای پیشگیرانه و هدایت‌شده در کشف تقلب‌ها فراهم آید. ماژول دوم این سیستم، که در شکل ۲ نشان داده شده است، نمایانگر ساختار انبار داده دو مرحله‌ای است که برای این منظور درنظر گرفته شده است. در مرحله اول، یک انبار داده مبتنی بر معماری ستاره‌ای[4] طراحی شده که داده‌های عملیاتی شرکت بیمه را از طریق فرآیند استخراج، تبدیل و بارگذاری[5] به این سیستم وارد می‌کند. این فرآیند طی الگوریتم‌های خاص پاک‌سازی داده، منجر به تکمیل داده‌های اولیه[6]، و حذف ناسازگاری‌ها و خطاهای موجود در آن‌ها می‌شود. در فرآیند پاک‌سازی، داده‌های ناسازگار، ناقص، و حاوی خطاها شناسایی و اصلاح یا حذف شدند تا اطمینان حاصل شود که داده‌ها با کیفیت و مطابق با استانداردهای لازم برای تحلیل‌های پیشرفته قرار دارند. این عملیات شامل اصلاح نواقص در داده‌ها، حذف رکوردهای تکراری و رفع ناسازگاری‌های موجود در مجموعه داده‌ها بود. به عنوان بخشی از این روند، اقداماتی چون پر کردن فیلد تخصص پزشک براساس رکوردهایی که تخصص آن‌ها ذکر شده، تکمیل فیلد نام و نوع ارائه‌دهنده خدمات درمانی (پزشک، مرکز درمانی، داروخانه و غیره) براساس کد مرجع و نام مرجع، و همچنین پر کردن تخصص ارائه‌دهنده (پزشک و سایر ارائه‌دهندگان)

1. Temporal
2. Cumulative
3. Differential
4. Star-Schema
5. Extract, Transform, and Load (ETL)
6. Data Imputation or Data augmentation

براساس فیلدهای مربوط به خدمات ارائه شده و رکوردهای مرتبط به ارائه‌دهنده انجام شد. انبار داده مرحله دوم، شامل ویژگی‌های استخراج شده از روی داده‌های انبار مرحله اول است که از نظر سه حالت تحلیل (دوره‌ای، تجمعی و تفاضلی) در مرحله قبل طراحی شده‌اند. به‌عنوان مثال، ویژگی‌هایی مانند «نسبت مراجعات یک بیمار به یک پزشک خاص» یا «نسبت ارائه یک خدمت خاص توسط پزشک به کل نسخه‌های او» می‌توانند نشانه‌هایی از الگوهای تکرارشونده یا غیرعادی باشند که در رفتارهای تقلب‌آمیز رایج‌اند. همچنین، مقایسه هزینه‌های دریافتی پزشک برای یک خدمت خاص با میانگین سایر پزشکان در همان خدمت یا تخصص، امکان تشخیص فعالیت‌های تقلب‌آمیز را فراهم می‌کند. در مجموع، ویژگی‌های مهندسی‌شده با هدف افزایش حساسیت مدل نسبت به الگوهای پنهان و پیچیده طراحی شده‌اند و نقش مهمی در بهبود دقت و کاهش نرخ مثبت کاذب در فرآیند کشف تقلب ایفا می‌کنند. از آنجا که نمرات ویژگی‌ها که نشان‌دهنده درجه انحراف رابطه بین بازیگران و خدمات/کالاها است، به جای استفاده از مقادیر اسمی، از نمرات استاندارد شده هنگام تعیین آن‌ها بهره گرفته می‌شود.

۳.۱.۳. موتور کشف تقلب هوشمند (ماژول ۳)

در این قسمت پس از معرفی الگوریتم درخت ایزوله و سپس الگوریتم IF، الگوریتم پیشنهادی توسعه داده می‌شود.

درخت ایزوله. براساس نظرات لیو[1] و همکاران (۲۰۰۸)، الگوریتم IF هر نقطه را به عنوان ناهنجار طبقه‌بندی می‌کند اگر طول مسیر کوتاهی در درخت ایزوله[2] داشته باشد. الگوریتم درخت ایزوله با انتخاب تصادفی یک ویژگی از مجموعه داده و مقداری در دامنه آن ویژگی کار می‌کند. سپس، با استفاده از آن مقدار و ویژگی، داده‌ها را به دو زیرمجموعه تقسیم می‌کند. این عمل، منجر به ایجاد دو گره فرزند برای گره کنونی می‌شود؛ که هر یک برای هر زیرمجموعه داده است. زیرمجموعه‌ای که کوچکتر از مقدار تقسیم است، به سمت چپ و زیرمجموعه بزرگ‌تر از آن، به سمت راست هدایت می‌شود. این فرآیند به صورت بازگشتی بر روی زیرمجموعه‌های داده‌ها تکرار می‌شود، تا زمانی که یا تمام نقاط داده در یک زیرمجموعه به یک کلاس خاص تعلق داشته باشند، یا عمق نهایی درخت حداکثر شود. پس از ساخت درخت‌های ایزوله، الگوریتم IF امتیاز ناهنجاری هر نقطه داده را با محاسبه عمق متوسط درختی که آن نقطه در آن ایزوله شده است، محاسبه می‌کند. سپس، این الگوریتم، نقاط داده با مسیر کوتاه‌تر به عنوان ناهنجار و نقاط با مسیر بلندتر به عنوان نقاط عادی طبقه‌بندی می‌شوند. تعاریف و فرمول‌های زیر از لیو و همکاران (۲۰۰۸) است:

تعریف ۱: فرض کنید T یک گره در یک درخت ایزوله است. این گره یا یک گره خارجی است که فرزندی ندارد، یا یک گره داخلی است که یک آزمایش دارد و دقیقاً دو فرزند (T_l, T_r) دارد. آزمایش X شامل یک ویژگی q و یک مقدار تقسیم p است، به طوری که آزمون q < p داده‌ها را به T_l و T_r تقسیم می‌کند. یک درخت ایزولاسیون می‌تواند از مجموعه داده‌های X ساخته شود، جایی که $X = \{x_1, x_2, ..., x_n\}$، با پیروی از فرآیندی مشابه آنچه در قسمت قبل توضیح داده شد. در اینجا، n تعداد ویژگی‌ها است.

با پیروی از فرآیند مشابهی که در بالا توضیح داده شد، درخت ایزوله یک درخت دودویی است که هر گره در آن یا دو فرزند دارد یا هیچ فرزندی ندارد. مجموعه داده‌های تقسیم‌نشده، که نمی‌توان بیشتر آن‌ها را تقسیم‌بندی کرد، به عنوان گره خارجی شناخته می‌شود و تعداد این گره‌ها برابر با تعداد نقاط داده در مجموعه، یعنی n است. گره‌ای که با دو فرزند خاتمه می‌یابد، گره داخلی است. تعداد گره‌های داخلی در یک درخت ایزوله برابر است با تعداد عملیات‌های تقسیم انجام شده منهای یک ($n-1$)، زیرا هر گره داخلی نشان‌دهنده یک عملیات تقسیم است و برای ایزوله کردن تمامی نقاط داده در مجموعه، نیاز است که $n-1$ عملیات تقسیم انجام شود. بنابراین، تعداد کل گره‌ها در یک درخت ایزوله برابر است با مجموع گره‌های خارجی و داخلی، که مقدار آن $2n-1$ است. این ویژگی، میزان حافظه مورد نیاز را محدود می‌کند و تنها با افزایش تعداد نقاط داده، خطی افزایش می‌یابد.

تعریف ۲: طول مسیر $h(x)$ مربوط به یک نقطه داده x با تعداد لبه‌هایی اندازه‌گیری می‌شود که x باید از ریشه درخت ایزوله عبور کند تا در نهایت در یک گره خارجی متوقف شود. امتیاز ناهنجاری برای یک نقطه داده از طریق سه مراحل به دست می‌آید: (۱) محاسبه طول مسیر $h(x)$ برای نقطه داده x در تمامی درخت‌های ایزوله. (۲) محاسبه متوسط طول مسیر برای نقطه داده x در تمام درخت‌های ایزوله که در آن نقطه ایزوله

1. Liu
2. Isolation Tree

شده است. (۳) محاسبه امتیاز ناهنجاری به صورت میانگین تمامی مقادیر به دست آمده در مرحله دوم، تقسیم بر تعداد درخت‌های موجود در جنگل. امتیاز ناهنجاری به دست آمده در دامنه صفر تا ۱ قرار دارد و امتیاز نزدیک به ۱ نشان می‌دهد که نقطه داده احتمالاً یک ناهنجاری است، در حالی که امتیاز نزدیک به صفر نشان می‌دهد که نقطه داده احتمالا یک نقطه عادی است [۹]. می‌توان متوسط طول مسیر درخت ایزوله را ابتدا با محاسبه $c(n)$ جایکه $c(n)$ برابر با میانگین طول مسیر $h(x)$ است و n تعداد ویژگی‌های مجموعه داده است، به دست آورد. از آنجا که درخت ایزوله ساختاری مشابه درخت جستجوی دودویی[1] (BST) دارد، محاسبه متوسط طول مسیر برای توقف در یک گره خارجی مشابه جستجو در BST است. بنابراین، متوسط طول مسیر می‌تواند از رابطه زیر محاسبه شود:

$$c(n) = 2H_{(n-1)} - \left(\frac{2(n-1)}{n}\right) \qquad (۱)$$

که در آن H(i) عدد هارمونیک است و می‌توان آن را با ln(i) + 0.5772156649 (ثابت اویلر-ماشرونی) محاسبه کرد. c(n) بیانگر میانگین طول مسیر است و برای نرمال‌سازی h(x) استفاده می‌شود. جایی که:

$$c(n) = \begin{cases} 2H_{(n-1)} - \left(\frac{2(n-1)}{n}\right) & n > 2 \\ 1 & n = 2 \\ 0 & otherwise \end{cases} \qquad (۲)$$

فرمول محاسبه نمره ناهنجاری برای نقطه داده x به صورت زیر است:

$$s(x,n) = 2^{-\frac{E(h(x))}{c(n)}} \qquad (۳)$$

که در آن $E(h(x))$ میانگین طول مسیر مورد انتظار از مجموعه‌ای از درخت‌ها است. با توجه به رابطه (۳)، هنگامی که $E(h(x)) \to c(n)$ ، پس $s \to 0.5$ ، هنگامی که $E(h(x)) \to 0$ ، پس $s \to 1$ ، و هنگامی که $E(h(x)) \to n-1$ ، پس $s \to 0$. با استفاده از نمره ناهنجاری محاسبه‌شده، الگوریتم قادر است یکی از این تصمیم‌ها را اتخاذ کند:

- اگر نمره ناهنجاری نقطه داده بسیار نزدیک به ۱ باشد، این نقطه به عنوان ناهنجار طبقه‌بندی می‌شود.
- اگر نمره ناهنجاری نقطه داده کمتر یا مساوی ۰.۵ باشد، این نقطه به عنوان نقطه عادی طبقه‌بندی می‌شود.
- اگر همه نقاط داده نمره ناهنجاری یکسان با ۰.۵ داشته باشند، این نشان دهنده این است که در مجموعه داده هیچ ناهنجاری مشخصی وجود ندارد و مدل تمامی نقاط داده را به عنوان نقاط عادی طبقه‌بندی می‌کند.
- اگر طول مسیر مورد انتظار E(h(x)) برابر با متوسط طول مسیر c(n) باشد، پس نمره ناهنجاری s = 0.5 است، صرف‌نظر از مقدار n.

الگوریتم‌های IF. در تشخیص ناهنجاری‌ها با استفاده از الگوریتم IF، دو مرحله اصلی وجود دارد: مرحله آموزش و مرحله آزمون. در مرحله آموزش، الگوریتم IF یک مجموعه از درخت‌های ایزوله را به صورت بازگشتی و با تقسیم مجموعه داده‌های آموزش در طول زمان ساخته و توسعه می‌دهد، تا زمانی‌که هر نقطه داده در برگ خود ایزوله شود یا حد ارتفاع درخت به آن برسد. حد ارتفاع درخت، براساس اندازه زیر نمونه[2] $\varphi : l = ceiling(\log_2 \varphi)$ تعیین شده است که به صورت تقریبی تقریباً میانگین ارتفاع درخت است. الگوریتم ۱ و ۲ در شکل ۳ و ٤ جزئیات مربوط به مرحله آموزش را نشان می‌دهند.

1. Binary Search Tree (BST)
2. Sub-sampling size

Algorithm 1: IF (D, t, φ)	
Input: $\boldsymbol{D}$ = ($x1$, $x2$, . . ., xn) – dataset, t – number of trees, φ – sub-sampling	
Output: a set of t isolation trees	
Initialize Forest	
set height limit $l = ceiling(\log_2 \varphi)$	**1:**
for i = 1 to t do	**2:**
D′ ← sample (D, φ)	**3:**
Forest ← Forest ∪ iTree $(D', 0, 1)$	**4:**
end for	**5:**
return Forest	**6:**

شکل ۳. شبه‌کد الگوریتم *IF*

دو پارامتر ورودی برای الگوریتم IF وجود دارد. اندازه زیر نمونه φ که کنترل‌کننده اندازه داده‌های آموزشی است و معمولاً به صورت پیش‌فرض برابر با $\varphi = 2^8 = 256$ است. این یک مقدار انعطاف‌پذیر است که IF در تشخیص ناهنجاری‌ها عملکرد خوبی دارد، اما همچنین این مقدار برای تمامی مجموعه‌های داده ثابت نیست و می‌تواند بنابر نوع مجموعه داده تغییر کند. در IF، مقدار اندازه زیرنمونه، معمولاً کمتر از کل داده‌ها است و مشخص می‌کند که در هر مرحله چند نمونه به صورت تصادفی برای ساخت درخت استفاده می‌شود. مهم‌ترین نقش آن در کنترل حجم داده‌های مورد استفاده برای ساخت هر درخت است، که بر سرعت و دقت الگوریتم تاثیر می‌گذارد. پارامتر دوم، تعداد درخت‌های ساخته شده t است که کنترل کننده اندازه مجموعه است و معمولاً مقدار آن ۱۰۰ تعیین می‌شود، چون طول مسیرها معمولاً زودتر از این مقدار همگرایی خوبی دارند [۳۷]. بعد از ساختن درخت‌ها، مدل برای مرحله ارزیابی آماده می‌شود.

Algorithm ۲: iTree (D, h, l)	
Input: D = ($x1$, $x2$, . . ., xn) – dataset, e – current tree height, l –height limit	
Output: an iTree t	
Initialize: t = ∅ (an empty tree)	
if e ≥ l or |X| ≤ 1 then	**1:**
return t	**2:**
Else	**3:**
let set Q to be a list of features in D	**4:**
randomly select a feature q_i (q_i ∈ Q)	**5:**
randomly select a split point p ∈ (min(q_i), max(q_i))	**6:**
Define D_l and D_r contain the sample of D, where q_i is smaller and larger than p	**7:**
D_l ← filter (D, q_i < p)	**8:**
D_r ← filter (D, q_i >= p)	**9:**
repeate iTree (D_l, h+ 1, l) and link the obtained tree as the left tree of t	**10:**
repeate iTree (D_r, h+ 1, l) and link the obtained tree as the right tree of t	**11:**
end if	**12:**

شکل ٤. شبه‌کد الگوریتم *IF*

الگوریتم پیشنهادی K-IF. شرایط واقعی نسخه‌های بیمه بسیار پیچیده است و باعث می‌شود تعاریف و توزیع داده‌ها بسیار سریع تغییر کنند، بنابراین لازم است هر زمان داده‌های بدون برچسب جدید جمع‌آوری شد، از آن‌ها به عنوان مجموعه آموزش برای به‌روزرسانی مدل استفاده کنیم.

در این حالت، تنها می‌توانیم با تکیه بر تجربه عملی، مقدار نرخ آلودگی(نسبت تقلب) را تعیین کنیم، اما نرخ آلودگی نادرست می‌تواند عملکرد IF را بی‌ثبات کند و منجر به دقت پایین و نرخ خطای بالا در تشخیص تقلب شود. اگرچه الگوریتم IF تا حدی مناسب‌تر از سایر الگوریتم‌ها برای داده‌های بزرگ بدون برچسب است، اما مانند سایر الگوریتم‌های بدون نظارت، عملکرد آن به شدت به تنظیمات نسبت ناهنجاری وابسته است. این روش باید آستانه‌ای پیدا کند تا نمره‌های ناهنجاری پیش‌بینی‌شده توسط مدل را به برچسب‌های مختلف تبدیل کند که این فرآیندی پرهزینه است، و همچنین نیازمند اطلاعاتی درباره نرخ آلودگی در داده‌های آموزش است که در سناریوهای داده‌های بزرگ ارائه آن مشکل‌ساز است. در این مطالعه، الگوریتم خوشه‌بندی K-means که مراحل آن در شکل ۵ آمده است، برای بهبود عملکرد IF در زمانی که نسبت ناهنجاری واقعی ناشناخته است، به کار گرفته شده است. الگوریتم K-means برخلاف الگوریتم‌هایی که تکیه بر دانش قبلی در خصوص ابرپارامترهایی مانند نسبت ناهنجاری واقعی دارند و در عمل پایدار نیستند، نیاز به دانستن چنین اطلاعاتی ندارد.

	Algorithm 3: K-means Algorithm
	Input: $F = (x_1, x_2, \ldots, x_m)$ - a dataset containing m instances
	Output: A set of anomalies and normal clusters
1:	**For** a range of K values, do:
2:	Calculate the Silhouette score;
3:	**End for**
4:	Plot out a K value vs. silhouette graph;
5:	Find out the optimal value of K, which has the highest score on the graph;
6:	Randomly choose K instances from F as the initial cluster centers;
7:	**Repeate**
8:	Reassign each instance to the cluster based on the mean of the instances in the cluster;
9:	Update the cluster means;
10:	**Until** no change occurs in the cluster means;
11:	**For** each cluster, **do**:
12:	Calculate the cluster center of each cluster to get the result $C = \{C_1, C_2, \ldots, C_K\}$
13:	Calculate the standard Euclidean distance d_k between C and C
14:	Dlist append d_k
15:	**end for**
16:	Specify a threshold value (tsv) determined by domain experts.
17:	**For** each cluster of F, do:
18:	If $d_k > tsv$, then:
19:	Label the cluster as abnormal;
20:	**Else**
21:	Label as normal;
22:	**End for**

شکل ۵. شبه‌کد الگوریتم *K-means*

در مقابل، الگوریتم پیشنهادی پژوهش حاضر که در شکل ٦ آمده است، تنها نیازمند دانش پایه‌ای درباره سناریو در یک کاربرد دنیای واقعی است تا بتواند عملکرد خوبی داشته باشد. برای مثال، در تشخیص ناهنجاری بیمه درمان، نرخ آلودگی معمول در این حوزه حدود ۱۰ تا ۱۵ درصد است، و کافی است پارامتر نرخ آلودگی را بالاتر از این مقدار تنظیم کنیم تا نتایج تشخیص خوبی حاصل شود.

	Algorithm 4 : K-IF Algorithm
	Input: $D = (x_1, x_2, \ldots, x_n)$ - dataset
	Output: the index list of normal values *Nlist*, the index list of abnormal values *Alist*
	Initialize: IF, *Nlist* = ∅, *Alist* = ∅, distance list *Flist* = ∅
1:	set the contamination parameter of IF to 2 times the background knowledge of the relevant field
2:	get anomaly score $A = (a_1, a_2, \ldots, a_n)$ and preliminary classification result $L = (l_1, l_2, \ldots, l_n)$ of D with IF
3:	**for** each label l_i in L **do**
4:	**if** $l_i = 1$ **then**
5:	*Nlist* append x_i
6:	**Else**
7:	*Alist* append x_i
8:	**end if**
9:	**End for**
10:	Calculate the cluster center C of *Nlist*
11:	**for** each sample s_i in *Alist* **do**
12:	calculate the standard Euclidean distance d_i between s_i and C
13:	Flist append d_i and a_i
14:	**end for**
15:	Maxa and Maxb ← box plot(*Flist*)
16:	*Flist* ← filter (Flist, d_i<Max)
17:	perform K-means clustering on *Flist*
18:	**for** each sample s_i in normal cluster(s) **do**
19:	*Nlist* append *si*
20:	*Alist remove si*
21:	**end for**
22:	**return *Nlist*, *Alist***

شکل ٦. شبه‌کد الگوریتم *K-IF*

مرحله اول: تعیین برچسب اولیه. در این مرحله، نمره ناهنجاری نمونه‌ها بکمک لگوریتم IF محاسبه و برچسب‌های اولیه مشخص و نمونه‌ها در دو خوشه نرمال و خوشه مشکوک به تقلب قرار می‌گیرند. در این مرحله، پارامتر نرخ آلودگی را دو برابر دانش پس‌زمینه در نظر می‌گیریم تا حداکثر تعداد نقاط ناهنجار به عنوان کاندید تقلب شناسایی شود. در این مرحله، هدف این است که کم‌ترین خطای طبقه‌بندی را داشته باشیم و همزمان نرخ تشخیص ناهنجاری را به حداکثر برسانیم.

مرحله دوم: محاسبه فاصله از مرکز خوشه نرمال. مرکز خوشه داده‌های نرمال در مرحله اولیه را محاسبه می‌کنیم و سپس فاصله اقلیدسی استاندارد شده از مقدار عناصر خوشه مشکوک تا مرکز خوشه نرمال را محاسبه می‌نماییم. فرض کنید $a = (x_{11}, x_{12}, \ldots, x_{1n})$ و $b = (x_{21}, x_{22}, \ldots, x_{2n})$ داده‌های مشاهده‌شده باشد، که در آن s_k انحراف معیار در هر بعد است و فاصله اقلیدسی استاندارد شده بین a و b به صورت رابطه زیر است:

$$d = \sqrt{\sum_{k=1}^{n} \left(\frac{x_{1k} - x_{2k}}{s_k} \right)^2} \qquad (٤)$$

هر چه فاصله تا مرکز خوشه کمتر باشد، داده‌ها به توزیع داده‌های نرمال نزدیک‌تر هستند، و هر چه فاصله بیشتر باشد، احتمال ناهنجار بودن داده بیشتر است.

مرحله سوم: فیلتر کردن براساس آستانه‌های بالا. مقادیر افراطی[1] در مقدار فاصله از مرکز خوشه نرمال از طریق نمودار جعبه‌ای [۱۲] فیلتر می‌شوند. فرض کنید Q_1 چارک بالایی و Q_3 چارک پایینی باشد. مشاهدات MAX و MIN را می‌توان به صورت زیر تعریف کرد:

$$MAX = Q_1 - 1.5 \times (Q_3 - Q_1) \quad (۵)$$

مقادیر افراطی با تغییر ویژگی‌های کلی داده‌ها، بر خوشه‌بندی بعدی تأثیر منفی می‌گذارند. بنابراین، در این مرحله، مقادیر دارای فاصله از مرکز خوشه نرمال بیشتر از MAX به عنوان ناهنجاری مستقیم لحاظ می‌شوند، و مقادیر زیر MIN به صورت نرمال در نظر گرفته می‌شود. چون نتیجه فاصله اقلیدسی همواره غیرمنفی است، ما مقادیر کمتر از MIN را در نظر نمی‌گیریم.

مرحله چهارم: خوشه بندی براساس K-means. داده‌های مربوط به نمونه‌های باقی‌مانده بکمک K-means برای خوشه‌های استفاده می‌شود. در این مرحله، از نمره سیلوئت برای تعیین مقدار بهینه K بر روی مجموعه داده‌ها استفاده می‌شود.

مرحله پنجم: محاسبه فاصله هر خوشه از مرکز خوشه نرمال. مرکز هر خوشه که توسط خوشه‌بندی K-means مشخص شد را محاسبه، و سپس فاصله استاندارد اقلیدسی آن از مرکز خوشه نرمال محاسبه می‌شود.

مرحله ششم: برچسب‌گذاری خوشه‌ها. خبرگان یک حد آستانه جهت تعیین خوشه (های) ناهنجار براساس این فاصله‌ها مشخص می کنند تا تمام داده‌ها به دو دسته نرمال و ناهنجار (تقلب) تقسیم شوند. آن‌ها با تحلیل داده‌های تاریخی و نمونه‌های تایید شده، میزان فاصله‌ای را که نشان‌دهنده تمایز قابل اعتماد بین نسخه‌های عادی و تقلبی است، مشخص می‌کنند.

در اجرای الگوریتم IF و K-Means، نکته قابل توجه این است که هر دو روش نسبت به پارامترهای تنظیم‌شده حساسیت دارند، الگوریتم IF بر مبنای ساخت درخت‌های تصمیم‌گیری و تقسیم تصادفی داده‌ها عمل می‌کند، و این ساختار سبب می‌شود که نویزهای تصادفی و موارد خارج از الگوهای قابل‌تشخیص، تأثیر نسبتاً کمی بر روی ساختار کلی درخت‌ها داشته باشد. در این مطالعه، پارامترهای اصلی مانند تعداد درختان و اندازه نمونه در هر درخت در IF با توجه به آزمایش‌های پیشین و تجربیاتی که از روی مجموعه داده‌های نمونه انجام شد، تنظیم شده است. همچنین، اثبات شده است که الگوریتم IF در مواجهه با نویزهای معمول در داده‌های با حجم بالا، به دلیل ساختار مقاومی که دارند، تا حد قابل قبولی مقاوم هستند. به همین دلیل، IF می‌تواند در تشخیص ناهنجاری‌های واقعی، که معمولاً از الگوهای رایج فاصله دارند، نسبتاً مقاوم باشد. علاوه براین، در الگوریتم پیشنهادی، با تنظیم نرخ آلودگی در یک مقدار نسبتا بالا و تفکیک خوشه مشکوک به عناصر نرمال و تقلبی بکمک الگوریتم K-Means، تاثیر نویزها تا حد زیادی در الگوریتم IF از بین می‌رود. الگوریتم K-Means به طور کلی حساسیت زیادی نسبت به نویز و نمونه‌های خارج از الگو دارد. اما، در صورت تنظیم دقیق پارامترهایی مثل K، می‌توان گفت K-Means در مقابل نویز مقاوم‌تر است، زیرا با انجام چندین تکرار و انتخاب بهترین نتیجه، اثر نویزها کاهش می‌یابد. براین این منظور، با استفاده از نمره سیلوئت، می‌توان تنظیمات و پارامترهای الگوریتم K-Means را به گونه‌ای بهبود داد که کمترین تأثیر را از نمونه‌های خارج از الگو یا نویز بپذیرد و در نتیجه، مقاومت کلی مدل در مقابل داده‌های ناپایدار یا متلاطم افزایش یابد.

٤.١.٣. بصری‌سازی و داشبورد (ماژول ٤)

پس از برچسب‌گذاری نسخه‌های پزشکی با کمک الگوریتم پیشنهادی، خروجی به صورت یک لیست شامل برچسب‌های نهایی هر نسخه پزشکی است که نشان‌دهنده تقلبی یا عادی بودن آن است. این برچسب‌ها در قالب یک ستون یا ویژگی اضافی به داده‌های ورودی در انبار داده مرحله دوم افزوده می‌شوند که در فرآیند مصورسازی مورد استفاده قرار می‌گیرد. در این ماژول سیستم کشف تقلب بیمه درمان، هدف اصلی تمرکز بر بصری‌سازی، یافتن و ارائه شواهد و ادله معتبر جهت تأیید و تبیین تصمیمات اتخاذ شده توسط این سیستم است. این ماژول نه تنها فرآیند تصمیم‌گیری را شفاف‌تر می‌کند، بلکه امکان تحلیل عمیق‌تر روابط بین بازیگران، خدمات/کالاها و رفتارهای غیرطبیعی را برای کارشناسان (کاربران) فراهم می‌سازد. هدف از این تحلیل‌ها، توانمندسازی کاربران در کسب اطلاع و درک بهتر از الگوهای زیرساختی و رفتاری مربوط به

1. Extreme

تقلب است، به گونه‌ای که بتوانند با اعتماد به داده‌های بصری، قضاوت‌های خود را در مورد پارامترهای ورودی و نتایج مرحله آموزش موتور هوشمند کشف تقلب، بازنگری و اصلاح کنند. این ماژول تحلیل‌هایی بر پایه شاخص‌های مختلف ارائه می‌دهد که سطح ناسازگاری بالای ادعاهای بیمه را تبیین و دلایل احتمالی آن را برای کاربران روشن می‌سازد. این ماژول، در قالب یک بسته نرم‌افزاری ارائه خواهد شد.

۲.۳. ارتباط میان اجزای چارچوب پیشنهادی

برای درک کامل و روشن‌تر نقش هر یک از ماژول‌های این چارچوب، همانطوری که در شکل ۲ مشاهده می‌شود، لازم است توجه داشت که این اجزا به صورت سلسله‌مرتب و در تعامل مستمر با یکدیگر عمل می‌کنند تا فرآیند تشخیص تقلب را به صورت پویا و انعطاف‌پذیر مدیریت کنند. در ابتدا، ماژول اول، یعنی مرحله‌ای که در آن دانش و فرضیات کارشناسان بیمه و پزشکان در خصوص روندهای تقلب شامل انواع تقلب و ویژگی‌های مورد نیاز برای کشف آن‌ها جمع‌آوری و مستندسازی می‌شود، نقش پایه‌ای و راهبردی برای سایر ماژول‌ها دارد. این ماژول، با تولید این فریم‌ورک‌های تخصصی توسط تیم‌های کارشناسان بیمه و پزشکی، اصول اولیه و داینامیک تقلب‌های بالقوه را مشخص می‌کند و امکان ایجاد ماژول‌های دیگر را فراهم می‌سازد. علاوه براین، در این ماژول در خصوص تعیین مقدار بهینه برخی از پارامترهای مربوط به الگوریتم‌های مورد استفاده مانند «حد آستانه» و «نرخ آلودگی» تصمیم‌گیری می‌شود تا موتور کشف تقلب در ماژول سوم بتواند با دقت بالاتری عمل کند. سپس، در ماژول دوم، داده‌ها را از پایگاه داده شرکت بیمه دریافت، پاکسازی و تکمیل و سپس به انبار داده مرحله اول در ماژول دوم منتقل می‌کند. سپس داده‌های غنی و ساختارمندتر را براساس فریم ورک‌های ماژول اول به شکل ویژگی‌های کشف تقلب تولید و در انبار داده مرحله دوم ذخیره می‌کند تا در اختیار موتور کشف تقلب در ماژول سوم قرار گیرد. این ساختار داده‌ای، نقش رابط میان دانش کارشناسان و موتور کشف تقلب را بر عهده دارد. همچنین ماژول دوم، نقش مهمی در تهیه اطلاعات مورد نیاز شامل مقادیر ویژگی‌های تقلب مربوط به نسخه‌های پزشکی و بازیگران مختلف دارد که در ماژول چهارم برای ارائه گزارش‌های تحلیلی و بصری‌سازی مورد استفاده قرار می‌گیرد.

در ادامه در ماژول سوم، موتور کشف تقلب، که هسته اصلی سیستم محسوب می‌شود، عمل می‌کند. این موتور ابتدا، با بهره‌گیری از الگوریتم IF، نسخه‌های پزشکی را به خوشه‌های نرمال و مشکوک تقسیم‌بندی می‌کند. سپس، به کمک الگوریتم خوشه‌بندی K-Means، نمونه‌های نرمال مرتبط با خوشه‌های مشکوک جدا شده و نمونه‌های تقلب شناسایی می‌شوند. پس از ارزیابی عملکرد الگوریتم، در صورت پایین بودن دقت موتور کشف تقلب، با برگشت به ماژول اول و با تغییر مقادیر حدآستانه و نرخ آلودگی در الگوریتم‌های مورد استفاده و یا حتی با توجه به ساختار منعطف چارچوب پیشنهادی جایگزین کردن الگوریتم‌ها موجود با الگوریتم‌های بهتر، مدلساز می‌توان درصدد بهبود دقت موتور کشف تقلب برآید. علاوه براین، در صورتی که ماژول سوم قادر نباشد نیازهای شرکت بیمه را در کشف تقلب‌های جدید برآورده کند، این امکان وجود دارد که ابتدا در ماژول اول این نوع تقلب‌ها و ویژگی‌های مورد نیاز برای کشف آن‌ها در قالب فریم ورک‌های تخصصی تعریف، سپس در ماژول دوم این ویژگی‌ها به انبار مرحله دوم اضافه، و نهایتاً در ماژول سوم الگوریتم‌ها را براساس انبار داده جدید اجرا کرد. در صورت عملکرد نامطلوب موتور کشف تقلب، این فرایند مجدد می‌تواند تکرار شود. نهایتا، ماژول چهارم، نقش حلقه اتصال و تسهیل‌کننده تعامل با کاربر و تصمیم‌گیری را برعهده دارد که در قالب ابزار تجسم و داشبورد مدیریتی نشان داده شده است. این ماژول اطلاعات مربوط به برچسب نسخه‌های پزشکی را به همراه داده‌های موجود در انبار داده مرحله یک و دو را دریافت و به کاربران امکان می‌دهد نتایج مدل و تحلیل‌های مختلف را در قالب گراف‌ها، نمودارها و جداول مشاهده و تحلیل کنند. این بخش، تحلیل‌های مختلف را براساس ویژگی‌های ادعاها و بازیگران به صورت تصویری و قابل فهم ارائه می‌دهد و به کاربر این امکان را می‌دهد تا با تنظیم پارامترها، افزودن روابط یا نشانگرهای جدید، فرآیند آموزش و تنظیم سیستم را در ماژول چهارم هدایت کند.

۳.۳. ارزیابی مدل‌ها

برای اهداف ارزیابی، شش معیار ارزیابی برای سنجش عملکرد مدل‌های یادگیری ماشین در نظر گرفته شده است. معیارهای ارزیابی شامل دقت (Accuracy)، فراخوانی (Recall)، صحت (Precision)، F1-Score و AUC-ROC هستند. Accuracy که در رابطه (۷) نشان داده شده است، نتیجه طبقه‌بندی‌های صحیح تقسیم بر تمام طبقه‌بندی‌ها است. با این حال، دقت همیشه یک معیار خوب نیست، به‌ویژه زمانی که داده‌ها

نامتوازن باشند. Recall، همان‌طور که در رابطه (۸) مشخص شده، نشان می‌دهد که مدل تا چه حد توانسته تمام موارد مثبت واقعی را شناسایی کند. Precision، همان‌طور که در رابطه (۹) بیان شده، نشان می‌دهد که از میان تمام نمونه‌هایی که مدل به‌عنوان ناهنجار یا مثبت شناسایی کرده، چند درصد واقعاً ناهنجار بوده‌اند [۱۰]. شاخص F1-Score میانگین هارمونیک دقت و حساسیت است [۴۱].

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \quad (۷)$$

$$\text{Sensitivity}(Recall) = \frac{TP}{TP + FN} \quad (۸)$$

$$Precision = \frac{TP}{TP + FP} \quad (۹)$$

$$F1 - Score = 2 \times \frac{\text{Precision} \times \text{Sensitivity}}{\text{Precision} + \text{Sensitivity}} \quad (۱۰)$$

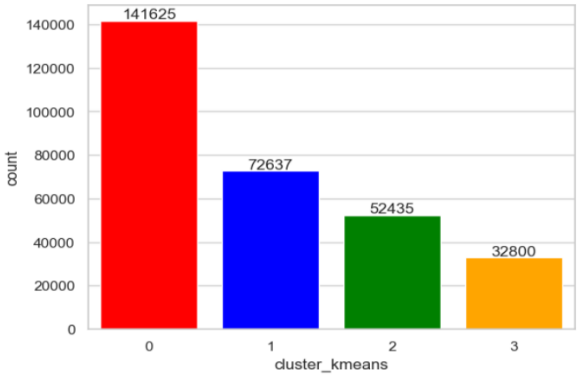
AUC-ROC: در زمینه کشف تقلب یا ناهنجاری در داده‌های بیمه، مالی یا پزشکی، AUC[1] به‌عنوان یک معیار قابل اعتماد برای سنجش کیفیت مدل‌های یادگیری ماشین مانند به کار گرفته می‌شود. این شاخص، مساحت زیر منحنی ROC[2] را اندازه‌گیری می‌کند و نشان‌دهنده توانایی مدل در تفکیک درست بین نمونه‌های نرمال و ناهنجار است. این منحنی نرخ مثبت‌های واقعی[3] را در برابر نرخ مثبت‌های کاذب[4] برای آستانه‌های مختلف نمایش می‌دهد. مقدار AUC عددی بین ۰ و ۱ است. مقدار ۱ نشان‌دهنده عملکرد عالی مدل در تفکیک ناهنجاری‌ها از داده‌های نرمال است. مقدار ۰.۵ معادل عملکرد تصادفی است [۱۰].

٤. تحلیل داده‌ها و یافته‌های پژوهش

٤.۱. پیاده‌سازی الگوریتم K-IF

در الگوریتم پیشنهادی K-IF، در مرحله اول ابتدا بکمک الگوریتم IF نمره ناهنجاری مجموعه داده‌های آموزش شامل ۷۹۳۵۸۹ نمونه‌ها/نسخه‌های پزشکی مربوط به شرکت بیمه دی محاسبه می‌شود. برای این منظور، نرخ آلودگی به عنوان پارامتر اصلی این الگوریتم ٤۰ درصد در نظر گرفته می‌شود که این مقدار دو برابر نسبت تقلبی است که توسط کارشناسان بیمه تعیین شده است تا بتوان نمونه‌های بیشتری را به عنوان نامزدهای تقلب جهت بررسی‌های عمیق‌تر در مراحل بعدی در اختیار داشت. سپس، این نمونه‌ها در دو خوشه نرمال و مشکوک به تقلب برچسب‌گذاری می‌شوند. در این مرحله، تعداد نمونه‌های مربوط به خوشه نرمال و خوشه مشکوک به ترتیب ٤۷٦۱۵۳ و ۳۱۷٤۳٦ نسخه‌های پزشکی تعیین شد. لازم به ذکر است برچسب نمونه‌های مربوط به خوشه نرمال به صورت قطعی بوده، اما برچسب نمونه‌های مربوط به خوشه مشکوک به صورت مقدماتی است و لازم است در مراحل بعدی به صورت قطعی مشخص شود. در مرحله دوم، مرکز خوشه نرمال محاسبه و سپس بکمک رابطه (٤) فاصله اقلیدسی استاندارد شده عناصر خوشه مشکوک از مرکز خوشه نرمال محاسبه می‌شود. در مرحله سوم، نمونه‌های خوشه مشکوک که مقدار فاصله آن‌ها از خوشه نرمال بالاتر از مقدار حد آستانه محاسبه شده در رابطه (۵) است، به عنوان نمونه‌های تقلب فیلتر می‌شوند. در این مرحله ۱۷۹۳۹ نسخه پزشکی فیلتر و به عنوان نمونه‌های تقلب شناسایی شد. در مرحله چهارم، برای تعیین برچسب ۲۹۹٤۷۹ نسخه باقی مانده در خوشه مشکوک، این نمونه‌ها بکمک الگوریتم K-Means جهت تعیین نسخه‌های تقلب و نرمال خوشه‌بندی می‌شوند. برای این منظور ابتدا بکمک نمره سیلوئت تعداد بهینه خوشه‌ها را مشخص می‌شود. پس از تعیین تعداد بهینه خوشه‌ها، الگوریتم K-Means با K=4 روی داده‌های آموزش مربوط به نمونه‌های مشکوک اجرا می‌شود. نتایج خوشه‌بندی این نمونه‌ها در شکل ۷ آورده شده است.

1. Area Under the Curve (AUC)
2. Receiver Operating Characteristic (ROC)
3. True Positive Rate
4. False Positive Rate



شکل ۷. خوشه‌بندی نقاط خوشه مشکوک بکمک *K-Means*

جهت چسب‌گذاری روی این چهار خوشه، پس از محاسبه مرکز هر یک از این خوشه‌ها، میانگین مجموع فاصله اقلیدوسی عناصر هر یک از این خوشه‌ها از مرکز خوشه نرمال محاسبه می‌شود. فواصل محاسبه شده به همراه میانگین ویژگی‌ها در هر یک از این خوشه‌ها در جدول ۱ آورده شده است.

جدول ۱. میانگین هر ویژگی برای هر خوشه و فاصله هر خوشه تا مرکز خوشه نرمال

خوشه	ویژگی ۱	ویژگی ۲	ویژگی ۳	...	ویژگی ۱۷	ویژگی ۱۸	ویژگی ۱۹	ویژگی ۲۰	فاصله تا مرکز خوشه نرمال
۰	۰.۳۹۵	۰.۳۲۷	−۰.۰۲۲	...	−۰.۰۹۹	۰.۲۰۱	۰.۱۳۰	۰.۲۱۹	**۲.۸۰**
۱	۰.۰۰۰	۰.۰۲۶	۱.۲۲۸	...	−۰.۰۸۸	−۰.۱۹۷	−۰.۲۰۶	۰.۰۸۳	**۳.۹۸**
۲	−۰.۰۷۹	−۰.۰۶۱	۷۰.۰۷	...	۰.۵۸۷	۲.۰۷۷	۱.۸۸۸	۰.۰۶۶	**۵.۹۳**
۳	−۰.۱۲۷	−۰.۱۵۴	۰.۲۴۹	...	۲.۵۱۵	−۰.۱۲۵	−۰.۱۲۰	−۰.۰۵۵	**۵.۳۲**

با توجه به فواصل محاسبه‌شده تا مرکز خوشه نرمال و میانگین ویژگی‌های مربوط به هر خوشه، امکان برچسب‌گذاری مستقیم این خوشه‌ها بکمک خبرگان فراهم می‌شود. فواصل بزرگ‌تر از مرکز خوشه نرمال و مقادیر بالاتر میانگین ویژگی‌ها در هر خوشه به معنی احتمال بالاتر تقلبی بودن نمونه‌ها در آن خوشه است. بر این اساس، خوشه‌های ۱، ۲ و ۳ به‌عنوان خوشه‌های تقلب و خوشه‌های ۰ به‌عنوان خوشه‌های نرمال برچسب‌گذاری شدند. تعداد نمونه‌های خوشه‌های ۱، ۲ و ۳ جمعاً ۱۵۷,۸۵۴ نسخه بوده که به‌عنوان نسخه‌های تقلب شناسایی می‌شوند. این تعداد نسخه تقلبی، با احتساب ۱۷,۹۳۹ نسخه تقلب که در مرحله سوم شناسایی شده بود، به ۱۷۵,۷۹۳ نسخه تقلبی می‌رسد. همچنین، تعداد نمونه‌های خوشه ۰، ۱۴۱,۶۲۵ نسخه پزشکی است و که به‌صورت نرمال برچسب‌گذاری می‌شوند. با احتساب ۴۷۶,۱۵۳ نسخه نرمال که در مرحله اول شناسایی شده بود، کل تعداد نسخه‌های تقلب به ۶۱۷,۷۷۸ نسخه می‌رسد. بدین ترتیب در حدود ۲۲ درصد از نسخه‌های پزشکی به عنوان نسخه تقلب برچسب‌گذاری می‌شوند.

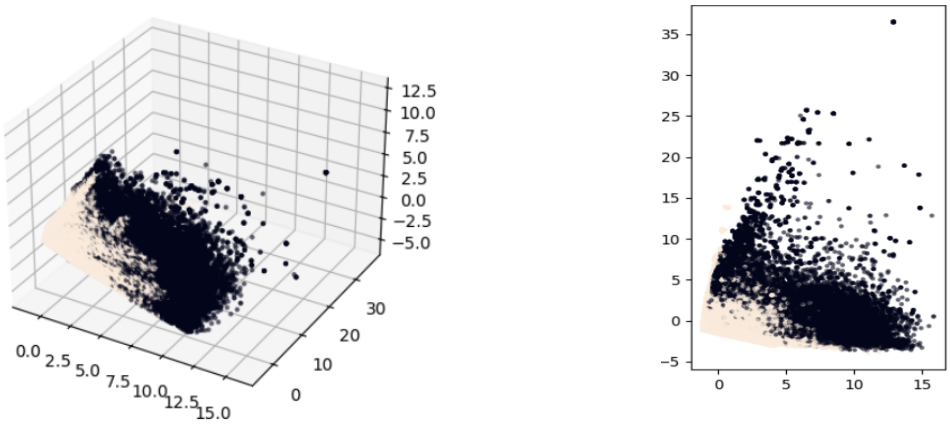
۲.٤. تحلیل عملکرد الگوریتم K-IF

در این قسمت عملکرد الگوریتم K-IF در مقایسه با الگوریتم‌های IF و AE برای شناسایی نقاط ناهنجار (تقلب) مورد بررسی قرار می‌گیرد. نخست، دو الگوریتم IF و K-IF، به‌طور مؤثری لبه‌ها و نقاط کم‌تعداد را تشخیص می‌دهند. در الگوریتم AE، مقادیر بالای loss نشان می‌دهد که نمونه‌ها از الگوی بازسازی شده توسط AE فاصله زیادی دارند؛ به عبارت دیگر به‌خاطر تفاوت‌های ساختاری از مدل گرایش خارج می‌شوند و احتمال وجود ناهنجاری در این نمونه‌ها بیشتر است. برای ارزیابی کارایی تشخیص تقلب در مطالعه حاضر، ابتدا از انجا که هیچ برچسب حقیقی برای مجموعه داده در دسترس نبود، با استفاده از الگوریتم‌های K-IF، IF و AE، برچسب‌های تقلب/نرمال نسخه‌های پزشکی پیش‌بینی می‌شود. دلیل مقایسه نتایج پیش‌بینی و تشخیص تقلب در دو الگوریتم K-IF و IF با الگوریتم AE، این است که این دو روش ویژگی‌های

مختلفی از داده‌ها را لحاظ می‌کنند: رویکردهای مبتنی بر IF یک روش مبتنی بر نمونه-محوری و انسجام‌یابی بدنه داده‌ها است که بر پایه شدت انحراف از رفتار عادی و عدم نیاز به بازسازی داده‌ها عمل می‌کند، در حالی که AE با یادگیری بازسازی داده‌ها از ورودی‌های نرمال، خطاهای بازسازی را به عنوان نشانگرهای غیرعادی اندازه‌گیری می‌کند؛ چنین تفاوتی امکان مقایسه بین دو منطق تشخیصی و ناتوانی نسبی هر کدام را در تشخیص تقلب در یک مجموعه داده با ابعاد بالا و پیچیدگی ذاتی آن‌ها و بدون برچسب‌های کامل نشان می‌دهد. در نتیجه، مقایسه این دو رویکرد به ارزیابی و تفسیر بهتر مدل‌های تشخیصی در برابر تغییرات داده، حساسیت به تغییرات ظریف در الگوها و قابلیت عمومی‌سازی آن‌ها برای نسخه‌های پزشکی می‌انجامد.

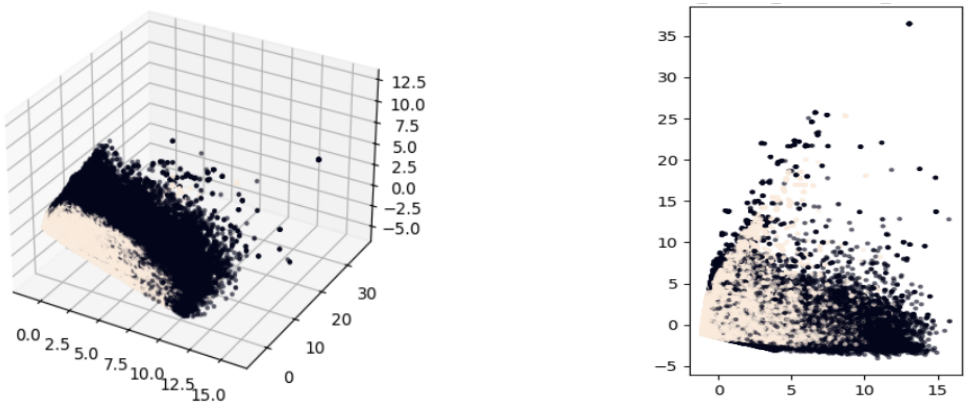
به منظور ارائه نمایش بصری برای مقایسه نتایج الگوریتم K-IF، IF و AE، مقادیر داده‌ها را با استفاده از روش PCA به ابعاد دو و سه کاهش دادیم تا بتوانیم توزیع نمونه‌های تقلب و نرمال را به‌صورت گرافیکی بررسی کنیم. لذا، نمودارهای دو بعدی و سه بعدی در شکل‌های ۸، ۹ و ۱۰ برای مقایسه نتایج پیش‌بینی این سه الگوریتم تهیه شد. نقاط نمونه‌ها به‌تفکیک براساس برآورد تقلب و نرمال به ترتیب به رنگ‌های سیاه و قهوه‌ای روشن نمایش داده شدند. با وجود نداشتن برچسب‌های واقعی، در این نمودارها توزیع تخمینی نمونه‌های تقلب در فضای با ابعاد پایین با دیدی از مناطق همگرا/پراکنده نمایش داده می‌شود. خوشه‌های تقلب و خطوط تشخیصی به‌وسیله این الگوریتم‌ها روی نمودارها قابل تشخیص‌اند و احتمال وجود ساختارهای متمایز بین دو حالت تقلب و نرمال را نشان می‌دهند. نمودارهای دو بعدی امکان تشخیص مناطق پرت و نمونه‌های شبیه‌به‌هم را فراهم می‌کند، در حالی که نمودارهای سه بعدی با نمایش سه بعدی امکان مشاهده الگوهای پیچیده‌تر و جداسازی بهتر بین دسته‌بندی‌ها را ارائه می‌دهند.

در نمودارهای شکل ۹ می‌توان مشاهده کرد که الگوریتم IF در یافتن لبه‌های داده و خوشه‌های بسیار کوچک عملکرد خوبی دارد. نتایج الگوریتم K-IF که در شکل ۸ قابل مشاهده است، تفاوت نقاط تقلب این الگوریتم با نقاط تقلب الگوریتم IF نسبتاً کم است، با این حال می‌توان نقاط بیشتری را در لبه‌ها به عنوان نقاط ناهنجار/تقلب پیدا کرد. این تفاوت بدلیل مدیریت نرخ آلودگی در الگوریتم K-IF و استفاده از الگوریتم K-Means در شناسایی کاندیدهای تقلب به عنوان نمونه‌های تقلب است. درحالی که الگوریتم AE نسبت به خوشه‌بندی‌هایی که از اکثریت دور هستند، حساسیت بالایی نشان می‌دهد؛ به همین دلیل همانطوری که در نمودارهای شکل ۱۰ مشاهده می‌شود نقاط تقلب شناسایی شده عمدتا فاصله بیشتری از اکثریت به عنوان مرکز خوشه نرمال دارند. در این نمودار، الگوریتم AE به نقاط دور از کل اکثریت داده‌ها حساس است، به‌عبارت دیگر تمایز میان الگوهای نادر یا خوشه‌های کم‌تعداد و خوشه‌های غالب را به‌طور نسبتاً قوی‌تری می‌تواند انجام بدهد. با این حال، برخلاف دو الگوریتم قبل قادر به شناسایی همه نمونه‌های تقلب در لبه‌ها نیست.



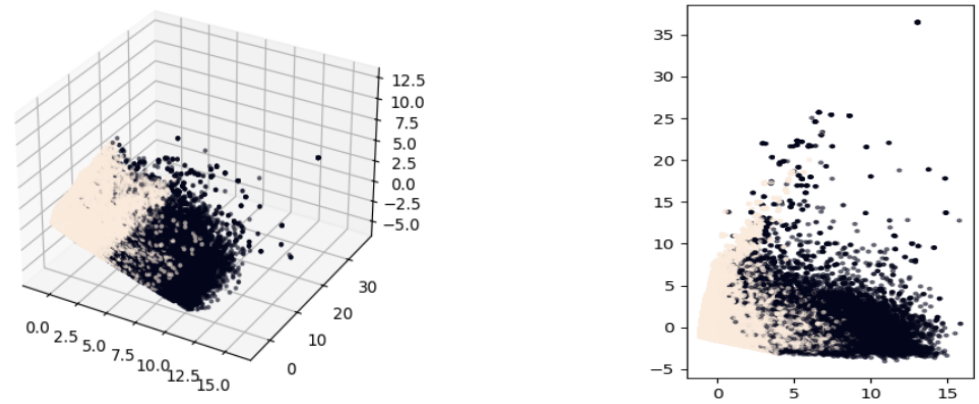
الف. نمودار دوبعدی ب. نمودار سه بعدی

شکل ۸. نقاط شناسایی شده به عنوان تقلب (آنومالی) با الگوریتم *K-IF*



الف. نمودار دوبعدی ب. نمودار سه بعدی

شکل ۹. نقاط شناسایی شده به عنوان تقلب (آنومالی) با الگوریتم *IF*



الف. نمودار دوبعدی ب. نمودار سه بعدی

شکل ۱۰. نقاط شناسایی شده به عنوان تقلب (آنومالی) با الگوریتم *AE*

۲.٤. ارزیابی الگوریتم‌های پیش‌بینی

۱.۲.٤. مجموعه داده‌ها

در این قسمت، برای ارزیابی و مقایسه عملکرد الگوریتم K-IF با سایر الگوریتم‌ها از سه مجموعه داده برای ارزیابی و مقایسه عملکرد الگوریتم K-IF با سایر الگوریتم‌ها استفاده خواهد شد. مجموعه داده اول (DataSet_1) مربوط به پروژه‌ای با عنوان «تشخیص کلاهبرداری بیمه سلامت و درمان با استفاده از یادگیری ماشین» است. دامنه پروژه محدود به داده‌های ساختارمند است و شامل ۵۰۰۰ ادعا با ویژگی‌های جمعیت‌شناسی، کدهای تشخیص و رویه‌ای، مبلغ ادعا، مالیات، شناسه ارائه‌دهنده و برچسب تقلب می‌شود. همچنین با استفاده از مهندسی ویژگی‌ها، به مجموعه ویژگی‌ها تاخیر ارسال، وضعیت تأیید، ورودی‌های تکراری و نشان‌گرهای ارائه‌دهنده مشکوک اضافه می‌شود. این پروژه در آدرس (https://github.com/124015001/healthcare-insurance-fraud-detection) در دسترس است. مجموعه داده دوم و سوم (DataSet_2 & DataSet_3) از پروژه‌ای تحت عنوان «کشف تقلب ارایه‌دهندگان پزشکی» از آدرس (https://www.kaggle.com/code/rohitrox/medical-provider-fraud-detection) گرفته شده است. در این پروژه، داده‌ها از سه فایل تشکیل شده‌اند: inpatient، outpatient و beneficiary. فایل اول و دوم شامل تاریخ‌های شروع و پایان ادعا، کدهای تشخیص ادعاها، کدهای رویه‌ای ادعاها، شناسه‌های پزشکان عمل و پزشک اصلی، مبالغ فرانشیز پرداختی و مبالغ بازپرداخت بیمه است. فایل سوم شامل داده‌های جمعیت‌شناسی بیمه‌شدگان از جمله جنسیت، نژاد، ایالت و شاخص‌های بیماری‌های مزمن است. در مراحل اولیه، ویژگی‌های جدیدی ایجاد و استانداردسازی شدند. برخی مقادیر تهی پر و برخی دیگر حذف شدند. پس از استخراج و ترکیب ویژگی‌های متنوع برای هر نسخه پزشکی، نهایتاً ویژگی‌های گوناگون برای هر پزشک محاسبه شده و به صورت رکوردهای جداگانه برای آموزش الگوریتم‌های با ناظر استفاده شدند. با

توجه به تعریف ویژگی‌های مختلف برای هر نسخه و برچسب‌دار بودن این نسخه‌ها و همچنین حجم مناسب داده‌ها، این مجموعه داده برای ارزیابی عملکرد الگوریتم K-IF به کار گرفته شد. این داده‌ها دارای ۱۵۷ ویژگی تعریف شده هستند. در ابتدا، برای آموزش و ارزیابی عملکرد الگوریتم‌های بدون ناظر، از تمام ویژگی‌ها استفاده شد. با این حال، به دلیل عملکرد غیرقابل قبول الگوریتم‌ها، مشخص شد که تعداد زیادی از ویژگی‌های تعریف‌شده ارتباط و همبستگی خوبی با برچسب‌های نسخه‌های پزشکی ندارند و این موضوع باعث کاهش دقت الگوریتم‌ها شده است. بنابراین، به‌منظور بهبود عملکرد، تصمیم گرفته شد مجموعه داده دیگر (DataSet_2) از روی همان مجموعه داده استخراج شود که تنها شامل ویژگی‌های مرتبط با برچسب‌های نسخه‌ها هستند. جدول ۲ مشخصات و اطلاعات مجموعه‌های داده استفاده شده برای ارزیابی الگوریتم‌ها را نشان می‌دهد.

جدول۲. اطلاعات مجموعه داده‌های مورد استفاده برای ارزیابی الگوریتم‌ها

نام مجموعه داده	تعداد رکوردها	تعداد ویژگی‌ها	نسبت تقلب
DataSet_1	۵۰۰۰	۹	۰.۱۰
DataSet_2	۵۵۸۲۱۱	۱۵	۰.۳۸
DataSet_3	۶۱۰۰۰۰	۱۵۷	۰.۲۵

۲.۲.٤. مقایسه عملکرد الگوریتم K-IF با الگوریتم‌های دیگر

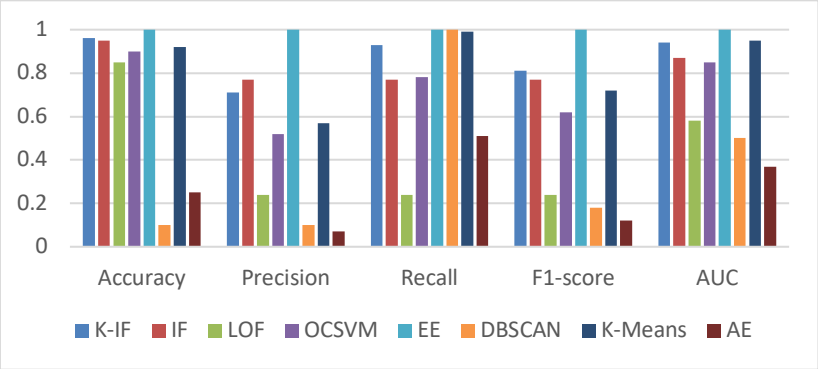
در این قسمت شش مدل یادگیری ماشین بر روی مجموعه داده اول تا سوم ارزیابی می‌شود. از پنج معیار برای ارزیابی عملکرد الگوریتم K-IF و مقایسه آن با عملکرد پنج مدل دیگر شامل IF، LOF، OCSVM، EE، DBSCAN، K-Means و AE استفاده می‌شود. در این قسمت، این هشت مدل با مجموعه‌های داده تعریف شده در جدول ۲ آموزش داده شده‌اند و براساس پنج شاخص عملکرد آن‌ها مقایسه و نمایش داده شده است. محدودیت شاخص Accuracy در ارزیابی عملکرد الگوریتم‌های کشف ناهنجاری (تقلب) این است که نسبت به کلاس‌بندی ناهمگون حساس نیست؛ به عنوان مثال، اگر تقلب تنها بخش کوچکی از داده‌ها باشد و مدل اکثر داده‌ها را به‌درستی نرمال تشخیص دهد، این شاخص می‌تواند بالا باشد در حالی که مدل تقلب‌ها را به خوبی تشخیص نمی‌دهد. همچنین، Accuracy به تفاوت بین FP و FN اهمیتی نمی‌دهد و از این رو برای سنجش کارایی در تشخیص تقلب که هدف اصلی است، شاخص‌های مانند Precision، Recall و F1-score یا AUC بسیار مناسب‌تر هستند. به همین دلیل، در این مطالعه از این شاخص‌ها جهت ارزیابی عملکرد الگوریتم‌ها استفاده خواهد شد. به طور کلی، هر چه Precision، Recall و f1-score بالاتر باشد، الگوریتم تشخیص ناهنجار (تقلب) بهتر عمل می‌کند. با این حال، Precision و Recall محدودیت‌هایی دارند. در حالت‌های افراطی، اگر تنها یک نقطه ناهنجار تشخیص داده شود، Precision صددرصد است در حالی که Recall بسیار پایین است؛ و اگر همه داده‌ها به عنوان نقاط ناهنجار تشخیص داده شوند، Recall به معنای صددرصد است و Precision بسیار پایین است. اهمیت هر یک از این دو شاخص به مسئله مورد بررسی بستگی دارد. در کشف تقلب بیمه درمان، به دلیل اهمیت بالای نرخ Recall، این شاخص در این مطالعه از Precision مهم‌تر در نظر گرفته شده است. در واقع، هزینه‌های مربوط به پیش‌بینی منفی کاذب بیشتر از هزینه‌های پیش‌بینی مثبت کاذب است. هزینه منفی کاذب شامل پرداخت اشتباه هزینه خسارت به بیمه‌شده است که معمولاً مبالغ بالایی دارد، در حالی که هزینه مثبت کاذب شامل بررسی مجدد پرونده‌ها به صورت دستی است که معمولاً ناچیز است. به عبارت دیگر، اگر نسخه سالم باشد و تقلب تشخیص داده شود، این می‌تواند به بیمه‌شده منتقل شود، بیمه‌شده مستندات را ارائه می‌دهد و در نهایت پرونده دوباره ارزیابی شده و حق بیمه بازگردانده می‌شود.

برای الگوریتم K-IF، ما تعداد estimators را برای هر سه مجموعه داده بهینه‌سازی کردیم. مقدار پیش‌فرض ۱۰۰ درخت است که طول مسیر به خوبی پیش از عبور از این عدد همگرا می‌شود. تعداد درخت‌ها یکی از هایپرپارامترهای این الگوریتم است. هایپرپارامتر دیگر نرخ آلودگی است که نسبت نقاط ناهنجار تشخیص داده شده در داده‌ها را تعیین می‌کند. مقدار آن از ۰ تا ۵۰ درصد متغیر است و در این الگوریتم، مقدار آن را دو برابر مقدار برآوردی در نظر گرفته می‌شود تا عملکرد مدل را در مقایسه با سایر مدل‌ها بررسی شود. برای سایر پارامترها و هایپرپارامترها،

مقادیر پیش‌فرض را نگه داشته می‌شود. همچنین مقدار پارامترها در پنج الگوریتم دیگر برای هر سه مجموعه داده به صورت بهینه در نظر گرفته شد. نتایج عملکرد الگوریتم‌ها بر روی مجموعه داده اول به ترتیب در جدول ۳ و شکل ۱۱ نشان داده شده است. با توجه به ستون شاخص Recall و AUC مشاهده می‌شود که اگرچه دو نسخه از الگوریتم IF نتایج بسیار خوبی ارائه می‌دهند، ولی عملکرد الگوریتم K-IF از الگوریتم IF بهتر است. همچنین عملکرد الگوریتم K-IF از نظر این دو شاخص از الگوریتم‌های LOF، OCSVM، DBSCAN و AE به صورت معناداری بهتر است. با این حال، الگوریتم‌های EE و K-Means نسبت به الگوریتم K-IF حداقل از نظر یک شاخص برتری معناداری دارند. نتایج نشان می‌دهد، در یک مجموعه داده کوچک و با نرخ آلودگی ۱۰ درصد عملکرد الگوریتم EE از نظر همه شاخص‌ها ۱۰۰ درصد است که عدد بسیار بالایی است و الگوریتم K-IF با امتیاز Recall و AUC به ترتیب برابر با ۹۳ و ۹۴ درصد بعد از الگوریتم K-Means با امتیاز ۹۹ و ۹۵ درصد در رتبه سوم قرار می‌گیرد. عملکرد سایر الگوریتم‌ها به غیر از OCSVM در حد قابل قبولی قرار ندارد.

جدول۳. تحلیل مقایسه عملکرد مدل‌های پیش‌بینی برای *DataSet_1*

مدل	Accuracy	Precision	Recall	F1-score	AUC	زمان (ثانیه)
K-IF	۰.۹۶	۰.۷۱	۰.۹۳	۰.۸۱	۰.۹۴	۱
IF	۰.۹۵	۰.۷۷	۰.۷۷	۰.۷۷	۰.۸۷	۱
LOF	۰.۸۵	۰.۲۴	۰.۲۴	۰.۲۴	۰.۵۸	۱
OCSVM	۰.۹۰	۰.۵۲	۰.۷۸	۰.۶۲	۰.۸۵	۱
EE	۱	۱	۱	۱	۱	۱
DBSCAN	۰.۱	۰.۱	۱	۰.۱۸	۰.۵	۱
K-Means	۰.۹۲	۰.۵۷	۰.۹۹	۰.۷۲	۰.۹۵	۱
AE	۰.۲۵	۰.۰۷	۰.۵۱	۰.۱۲	۰.۳۷	۵



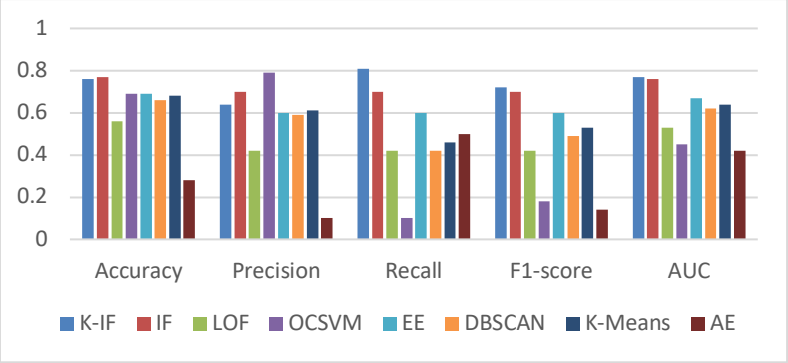
شکل ۱۱. مقایسه نتایج مدل‌های پیش‌بینی برای *DataSet_1*

همچنین شاخص‌های عملکرد هشت مدل بر روی مجموعه داده دوم برای یافتن مؤثرترین مدل برای کشف تقلب در بیمه درمان محاسبه شده است که نتایج آن به ترتیب در جدول ٤ و شکل ۱۲ نشان داده شده است. بالاترین Recall بین هشت مدل توسط K-IF بدست آمده و برابر با ۸۱ درصد است که عدد خوبی برای یک مدل یادگیری بدون ناظر است. دومین عملکرد بالا از نظر این دو شاخص برای مدل IF برابر با ۷۰ درصد است و سپس مدل EE است که برابر با ٦۰ درصد شده است. نهایتاً، معیار AUC نیز برای همه مدل‌های پیش‌بینی محاسبه و گزارش شده است. این شاخص برای شناسایی مؤثرترین مدل در کشف تقلب در مجموعه داده بیمه درمان بسیار حائز اهمیت است. مشخص است که مدل K-IF از نظر امتیاز AUC بر سایر مدل‌ها برتری قابل توجهی دارد. مدل IF در رتبه دوم است و EE در رده سوم قرار دارد. به طور کلی، این سه

مدل پیش‌بینی Recall و AUC بسیار قابل توجهی نشان می‌دهند و حاکی از قدرت هر سه مدل در یک مجموعه داده متوسط برای کشف تقلب در بیمه درمان هستند. سایر مدل‌های پیش‌بینی عملکرد پایینی از نظر شاخص‌های عملکردی از خود نشان می‌دهند.

جدول ٤. تحلیل مقایسه عملکرد مدل‌های پیش‌بینی برای *DataSet_2*

مدل	Accuracy	Precision	Recall	F1-score	AUC	زمان (ثانیه)
K-IF	۰.۷۶	۰.۶۴	**۰.۸۱**	۰.۷۲	۰.۷۷	۳۸
IF	۰.۷۷	۰.۷۰	۰.۷۰	۰.۷۰	۰.۷۶	۵۶
LOF	۰.۵۶	۰.۴۲	۰.۴۲	۰.۴۲	۰.۵۳	۳۸
OCSVM	۰.۶۹	۰.۷۹	۰.۱۰	۰.۱۸	۰.۴۵	۶۵۰۰
EE	۰.۶۹	۰.۶۰	۰.۶۰	۰.۶۰	۰.۶۷	۲۶۳
DBSCAN	۰.۶۶	۰.۵۹	۰.۴۲	۰.۴۹	۰.۶۲	۲۴۷
K-Means	۰.۶۸	۰.۶۱	۰.۴۶	۰.۵۳	۰.۶۴	۱۰
AE	۰.۲۸	۰.۱۰	۰.۵۰	۰.۱۴	۰.۴۲	۳۸۴۰

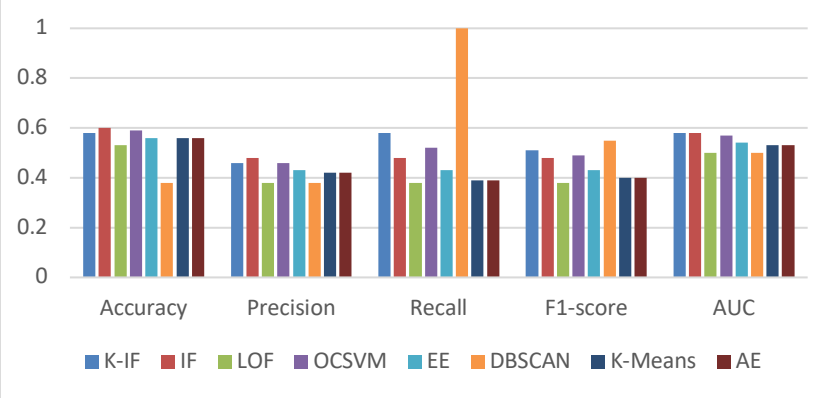


شکل ۱۲. مقایسه نتایج مدل‌های پیش‌بینی برای *DataSet_2*

همچنین شاخص‌های عملکرد هشت مدل بر روی مجموعه داده سوم برای یافتن بهترین مدل برای کشف تقلب در بیمه درمان محاسبه شده است که نتایج آن به ترتیب در جدول ۵ و شکل ۱۳ نشان داده شده‌اند. بالاترین Recall بین هشت مدل توسط DBSCAN بدست آمده و برابر با ۱۰۰ درصد است که عدد بسیار خوبی برای یک مدل یادگیری بدون ناظر است. با این حال، این مدل دارای Precision خیلی پایین ۳۸ درصد است که عملکرد آن را غیرقابل قبول می‌کند. در واقع این مدل بیشتر نسخه‌های پزشکی را به عنوان تقلب تشخیص می‌دهد که منجر به عملکرد بالای شاخص Recall و عملکرد پایین شاخص Precision شده است. بهترین عملکرد از نظر دو شاخص Precision و Recall برای مدل K-IF به ترتیب برابر با ٤٦ و ٥٨ درصد است و سپس مدل IF که به ترتیب برابر با ٤٨ و ٤٨ درصد است. نهایتاً، معیار AUC نیز برای همه مدل‌های پیش‌بینی محاسبه و گزارش شده است. مشخص است که مدل K-IF و مدل IF که از نظر امتیاز AUC برابر بوده بر سایر مدل‌ها برتری قابل توجهی دارند. مدل OCSVM از نظر این شاخص در رتبه دوم است و EE در رده سوم قرار می‌گیرد. به طور کلی، مدل K-IF همزمان از نظر هر دو شاخص Recall و AUC در یک مجموعه داده بزرگ عملکرد بهتری نسبت به سایر مدل‌ها در کشف تقلب در بیمه درمان نشان می‌دهد. سایر مدل‌های پیش‌بینی عملکرد پایین‌تری از نظر شاخص‌های عملکردی از خود نشان می‌دهند.

جدول ۵. تحلیل مقایسه عملکرد مدل‌های پیش‌بینی برای *Data_Set_3*

مدل	Accuracy	Precision	Recall	F1-score	AUC	زمان (ثانیه)
K-IF	۰.۵۸	۰.۴۶	**۰.۵۸**	۰.۵۱	**۰.۵۸**	۶۴۴
IF	۰.۶۰	۰.۴۸	۰.۴۸	۰.۴۸	**۰.۵۸**	۶۲۳
LOF	۰.۵۳	۰.۳۸	۰.۳۸	۰.۳۸	۰.۵۰	۴۳۴۹
OCSVM	۰.۵۹	۰.۴۶	۰.۵۲	۰.۴۹	۰.۵۷	۸۳۹۳
EE	۰.۵۶	۰.۴۳	۰.۴۳	۰.۴۳	۰.۵۴	۱۶۳۳
DBSCAN	۰.۳۸	۰.۳۸	**۱**	۰.۵۵	۰.۵۰	۱۴۳۱
K-Means	۰.۵۶	۰.۴۲	۰.۳۹	۰.۴۰	۰.۵۳	۱۲۰
AE	۰.۵۶	۰.۴۲	۰.۳۹	۰.۴۰	۰.۵۳	۳۵۲۶



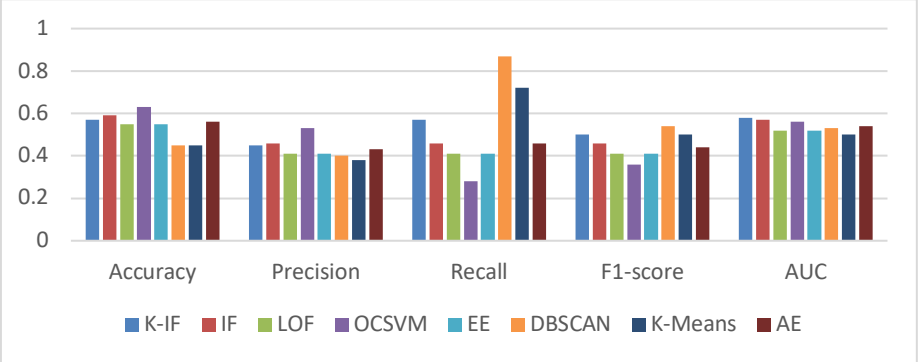
شکل ۱۳ مقایسه نتایج مدل‌های پیش‌بینی برای *Data_Set_3*

در ادامه، برای کاهش زمان محاسباتی الگوریتم‌های مختلف بر روی مجموعه داده سوم، بکمک رویکرد PCA تعداد متغیرهای این مجموعه داده از ۱۵۷ ویژگی به ۱۵ ویژگی کاهش داده شد. شاخص‌های عملکرد هشت مدل بر روی این مجموعه داده فشرده شده برای یافتن بهترین مدل برای کشف تقلب در بیمه درمان محاسبه شده است که نتایج آن به ترتیب در جدول ٦ و شکل ١٤ نشان داده شده‌اند. بالاترین Recall بین شش مدل پیش‌بینی‌شده مجددا توسط DBSCAN بدست آمده و برابر با ۸۷ درصد است که عدد بسیار خوبی برای یک مدل یادگیری بدون ناظر است. با این حال، این مدل دارای Precision خیلی پایین ٤٠ درصد است که عملکرد آن را غیرقابل قبول می‌کند. بهترین عملکرد از نظر دو شاخص Precision و Recall برای مدل K-IF به ترتیب برابر با ٤٥ و ٥٧ درصد است و سپس مدل IF که به ترتیب برابر با ٤٦ و ٤٦ درصد است. نهایتاً، معیار AUC نیز برای همه مدل‌های پیش‌بینی محاسبه و گزارش شده است. مشخص است که مدل K-IF و سپس مدل IF به ترتیب با امتیار ۵۸ و ۵۷ درصد دارای AUC برابر بوده بر سایر مدل‌ها برتری دارند. به طور کلی، مدل K-IF همزمان از نظر هر دو شاخص Recall و AUC در یک مجموعه داده بزرگ فشرده شده عملکرد بهتری نسبت به سایر مدل‌ها در کشف تقلب بیمه درمان نشان می‌دهد. سایر مدل‌های پیش‌بینی عملکرد پایین‌تری از نظر شاخص‌های عملکردی از خود نشان می‌دهند. مشاهده می‌شود که با فشرده‌سازی مجموعه داده‌ها بکمک PCA شاخص Recall برای الگوریتم K-IF و IF به ترتیب با ۱ و ۲ درصد کاهش به ۵۷ و ٤٦ درصد رسیده است. با این حال شاخص AUC برای الگوریتم K-IF بدون تغییر و برای الگوریتم IF با یک درصد کاهش به ۵۷ درصد رسیده است. این نتایج نشان می‌دهد که با بکارگیری رویکرد PCA در این دو الگوریتم پایداری نتایج حفظ می‌شود و کاهش ابعاد داده‌ها تاثیر معناداری روی عملکرد مدل‌ها نمی‌گذارد. این یافته‌ها بر روی مجموعه داده‌های با ابعاد بالاتر احتمالا تقویت شود. از سوی دیگر، این تغییرات برای الگوریتم LOF و EE به صورت مثبت گزارش شده است و به ترتیب باعث بهبود ۳ و ۲ درصدی شاخص Recall و AUC در الگوریتم LOF و بهبود ۷ و ۱ درصدی این دو شاخص در الگوریتم EE شده

است. به نظر می‌رسد عملکرد این دو الگوریتم در مجموعه داده‌های با تعداد ویژگی‌های بالا بسیار تحت تاثیر ابعاد داده‌های مسئله قرار می‌گیرند و بکارگیری رویکرد PCA تاثیر مثبتی روی عملکرد آن‌ها دارد. با این حال، بکارگیری رویکرد PCA تاثیر منفی روی عملکرد الگوریتم‌های OCSVM و EE از نظر هر دو شاخص Recall و AUC دارد. همچنین، یافته‌ها تاثیر فشرده‌سازی داده‌ها بر روی عملکرد الگوریتم‌های DBSCAN و K-Means را به صورت متفاوتی گزارش می‌دهد، به طوری که باعث بدتر شدن ۱۳ درصدی شاخص Recall و بهبود ۳ درصدی شاخص AUC برای الگوریتم DBSCAN و بهبود ۳۳ درصدی شاخص Recall و کاهش ۳ درصدی شاخص AUC برای الگوریتم K-Means شده است. به نظر این دو الگوریتم نیز به شدت تحت تاثیر کاهش ابعاد داده‌ها توسط PCA قرار دارند.

جدول ٦. تحلیل مقایسه عملکرد مدل‌های پیش‌بینی ترکیب شده با *PCA* برای *Data_Set_3*

مدل	Accuracy	Precision	Recall	F1-score	AUC	زمان (ثانیه)
K-IF	۰.۵۷	۰.۴۵	**۰.۵۷**	۰.۵۰	۰.۵۸	۳۸
IF	۰.۵۹	۰.۴۶	**۰.٤٦**	۰.۴۶	۰.۵۷	۴۲
LOF	۰.۵۵	۰.۴۱	۰.۴۱	۰.۴۱	۰.۵۲	۱۵۷۰
OCSVM	۰.۶۳	۰.۵۳	۰.۲۸	۰.۳۶	۰.۵۶	۶۷۳۳
EE	۰.۵۵	۰.۴۱	۰.۴۱	۰.۴۱	۰.۵۲	۱۷۳
DBSCAN	۰.۴۵	۰.۴۰	۰.۸۷	۰.۵۴	۰.۵۳	۵۸۶
K-Means	۰.۴۵	۰.۳۸	۰.۷۲	۰.۵۰	۰.۵۰	۱۰
AE	۰.۵۶	۰.۴۳	۰.۴۶	۰.۴۴	۰.۵۴	۳۸۵۰



شکل ١٤. مقایسه نتایج مدل‌های پیش‌بینی ترکیب شده با *PCA* برای *Data_Set_3*

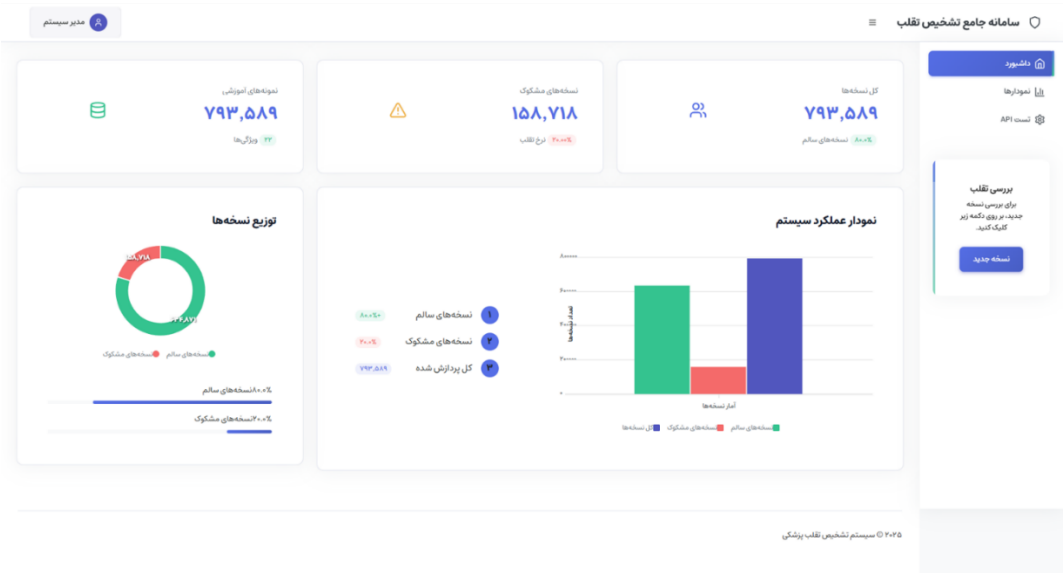
همچنین نتایج عملکرد هشت الگوریتم از نظر کارایی محاسباتی برای هر سه مجموعه داده در جداول (٦)-(٣) گزارش شده است. نتایج نشان می‌دهد که زمان محاسباتی همه الگوریتم‌ها به استثنای الگوریتم EE در مجموعه داده اول بدلیل تعداد پایین رکوردها به حدود ۱ ثانیه می‌رسد. در مجموعه داده دوم و سوم الگوریتم K-Means کمترین زمان محاسباتی را دارد و سپس دو الگوریتم K-IF و IF با زمان محاسباتی تقریبا یکسان در رتبه‌های بعدی قرار دارند. از سوی دیگر، نتایج نشان می‌دهد که این سه الگوریتم از نظر زمان محاسباتی در مجموعه داده‌های با تعداد رکورد بالا تفاوت معناداری با دیگر الگوریتم‌ها به ویژه با الگوریتم‌های OCSVM و AE دارند. می توان نتیجه‌گیری کرد با توجه به عملکرد بهتر الگوریتم K-IF در مجموعه داده‌های بزرگ از نظر شاخص‌های Recall و AUC و همچنین کارای محاسباتی بالای آن نسبت به سایر الگوریتم‌ها، این الگوریتم توجیه منطقی در کشف تقلب بیمه درمان دارد. علاوه براین، با توجه به پایداری نتایج این الگوریتم نسبت به کاهش ابعاد داده‌ها بکمک رویکرد PCA در مجموعه داده‌های بزرگ، این الگوریتم می تواند به عنوان یک راه حل اقتصادی برای داده‌های با حجم بالا توصیه شود.

۳.٤. نرم‌افزار کشف تقلب طراحی شده

۱.۳.٤. مشخصات و محیط نرم‌افزار

نرم‌افزار کشف تقلب بیمه درمان که در پژوهش حاضر طراحی شده است، «سیستم هوشمند تشخیص تقلب بیمه درمان» را به عنوان یک راه‌حل جامع و پیشرفته برای شرکت‌های بیمه معرفی می‌کند که مبتنی بر ترکیب یادگیری ماشین و تحلیل داده‌ها است. طراحی این نرم‌افزار از دو بخش اصلی API Backend (Flask) و Portal Frontend (React) تشکیل شده است. بخش بک‌اند از زبان برنامه نویسی Python 3.x با فریم‌ورک Flask استفاده و معماری آن بر پایه RESTful API باBlueprint Pattern بنا شده است تا یک API قدرتمند و گسترش‌پذیر فراهم کند. بخش فرانت‌اند نیز از زبان برنامه‌نویسی TypeScript با فریم‌ورک React ۱۹ استفاده کرده است.

بخش Portal Frontend با استفاده از React و TypeScript طراحی شده است تا رابط کاربری کاربرپسند و واکنش‌گرا ارائه دهد. این مولفه از ابزارهای مدرن مانند Vite برای بیلد، React Router DOM برای مسیردهی، Axios برای مدیریت درخواست‌ها و Recharts برای نمودارها به‌طور یکپارچه استفاده می‌کند تا تجربه کاربری بهینه‌ای ارائه شود. با تمرکز بر مقیاس‌پذیری بالا و کارایی، این سیستم نه تنها نمایشگرهای گرافیکی دقیق از نتایج تحلیل‌ها را فراهم می‌کند بلکه امکان اتصال امن و کارآمد به Backend و مدیریت داده‌های بیمه درمان را نیز فراهم می‌آورد. برای ذخیره‌سازی داده‌ها از MariaDB استفاده است؛ تبعیت از استانداردهای SQL و کارایی مناسب، قابلیت مقیاس‌پذیری و تراکنش‌های ACID قوی، و ساده‌بودن نگهداری با اکوسیستم گسترده از مزایای این ابزار است. MariaDB نسخه منبع‌باز با بهبودهای امنیتی و عملکردی نسبت به MySQL است، از موتورهایی مانند InnoDB پشتیبانی می‌کند و امکان مدیریت کلیدهای خارجی و قفل‌گذاری سطح سطور را فراهم می‌کند. همچنین با جامعه فعال و ابزارهای گوناگون، نگهداری، گزارش‌گیری و پایش داده‌ها را تسهیل می‌کند. مدل تشخیص تقلب با الگوریتم IF پیاده‌سازی شده است تا بتواند به‌طور هوشمند و کارآمد الگوهای تقلبی را استخراج و تشخیص دهد. همچنین این سیستم از ۲۰ شاخص ریسک برای ارزیابی نسخه‌های پزشکی، و از مجموعه‌ای از کنترل‌های امنیتی، اعتبارسنجی و پاک‌سازی داده‌ها بهره می‌برد. امکان بهره‌برداری از نرم‌افزار بر روی سرور اختصاصی فراهم شده و یک دامنه اینترنتی نیز برای دسترسی به آن در نظر گرفته شده است. نمای محیط کاربری نرم‌افزار در شکل ۱۵ نمایش داده شده است.

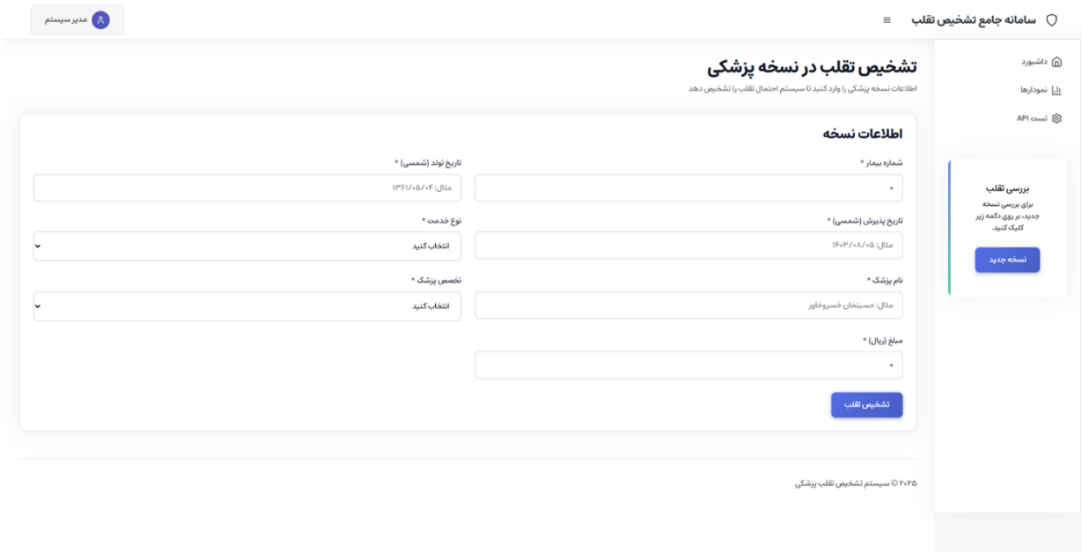


شکل ۱۵. نمای کلی محیط کاربری نرم‌افزار کشف تقلب بیمه درمان

همچنین امکان نمایش گزارشات مختلف به صورت نمودارهای تحلیلی در نرم‌افزار کشف تقلب بیمه درمان فراهم شده است.

۲.۳.٤. کشف تقلب در نرم‌افزار طراحی شده

در ادامه نحوه پیش‌بینی تقلبی یا نرمال بودن یک نسخه پزشکی توضیح داده می‌شود. شکل ۱٦ صفحه ورود اطلاعات نسخه پزشکی را نشان می‌دهد که شامل اطلاعات لازم برای پیش‌بینی بر روی نرم‌افزار است. این اطلاعات شامل شناسه بیمار (ID)، نام بیمار، تاریخ تولد، تاریخ پذیرش، نوع خدمت ارائه شده، نام و تخصص پزشک معالج و مبلغ کل نسخه می‌شود.



شکل ۱٦. محیط نرم‌افزار برای تشخیص تقلب در یک نسخه پزشکی

جهت استفاده از این اطلاعات برای پیش‌بینی تقلبی یا نرمال بودن این نسخه، لازم است این اطلاعات به مجموعه داده‌های قبلی نسخه‌های پزشکی (داده‌های آموزش) افزوده شود تا بتوان ویژگی‌های مربوط به این نسخه را محاسبه کرد. این ادغام ضروری است، چرا که بیشتر ویژگی‌ها به صورت دوره‌ای یا تجمعی هستند و لازم است اطلاعات نسخه‌های قبلی نیز در نظر گرفته شود. برای مثال، برای محاسبه ویژگی دفعات مراجعه بیمار با کد "٤٨٩٢٨" به پزشک با کد "٤٥٢٣" لازم است نسخه‌های قبلی مربوط به مراجعه این بیمار نزد این پزشک دیده شود. پس از محاسبه این ویژگی‌ها بکمک توابعی که از قبل برای این منظور بهینه شده است، این ویژگی‌ها بی‌مقیاس و به صورت یک آرایه نرمالیزه وارد تابع پیش‌بینی الگوریتم K-IF می‌شوند. علاوه بر پیش‌بینی نسخه‌های پزشکی، امکان محاسبه و نمایش مقدار ویژگی‌ها به صورت شاخص ریسک نیز در نرم‌افزار فراهم شده است.

۵. نتیجه‌گیری و پیشنهادها

در پژوهش حاضر یک چارچوب هوشمند ماژولار کشف تقلب بر پایه الگوریتم‌های یادگیری بدون نظارت برای کشف تقلب و رفتارهای سوءاستفاده‌گرانه در صنعت بیمه سلامت توسعه داده شد که ترکیبی از الگوریتم‌های IF و K-Means است. این چارچوب مستقل از بازیگران و خدمات است، قابل پیکربندی و توسعه بوده و به‌راحتی در محیط پویا تقلب و رفتارهای تخریبگر قابل انطباق است. همچنین شامل ابزار تجسمی است که زمان لازم برای کاربران را در فرایند یافتن و نظارت بر بازیگران پس از آنکه نرم‌افزار طراحی شده کاربر را از وجود ادعاهای پرریسک مطلع کرد، به‌طور قابل توجهی کاهش می‌دهد. چارچوب توسعه‌یافته از جمله معدود ابزارهایی است که تقریباً تمامی بازیگران و خدمات حوزه بیمه درمان را در بر می‌گیرد و راه‌حلی ارائه می‌دهد که می‌تواند پاسخگوی نیاز به ابزار پشتیبانی تصمیم‌گیری باشد که بتواند معیارهای ریسک را به تراکنش‌های خسارات بیمه‌ای تخصیص دهد.

این چارچوب شامل چهار ماژول کلیدی است. نخستین ماژول، مرحله‌ای را در بر می‌گیرد که در آن دانش و دیدگاه‌های کارشناسان درباره هدف و فرضیات فرایند یادگیری ماشین گنجانده می‌شود و این امر از طریق فریم‌ورک‌های تولیدی توسط تیمی از کارشناسان بیمه و پزشکی در خصوص انواع تقلب و ویژگی‌های لازم جهت شناسایی آن‌ها محقق می‌شود. ماژول دوم، یک انبار داده دو مرحله‌ای است که براساس بازیگران، خدمات/کالاها و ویژگی‌های استخراج‌شده از فریم‌ورک تقلب در ماژول اول ساخته می‌شود. در واقع، در چارچوب پیشنهادی، با در هم آمیختن تجربه کارشناسان، فریم ورک تقلب تهیه و بازیگران و ویژگی‌های مربوط به فریم‌ورک انواع تقلب انتخاب می‌شوند. سپس از کارشناسان خواسته می‌شود نظر خود را در رابطه با انواع تقلب و ویژگی‌های لازم برای کشف آن‌ها بیان کنند. اجماع کارشناسان روی تقلب‌های مهم و قابل شناسایی از روی داده‌های موجود و سپس تعریف ویژگی‌های لازم برای کشف آن‌ها طی چند آزمایش در نظر گرفته می‌شود. تا زمانی که نتایج آزمایش‌های مربوط به اجرای الگوریتم‌ها در ماژول سوم از لحاظ شاخص‌های ارزیابی رضایت‌بخش باشند، تکرارهای آزمایشی ادامه می‌یابد. در نهایت، بیست ویژگی مهم به‌عنوان ویژگی‌های تاثیرگذار در کشف تقلب استخراج و انتخاب شدند. سومین ماژول، موتور کشف تقلب است که از الگوریتم پیشنهادی K-IF برای کشف تقلب استفاده می‌کند. چهارمین ماژول، ابزارهای تجسم و داشبورد مدیریتی است که تحلیل‌های گوناگون را بر پایه ورودی‌های مربوط به ادعاها و/یا بازیگران ارائه می‌دهد و امکان تعامل کاربر با آموزش موتور کشف تقلب هوشمند را فراهم می‌کند.

در ماژول سوم به عنوان موتور کشف تقلب، چند مدل پیش‌بینی تقلب به کار گرفته شد، تا بهترین تقلب در بیمه درمان انتخاب شود. در این پژوهش، الگوریتم K-IF به دلیل چندین مزیت نسبت به سایر تکنیک‌های طبقه‌بندی ترجیح داده شده است. این الگوریتم امکان تفکیک ادعاها را به دو دسته معتبر و تقلبی با دقت بهتر را فراهم می‌کند، که به بهبود تعمیم‌پذیری مدل هوشمند تشخیص تقلب بیمه درمان کمک می‌کند. به عنوان یکی از مزایای قابل توجه دیگر، قادر به شناسایی نقاط بیشتری در لبه‌ها هست که احتمال وجود ناهنجاری در آن‌ها بالا است. از سوی دیگر، یکی از ضعف‌های الگوریتم IF وابستگی شدید آن به در اختیار داشتن اطلاعات زمینه‌ای در خصوص نرخ آلودگی است. روش پیشنهادی نسبت به الگوریتم‌های دیگر هم زمان محاسبه را کاهش و هم عملکرد پیش‌بینی را بهبود می‌دهد و ضعف IF در نیاز به اطلاعات پس زمینه مانند نرخ آلودگی برای مجموعه داده را نیز برطرف می‌کند. برای حل این مشکل، در الگوریتم K-IF پیشنهادی ابتدا مقدار این پارامتر در الگوریتم IF دو برابر مقدار برآوردی در نظر گرفته می‌شود تا بتوان نمونه‌های مشکوک را شناسایی کرد. سپس بکمک الگوریتم K-Means با خوشه‌بندی نمونه‌های مشکوک، نمونه‌های تقلب مشخص می‌شود. نتایج نشان می‌دهد که الگوریتم K-IF می‌تواند نقاط بیشتری در لبه‌ها را نسبت به الگوریتم IF و AE به عنوان نقاط تقلب پیدا کند. علاوه براین، بررسی عملکرد الگوریتم پیشنهادی بر روی سه مجموعه داده واقعی برچسب‌دار نشان‌دهنده برتری عملکرد مدل K-IF نسبت به سایر الگوریتم‌ها است؛ نتایج نشان می‌دهد عملکرد خوب الگوریتم طراحی شده بدلیل قدرت الگوریتم IF در تفکیک نمونه‌ها، تنظیم دقیق پارامترهای این الگوریتم با استفاده از روش پیشنهادی برای تعیین مقدار نرخ آلودگی در الگوریتم IF و استفاده از نمره سیلوئیت برای تعیین تعداد خوشه بهینه در الگوریتم K-Means است. این نتایج نسبت به نتایج به‌دست‌آمده از الگوریتم‌های دیگر بهتر بود. به طور کلی، این مطالعه پتانسیل قابل توجه الگوریتم‌های هوش مصنوعی را در شناسایی تقلب در سیستم‌های بیمه درمانی مورد تأکید قرار داده و با طراحی و پیاده‌سازی یک چارچوب هوشمند و ماژولار، گامی مؤثر در جهت کشف این نوع تخلفات برداشته است.

از سوی دیگر، نتایج نشان می‌دهد که طبقه‌بندی خوب الگوریتم‌های مختلف در مجموعه داده اول و دوم نسبت به مجموعه داده سوم به دلیل انتخاب دقیق ویژگی‌ها برای آموزش و توسعه مدل به دست آمده است. به طوری که نتایج آزمایش مجموعه داده‌های مختلف در الگوریتم‌های کشف تقلب مختلف بسته به کیفیت مجموعه ویژگی‌های تعریف شده جهت کشف تقلب تغییر می‌کند. بنابراین استخراج ویژگی‌های مناسب جهت کشف تقلب بر روی مجموعه داده تأثیر قابل توجهی بر نتایج طبقه‌بندی دارد. همچنین، هم‌افزایی بین داده‌های جمعیتی، مالی و خدماتی می‌تواند حساسیت مدل را نسبت به رفتارهای تقلبی که در یک جنبه ظاهر می‌شود، افزایش دهد. در این راستا، یکی از محدودیت‌های اساسی توسعه نرم‌افزار مبتنی بر چارچوب پیشنهادی برای شناسایی تقلب با دقت بالا، کمبود داده‌های باکیفیت از لحاظ جمعیتی، مالی و خدماتی است. در این راستا، لازم است اطلاعات متنوعی درباره بازیگران مختلف در هر تراکنش درمانی ثبت شود، از جمله نوع بیماری، تشخیص، تخصص پزشکی، داروهای تجویز شده به تفکیک، قیمت‌های کالاها و خدمات در رکوردها و تراکنش‌های مربوط به هر

نسخه پزشکی. به‌عنوان مثال، تنها ثبت «نوع خدمت» به‌همراه قیمت یا توصیف کلی و بدون ذکر ویزیت خاص یا نام دارو، به شدت می‌تواند امکان شناسایی رفتارهای ناهنجار را کاهش دهد و باعث کاهش یادگیری الگوهای تقلب شود؛ بنابراین داده‌های دقیق و با سطح جزئیات مناسب، همراه با استانداردهای مربوط به حریم خصوصی و انسجام ثبت، برای آموزش مدل‌های تشخیص تقلب ضروری است. با توجه به داده‌های موجود، ما الگوی آماری غیرعادی را از الگوی افزایش فعالیت بازیگران، هزینه صرف‌شده برای یک خدمت، تطابق جنسیت و سن با تشخیص، تطابق سن و جنسیت با خدمات، تطابق خدمت با تخصص ارائه دهندگان، و ویژگی‌های دیگر شناسایی کردیم.

نهایتا، چارچوب پیشنهادی توسعه یافته به صورت یک بسته نرافزاری به‌عنوان بخشی از خدمات مدیریت ادعای بیمه درمان که برای شرکت‌های بیمه خصوصی توسعه و عرضه می‌شود. این ترکیب قدرتمند از مدل‌های یادگیری ماشین، معماری خوب و رابط کاربری مدرن، قابلیت گسترش‌پذیری بالا و پاسخگویی سریع به نیازهای سازمانی را فراهم می‌آورد و می‌تواند به عنوان یک راهکار جامع برای کشف تقلب در حوزه بیمه درمان به کار گرفته شود. با توجه به قابلیت‌های فنی و تحلیلی چارچوب پیشنهادی، پیشنهاد می‌شود مدیران شرکت‌های بیمه بسته نرم‌افزاری طراحی شده را به‌عنوان بخشی از راهبرد تحول دیجیتال در مدیریت ادعاهای درمانی به کار گیرند. بهره‌گیری از الگوریتم‌های یادگیری ماشین مانند K-IF، همراه با معماری ماژولار و رابط کاربری مدرن، امکان شناسایی دقیق و سریع تقلب‌های بیمه‌ای را فراهم کرده و ریسک‌های مالی ناشی از ادعاهای غیرواقعی را به‌طور مؤثر کاهش می‌دهد. همچنین، استفاده از پایگاه داده امن و ابزارهای تجسم تعاملی، زمینه‌ساز تصمیم‌گیری آگاهانه و مبتنی بر داده برای مدیران ارشد خواهد بود. از منظر مدیریتی، توصیه می‌شود این سیستم به‌صورت یکپارچه با سامانه‌های موجود در شرکت ادغام شود و در قالب یک مرکز تحلیل تقلب، زیر نظر واحد فناوری اطلاعات و با همکاری واحدهای حقوقی و پزشکی به کار گرفته شود. مستندسازی کامل، رعایت استانداردهای امنیتی و قابلیت گسترش‌پذیری بالا، این نرم‌افزار را به یک سرمایه راهبردی برای توسعه پایدار و افزایش اعتماد مشتریان تبدیل می‌کند. اجرای این راهکار می‌تواند به بهبود کارایی عملیاتی، ارتقاء شفافیت فرآیندها و کاهش هزینه‌های ناشی از تقلب در صنعت بیمه درمان منجر شود.

تحلیل دقیق فرایندهای چارچوب ماژولار پیشنهادی، به‌ویژه جزئیات فرایند بازنگری نظر کارشناسان – مثلاً انواع تقلب و ویژگی‌ها چه هستند و چگونه بر فرایند تصمیم‌گیری تأثیر می‌گذارند، آیا صرفاً یک فرایند تجربی است یا یک فرایند یادگیری ساخت‌یافته یا نسخه‌ای ترکیبی به کار گرفته شده است یا خیر- برای کاربردهای آتی ارزشمند خواهد بود. از سوی دیگر، تحقیقات آینده باید علاوه بر استفاده PCA به کاربرد تکنیک‌های بیشتری از مهندسی ویژگی مانند تحلیل آماری و الگوریتم‌های فراابتکاری مانند الگوریتم ژنتیک بپردازند تا ویژگی اثرگذارتری در کشف تقلب انتخاب شود. در کارهای آینده، تمرکز محققان می تواند بر توسعه الگوریتم‌های مختلف برای تشخیص نابهنجاری باشد، زیرا این کار می‌تواند به کاهش میزان داده‌های برچسب‌خورده مورد نیاز برای آموزش الگوریتم‌های تشخیص تقلب کمک کند و هم‌زمان دقت بالای تشخیص تقلب را حفظ کند. از جمله این پیشنهادها، توسعه مدل‌های پیش‌بینی بیشتری مانند یادگیری عمیق است تا بتوانند روی مجموعه‌های داده بیمه درمان اعمال شده و یافته‌های قوی‌تری ارائه دهند. همچنین، استفاده از ترکیب الگوریتم‌های مبتنی بر گراف با روش‌های یادگیری ماشین می‌تواند به شناسایی بهتر تقلب‌های گروهی و پیچیده بینجامد. در آینده، کاربرد مدل‌ها و فناوری‌های استخراج ویژگی‌های جدیدتر جهت کاهش زمان محاسبه مقدار ویژگی‌های هر نمونه جدید نیز می‌تواند برای اثربخشی آن بر داده‌های دیده نشده بررسی شود.

تعارض منافع. برای ارائه مطالب و نگارش این مقاله هیچ‌گونه کمک مالی از هیچ فرد، نهاد و سازمانی دریافت نشده است و نتایج و دستاوردهای این مقاله به نفع یا ضرر سازمان یا فردی خاص نخواهد بود. حضور نویسندگان در این پژوهش به‌عنوان شاهدی بی‌طرف ولی متخصص بوده است و نویسندگان هیچ‌گونه تعارض منافعی ندارند.

منابع

1. AbuAlghanam, O., Alazzam, H., Alhenawi, E. A., Qatawneh, M., & Adwan, O. (2023). Fusion-based anomaly detection system using modified isolation forest for internet of things. *Journal of Ambient Intelligence and Humanized Computing*, *14*(1), 131-145.

2. Bauder, R., Khoshgoftaar, T. M., & Seliya, N. (2017). A survey on the state of healthcare upcoding fraud analysis and detection. *Health Services and Outcomes Research Methodology*, *17*, 31-55.
3. Busch, R. S. (2012). *Healthcare fraud: auditing and detection guide*. John Wiley & Sons.
4. Chandralekha, E., Vinodhini, S., Kandasamy, V., & Rama, P. (2025). Heart Rate Anomaly Detection in Healthcare Using Elliptic Envelope and Local Forest. *Procedia Computer Science*, *258*, 1677-1687.
5. Debener, J., Heinke, V., & Kriebel, J. (2023). Detecting insurance fraud using supervised and unsupervised machine learning. *Journal of Risk and Insurance*, *90*(3), 743-768.
6. De Meulemeester, H., De Smet, F., van Dorst, J., Derroitte, E., & De Moor, B. (2025). Explainable unsupervised anomaly detection for healthcare insurance data. *BMC Medical Informatics and Decision Making*, *25*(1), 14.
7. Dharmadhikari S., (2025) Insurance Claims Management Market Report 2025 (Global Edition). Site: https://www.cognitivemarketresearch.com/insurance-claims-management-market-report.
8. du Preez, A., Bhattacharya, S., Beling, P., & Bowen, E. (2025). Fraud detection in healthcare claims using machine learning: A systematic review. *Artificial Intelligence in Medicine*, *160*, 103061.
9. Ding, Z., & Fei, M. (2013). An anomaly detection approach based on isolation forest algorithm for streaming data using sliding window. *IFAC Proceedings Volumes*, *46*(20), 12-17.
10. Ding, K., Zhou, Q., Tong, H., & Liu, H. (2021, April). Few-shot network anomaly detection via cross-network meta-learning. In *Proceedings of the web conference 2021* (pp. 2448-2456).
11. Ester, M., Kriegel, H. P., Sander, J., & Xu, X. (1996, August). A density-based algorithm for discovering clusters in large spatial databases with noise. In *kdd* (Vol. 96, No. 34, pp. 226-231).
12. Frigge, M., Hoaglin, D. C., & Iglewicz, B. (1989). Some implementations of the boxplot. *The American Statistician*, *43*(1), 50-54.
13. Feng, Y., Cai, W., Yue, H., Xu, J., Lin, Y., Chen, J., & Hu, Z. (2022). An improved X-means and isolation forest based methodology for network traffic anomaly detection. *Plos one*, *17*(1), e0263423.
14. Hamid, Z., Khalique, F., Mahmood, S., Daud, A., Bukhari, A., & Alshemaimri, B. (2024). Healthcare insurance fraud detection using data mining. *BMC Medical Informatics and Decision Making*, *24*(1), 112.
15. Johnson, M. E., & Nagarur, N. (2016). Multi-stage methodology to detect health insurance claim fraud. *Health care management science*, *19*, 249-260.
16. Jones, P. J., James, M. K., Davies, M. J., Khunti, K., Catt, M., Yates, T., ... & Mirkes, E. M. (2020). FilterK: A new outlier detection method for k-means clustering of physical activity. *Journal of biomedical informatics*, *104*, 103397.
17. Jafarnejad Chaghoshi, A., Khani, A. M., & Rezasoltani, A. (2024). Risk Modeling in Banking Services for the Blind Using Fuzzy FMEA and Graph Neural Network (GNN). *Journal of Industrial Management Perspective, 14*(4), 223-255. *(In Persian).*
18. Kose, I. (2020). An Ontology-Based Medical Information Management System for Electronic Claim Processing Systems.
19. Kose, I., Gokturk, M., & Kilic, K. (2015). An interactive machine-learning-based electronic fraud and abuse detection system in healthcare insurance. *Applied Soft Computing*, *36*, 283-299.
20. Kumar, M., Ghani, R., & Mei, Z. S. (2010, July). Data mining to predict and prevent errors in health insurance claims processing. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 65-74).
21. Keshvari Kamran, J., Keramati, M., Toloie Eshlaghy, A., & Amin Mousavi, S. A. (2024). Presenting the Elements and Reinforcement Learning Methodology of Hospital Accreditation Based on the Agent-Based Conceptual Model. *Journal of Industrial Management Perspective*, 14(1), 238-264. *(In Persian).*
22. Liou, F. M., Tang, Y. C., & Chen, J. Y. (2008). Detecting hospital fraud and claim abuse through diabetic outpatient services. *Health care management science*, *11*, 353-358.
23. Luo, W., & Gallagher, M. (2010, December). Unsupervised DRG upcoding detection in healthcare databases. In *2010 IEEE International Conference on Data Mining Workshops* (pp. 600-605). IEEE.
24. Leung, K., & Leckie, C. (2005, January). Unsupervised anomaly detection in network intrusion detection using clusters. In *Proceedings of the Twenty-eighth Australasian conference on Computer Science-Volume 38* (pp. 333-342).

25. Laskar, M. T. R., Huang, J. X., Smetana, V., Stewart, C., Pouw, K., An, A., ... & Liu, L. (2021). Extending isolation forest for anomaly detection in big data via K-means. *ACM Transactions on Cyber-Physical Systems (TCPS)*, *5*(4), 1-26.
26. Lu, J., Lin, K., Chen, R., Lin, M., Chen, X., & Lu, P. (2023). Health insurance fraud detection by using an attributed heterogeneous information network with a hierarchical attention mechanism. *BMC Medical Informatics and Decision Making*, *23*(1), 62.
27. Leevy, J. L., Salekshahrezaee, Z., & Khoshgoftaar, T. M. (2024, July). A Review of Unsupervised Anomaly Detection Techniques for Health Insurance Fraud. In *2024 IEEE 10th International Conference on Big Data Computing Service and Machine Learning Applications (BigDataService)* (pp. 141-149). IEEE.
28. Li, Y., Yan, C., Liu, W., & Li, M. (2018). A principle component analysis-based random forest with the potential nearest neighbor method for automobile insurance fraud identification. *Applied Soft Computing*, *70*, 1000-1009.
29. Lee, I., & Shin, Y. J. (2020). Machine learning for enterprises: Applications, algorithm selection, and challenges. *Business Horizons*, *63*(2), 157-170.
30. Matloob, I., Khan, S., Rukaiya, R., Alfrahi, H., & Ali Khan, J. (2025). Healthcare fraud detection using adaptive learning and deep learning techniques. *Evolving Systems*, *16*(2), 72.
31. Mohanta, A., & Panigrahi, S. (2023). Health Insurance Fraud Detection Using Feature Selection and Ensemble Machine Learning Techniques. In *Advances in Distributed Computing and Machine Learning: Proceedings of ICADCML 2023* (pp. 197-207). Singapore: Springer Nature Singapore.
32. Massi, M. C., Ieva, F., & Lettieri, E. (2020). Data mining application to healthcare fraud detection: a two-step unsupervised clustering method for outlier detection with administrative databases. *BMC medical informatics and decision making*, *20*(1), 160.
33. Machireddy, J. (2025). Automation in healthcare claims processing: Enhancing efficiency and accuracy.
34. Naidoo, K., & Marivate, V. (2020, April). Unsupervised anomaly detection of healthcare providers using generative adversarial networks. In *Conference on e-Business, e-Services and e-Society* (pp. 419-430). Cham: Springer International Publishing.
35. Ortega, P. A., Figueroa, C. J., & Ruz, G. A. (2006). A Medical Claim Fraud/Abuse Detection System based on Data Mining: A Case Study in Chile. *DMIN*, *6*, 26-29.
36. Pooya, A. R., & Javan Rad, E. (2014). Implementation of Neural Networks in Group Technology and Its Comparison to the Results of K-means, Similarity Coefficient Method and Rank Order Clustering. *Journal of Industrial Management Perspective, 3*(4), 39-62. *(In Persian).*
37. Puggini, L., & McLoone, S. (2018). An enhanced variable selection and Isolation Forest based methodology for anomaly detection with OES data. *Engineering Applications of Artificial Intelligence*, *67*, 126-135.
38. Ripan, R. C., Sarker, I. H., Hossain, S. M. M., Anwar, M. M., Nowrozy, R., Hoque, M. M., & Furhad, M. H. (2021). A data-driven heart disease prediction model through K-means clustering-based anomaly detection. *SN Computer Science*, *2*(2), 112.
39. Samariya, D., Ma, J., Aryal, S., & Zhao, X. (2023). Detection and explanation of anomalies in healthcare data. *Health Information Science and Systems*, *11*(1), 20.
40. Shamitha, S. K., & Ilango, V. (2020, July). A time-efficient model for detecting fraudulent health insurance claims using artificial neural networks. In *2020 International Conference on System, Computation, Automation and Networking (ICSCAN)* (pp. 1-6). IEEE.
41. Sasaki, Y. (2007). The truth of the F-measure. *Teach tutor mater*, *1*(5), 1-5.
42. Sisko, A. M., Keehan, S. P., Poisal, J. A., Cuckler, G. A., Smith, S. D., Madison, A. J., ... & Hardesty, J. C. (2019). National health expenditure projections, 2018–27: economic and demographic trends drive spending and enrollment growth. *Health affairs*, *38*(3), 491-501.
43. Suroor, N., & Misra, T. (2024). Medical Insurance Fraud Detection. In *Deep Learning in Internet of Things for Next Generation Healthcare* (pp. 182-193). Chapman and Hall/CRC.
44. Suesserman, M., Gorny, S., Lasaga, D., Helms, J., Olson, D., Bowen, E., & Bhattacharya, S. (2023). Procedure code overutilization detection from healthcare claims using unsupervised deep learning

methods. *BMC Medical Informatics and Decision Making*, *23*(1), 196.
45. Sun, J., Li, Y., Chen, C., Lee, J., Liu, X., Zhang, Z., ... & Xu, W. (2020, April). FDHelper: assist unsupervised fraud detection experts with interactive feature selection and evaluation. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (pp. 1-12).
46. Spiekermann, D., & Keller, J. (2021). Unsupervised packet-based anomaly detection in virtual networks. *Computer Networks*, *192*, 108017.
47. Vishwakarma, M., & Kesswani, N. (2023). A new two-phase intrusion detection system with Naïve Bayes machine learning for data classification and elliptic envelop method for anomaly detection. *Decision Analytics Journal*, *7*, 100233.
48. Wynia, M. K., Cummins, D. S., VanGeest, J. B., & Wilson, I. B. (2000). Physician manipulation of reimbursement rules for patients: between a rock and a hard place. *Jama*, *283*(14), 1858-1865.
49. Yamanishi, K., Takeuchi, J. I., Williams, G., & Milne, P. (2000, August). On-line unsupervised outlier detection using finite mixtures with discounting learning algorithms. In *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 320-324).
50. Yoo, Y., Shin, J., & Kyeong, S. (2023). Medicare fraud detection using graph analysis: a comparative study of machine learning and graph neural networks. *IEEE Access*, *11*, 88278-88294.