



Summer 2024, 5 (2), 11-21

DOI: [10.71856/impcs.2024.1121950](https://doi.org/10.71856/impcs.2024.1121950)

Received: 16 Apr 2024

Accepted: 8 Jun 2024

مقاله پژوهشی

A MLP Neural Network Approach for Validation of Health Insurance Customers in Big Data Environments

Saeed Shouri^{1*}, Ali Cheshomi², Ahmad Safi³, Rasoul Ramezani⁴, Taghi Ebrahimi Salari⁵

1. PhD. Student, Ferdowsi University of Mashhad, Mashhad, Iran. *Corresponding Author, Email: saeedshouri63@gmail.com
2. Assistant Professor, Ferdowsi University of Mashhad, Mashhad, Iran. a.cheshomi@um.ac.ir
3. Associate Professor, Ferdowsi University of Mashhad, Mashhad, Iran. spring05@um.ac.ir
4. Assistant Professor, The University of Texas at Dallas, Dallas, US. rasoul.ramezani@utdallas.edu
5. Associate Professor, Ferdowsi University of Mashhad, Mashhad, Iran. ebrahimi@um.ac.ir

Abstract

Introduction: Inadequate risk assessment of policyholders is a significant challenge in the health insurance industry. This research proposes a customer validation model for health insurance plans, focusing specifically on employees of the East Iran Oil Company.

Method: In this study, a Multi-Layer Perceptron (MLP) neural network with six steps is employed for customer validation. Weight training, a crucial step in neural network implementation, determines the influence of explanatory variables on the model's output. The trained model is then used for validation.

Results: Validation results indicate that health insurance claim variables and specific disease variables have the highest impact on unhealthy customer classification. Notably, the validation process identified approximately 1.8% of the population as "unhealthy." This seemingly small group accounts for a disproportionately high 17.4% of the company's total health insurance claims, despite currently being classified and charged premiums as healthy individuals.

Discussion: The proposed validation model offers a practical approach for insurance companies to assess customer risk profiles and tailor premiums accordingly. This approach promotes a more equitable and sustainable insurance system.

JEL classification: C52, I13, C45

Keywords: Validation, health insurance, Claims-Based Risk, MLP neural network

دوره پنجم، تابستان ۱۴۰۳
شماره دوم، صص: ۱۱-۲۱

تاریخ دریافت: ۱۴۰۳/۰۱/۲۸
تاریخ پذیرش: ۱۴۰۳/۰۳/۱۹

رویکرد شبکه عصبی MLP برای اعتبارسنجی مشتریان بیمه سلامت در محیط‌های کلان داده

سعید شعوری^{۱*}، علی چشمی^۲، احمد سیفی^۳، رسول رضانی^۴، تقی ابراهیمی سالاری^۵

۱. دانشجوی دکتری، گروه اقتصاد، دانشگاه فردوسی مشهد، مشهد، ایران. (نویسنده مسئول) saeedshouri63@gmail.com
۲. استادیار، گروه اقتصاد، دانشگاه فردوسی مشهد، مشهد، ایران. a.cheshomi@um.ac.ir
۳. دانشیار، گروه اقتصاد، دانشگاه فردوسی مشهد، مشهد، ایران. spring05@um.ac.ir
۴. استادیار، گروه اقتصاد، دانشگاه تگزاس، دالاس، آمریکا. rasoul.ramezani@utdallas.edu
۵. دانشیار، گروه اقتصاد، دانشگاه فردوسی مشهد، مشهد، ایران. ebrahimi@um.ac.ir

چکیده: ارزیابی ناکافی ریسک بیمه‌گذاران یک چالش مهم در صنعت بیمه سلامت است. این تحقیق یک مدل اعتبارسنجی مشتری برای طرح‌های بیمه سلامت با تمرکز ویژه بر کارکنان شرکت نفت شرق ایران پیشنهاد می‌کند. در این مطالعه، یک شبکه عصبی پرسپترون چندلایه (MLP) مشتمل بر شش مرحله برای اعتبارسنجی مشتری استفاده می‌شود. آموزش وزن‌ها به عنوان گامی حیاتی در پیاده‌سازی شبکه عصبی، تأثیر متغیرهای توضیحی را بر خروجی مدل تعیین می‌کند. سپس مدل آموزش‌دیده برای اعتبارسنجی استفاده می‌شود. نتایج اعتبارسنجی نشان می‌دهد که متغیرهای هزینه پرداختی بیمه سلامت و متغیرهای بیماری خاص بیشترین تأثیر را بر طبقه‌بندی مشتریان ناسالم دارند. نتیجه حائز اهمیت پس از اعتبارسنجی بیانگر این است که حدود ۱,۸٪ از جمعیت به‌عنوان "ناسالم" شناسایی شده است. این گروه به‌ظاهر کوچک، ۱۷,۴۷ درصد کل پرداختی شرکت بیمه بابت مطالبات بیمه درمان را تشکیل می‌دهند این در حالی است که اکنون حق بیمه این افراد برابر افراد سالم می‌باشد. مدل اعتبارسنجی پیشنهادی یک رویکرد عملی برای شرکت‌های بیمه ارائه می‌کند تا پروفایل ریسک مشتری را ارزیابی کرده تا حق بیمه‌ها را بر این اساس تنظیم کنند. این رویکرد سیستم بیمه‌ای عادلانه‌تر و پایدارتر را ترویج می‌کند.

طبقه‌بندی JEL: C52, I13, C45

واژه‌های کلیدی: اعتبارسنجی، بیمه سلامت، ریسک مبتنی بر خسارت، شبکه عصبی MLP

۱. مقدمه

از طریق ارزیابی مبنی بر ویژگی‌های افراد بیمه‌شده و با استفاده از شبکه عصبی پرسپترون چندلایه یا MLP^1 انجام‌دهیم.

مقاله به شرح زیر تنظیم شده‌است: در بخش اول، اعتبارسنجی بیمه سلامت تجزیه و تحلیل شده‌است. در قسمت دوم پیشینه تحقیق بررسی و در قسمت سوم مبانی نظری تحقیق ارائه شده‌است. بخش چهارم به معرفی داده‌های تحقیق می‌پردازد و بخش پنجم به جزئیات طراحی و مدل‌سازی فرآیند اعتبارسنجی می‌پردازد. بخش پایانی یافته‌ها را خلاصه می‌کند و توصیه‌هایی نیز ارائه می‌دهد.

۲. پیشینه تحقیق

مطالعات زیادی در مورد اعتبارسنجی در حوزه‌های مختلف انجام شده‌است که در این بخش مطالعات انجام شده در زمینه اعتبارسنجی، طبقه‌بندی و شناسایی عوامل ریسک مربوط به بخش بیمه سلامت اشاره می‌شود: الکساندر بهرنت و همکاران [۳] (20۲۰) به بررسی مبانی اعتبارسنجی مطالبات بیمه سلامت می‌پردازند. آن‌ها دو رویکرد را برای اعتبارسنجی بیمه سلامت معرفی می‌کنند. رویکرد اول اعتبارسنجی مبتنی بر مدل و رویکرد دوم اعتبارسنجی مبتنی بر طبقه‌بندی می‌باشد. پیش فرض اصلی رویکرد اول این است که منظور از اعتبار داده‌ها نیستند بلکه ویژگی و تفسیر داده‌ها و پیامدهای تحلیل‌هایی است که روی داده‌ها انجام می‌شود. فرض اصلی در رویکرد دوم این است که نتایج روش‌های توصیفی یا پیچیده در تحقیقات مراقبت‌های بهداشتی معمولاً بر روی زیرگروه‌های قابل مقایسه متمرکز می‌شوند. داده‌ها با زیرگروه مربوطه در داده‌های مطالبات بیمه سلامت از نظر بیماری‌های همراه مرتبط، پیامدهای اندازه‌گیری شده قابل مقایسه هستند.

دیان، جورج [۶] (2014) طبقه‌بندی ریسک به استفاده از ویژگی‌های قابل مشاهده و عینی که توسط بیمه‌گران برای گروه‌بندی متقاضیان بیمه با خسارت‌های مورد انتظار مشابه و محاسبه حق بیمه مربوطه و در نتیجه کاهش اطلاعات نامتقارن اشاره دارد. در این مطالعه از یک مدل غربالگری متعارف قرارداد بیمه استفاده شده تا این مبادلات را در طیف وسیعی از محیط‌های اطلاعاتی مطالعه شود. مرینر، وندي [۷] (2013) در تحقیقی به نقش بیمه در تعریف مسئولیت ریسک درمان و هزینه‌های آن می‌پردازد که بر اساس بخشی از نتایج این تحقیق بیماری‌ها، وضعیت بهداشت و سلامت، سن و جنسیت افراد از جمله ریسک‌هایی هستند که بر حق بیمه و تعهدات پوشش‌های مختلف بیمه درمان تاثیر دارد.

با بررسی مطالعات در زمینه بیمه سلامت و درمان در ایران مشاهده شد که این مطالعات در سه سطح کلی قابل طبقه‌بندی هستند. گروه اول مطالعات در سطح اقتصاد کلان، گروه دوم مطالعات در سطح اقتصاد خرد و گروه سوم مطالعات تطبیقی هستند.

هدف بیمه ارائه ابزاری مقرون به صرفه برای محافظت در برابر ضررهای مالی بالقوه بزرگ برای بیمه‌شده، خانواده او، اشخاص ثالث یا جامعه به عنوان یک کل (مانند هزینه‌های پزشکی و کاهش بار بر دوش بیمه‌گذار) است. بنابراین، بیمه عموماً از نظر اجتماعی پدیده‌ای مطلوب است [۱]. از طرفی یکی از مهمترین شاخه‌های صنعت بیمه، بیمه درمان است که در میان سایر موضوعات بیمه‌ای از اهمیت ویژه‌ای برخوردار است و این اهمیت از آنجاست که ارتباط مستقیم با سلامت و رفاه جامعه دارد. از طرف دیگر تلاش‌های مداوم علیه سوءاستفاده و تقلب پزشکی، شامل اقداماتی برای کاهش استفاده نامناسب از مراقبت‌های بهداشتی است که توسط پرداخت‌کنندگان دیگر تأمین می‌شود که این فرآیند پرهزینه است [۲]. اما تعیین حق بیمه درمان بر اساس ارزیابی صحیح ریسک می‌تواند به کاهش سوءاستفاده شرکت‌های بیمه منجر شود. اخیراً مطالبات بیمه سلامت به عنوان یکی از واقعیت‌های جامعه در تحقیقات مراقبت‌های بهداشتی و بهبود کیفیت، دوباره مورد توجه جامعه علمی قرار گرفته‌است. تا کنون، مطالعات بسیار کمی در دسترس است که اعتبار مطالبات بیمه سلامت را مورد اعتبارسنجی قرار داده باشد [۳].

یکی از مهم‌ترین اهداف و دغدغه‌های تمام شرکت‌های بیمه، پیدا کردن رابطه درست و منطقی بین حق بیمه پرداختی مشتریان و هزینه واقعی ادعای خسارات افراد بیمه شده‌است که علاوه بر حفظ منافع شرکت‌های بیمه، برای مشتریان بیمه نیز سودمند است. شرکت‌ها با اعلام نرخ واقعی حق بیمه در قبال هر فرد، می‌توانند در رقابتی درست با سایر شرکت‌های بیمه، سهمی در بازار داشته باشند و در مقابل، متقاضیان بیمه سلامت نیز به تناسب ریسکی که ایجاد می‌نمایند هزینه پرداخت می‌کنند و در صورتی که فردی کم‌ریسک باشد هزینه گزاف پرداخت نخواهد کرد.

اعتبارسنجی یک رویکرد علمی است که با استفاده از اطلاعات وضعیت جاری و سابقه متقاضیان بیمه، تجزیه و تحلیل می‌گردد و امتیاز اعتباری بر مبنای میزان ریسک سلامتی، امتیازدهی و طبقه‌بندی می‌شود. لازم به ذکر است که اعتبارسنجی مدل نه‌تنها دقت یک مدل محاسباتی را ارزیابی می‌کند، بلکه کمک می‌کند تا فرآیند مدل بر اساس نتایج اعتبارسنجی، بهبود یابد [۴]. به عبارتی، اعتبارسنجی چیزی جز سازماندهی مناسب نیست و مهمترین ویژگی در این تعریف درجه اطمینان بالا، فرآیند خاص، ثابت و مشخصات از پیش تعیین شده است [۵]. امتیاز بیمه بر اساس نتایج اعتبارسنجی افراد است و بر حق بیمه‌ای که برای برخورداری از خدمات بیمه سلامت پرداخت می‌کنند، تأثیر می‌گذارد. امتیاز پایین نشان‌دهنده ریسک بالاتر افراد برای شرکت‌های بیمه است بنابراین منجر به تقاضای حق بیمه بالاتر خواهد شد. امتیاز بالا به دلیل ریسک کمتر، باعث کاهش حق بیمه می‌شود. به‌طورکلی در این مطالعه سعی داریم اعتبارسنجی مشتریان بیمه سلامت را برای کارکنان شرکت نفت شرق ایران

بیشتر مراقبت‌های پزشکی و به اصطلاح آزمایش‌های عمل‌گرایانه فرار گرفته- است [18]. یکی از مهمترین معیارهای اعتبارسنجی، تعیین درجه ریسک مشتریان می‌باشد. در واقع، یک ریسک را می‌توان به‌عنوان یک عدد تصادفی X تعریف کرد که نتیجه (یا تحقق) واقعی آن ناشناخته است [19]. افرادی که به مجموعه‌ای از بیمه‌نامه‌های بیماری تعلق دارند، جمعیت نسبتاً ناهمگونی را تشکیل می‌دهند، به‌ویژه با توجه به خطرات مرتبط با سلامت. ارمانو پیکاتو [20] (۲۰۱۴) در کتاب بیمه سلامت عوامل ریسک بیمه سلامت را می‌توان به صورت زیر طبقه‌بندی کرد:

✓ عوامل ریسک عینی^۲ و ویژگی‌های جسمانی بیمه‌شده به ویژه:

سن، جنسیت، سوابق سلامت، شغل می‌باشد.

✓ در میان عوامل ریسک ذهنی^۳، نگرش شخصی نسبت به

سلامت را تشخیص می‌دهیم که تقاضای فردی برای درمان‌های

پزشکی و در نتیجه درخواست مزایای بیمه را تعیین می‌کند.

در اقتصادهای پیشرفته ریسک مانند یک کالا در بازارهای مالی، خرید و فروش می‌شوند. شرکت‌های بیمه در قبال دریافت مبلغی تحت عنوان "حق بیمه" ریسک یک فعالیت را پوشش می‌دهند. متغیر ریسک را با X نشان می‌دهیم که غیرمنفی است و کلیه خسارت‌های احتمالی که می‌تواند متوجه شرکت بیمه شود را شامل می‌شود و حق بیمه را تابعی از X و به صورت $\pi(X)$ نشان می‌دهیم. X یک متغیر تصادفی است که با RV نشان- داده می‌شود. یک RV نتیجه جمع تعداد محدود n تا از RV s است که $X = S_n = \sum_{i=1}^n X_i$ یک دنباله از RV s مستقل و غیرمنفی و دارای توزیع یکسان است که در آن X_i مطالبات آم را نشان می‌دهد و در نهایت X_i دارای توزیع پواسن می‌باشد:

$$X = S_n = \sum_{i=1}^n X_i, N \sim Poisson(\lambda), \lambda > 0 \quad (1)$$

که S جمع X_i ها است طوری که X_i در N زمان تصادفی رخ داده است و X_i ها مستقل از N هستند بدین معنی که شمار مطالبات بیمه رخ داده و مقدار پولی که مرتبط نیستند.

پترو لیما رومز [۲۱] (۲۰۱۷) اصل مطلوبیت مورد انتظار برای بیمه‌گر را بیان می‌کند و تصریح می‌کند که یک بیمه‌گر با مطلوبیت U و سرمایه W باید قراردادی را در برابر ریسک X و با حق بیمه $\pi(X)$ بپذیرد اگر و فقط اگر:

$$E[u(w + \pi(x) - x)] \geq u(w). \quad (2)$$

مطلوبیتی که شرکت بیمه به سرمایه اولیه خود نسبت می‌دهد باید کمتر یا مساوی با ارزش مورد انتظار مبالغ پولی مختلف ممکن در آینده باشد. تعریف ۱: طبق مطلوبیت مورد انتظار حداقل ارزش $\pi(X)$ که یک بیمه‌گر با سرمایه اولیه w باید هزینه کند برای پوشش ریسک X حل معادله زیر است:

در مطالعات اخیر گروه اول هادیان و همکاران [۸] (۲۰۰۷)، دقیقی و همکاران [۹] (۲۰۱۰) با استفاده از روش پنل دیتا، بیان کردند درآمد سرانه، مخارج بهداشتی انتظاری و نرخ تورم مهمترین عوامل تعیین‌کننده تابع تقاضای بیمه سلامت هستند. عباسی و تقی‌آبادی [۱۰] (۲۰۱۱) نشان- دادند که یک ارتباط بلندمدتی بین درآمد سرانه و نرخ باسوادگی با تقاضای بیمه مکمل وجود دارد اما قیمت بیمه سلامت، نرخ تورم و بیکاری تأثیر معنی‌داری ندارند.

در مطالعات گروه دوم، نخعی و کامویی [۱۱] (۲۰۱۰) ابراهیم‌زاده و همکاران [۱۲] (۲۰۱۵) نصرت‌نژاد و همکاران [۱۳] (۲۰۱۴) با رویکرد اقتصاد خرد نشان دادند درآمد خانوار، سن و تحصیلات سرپرست خانوار، هزینه‌های درمانی خانوار عوامل اصلی مؤثر بر تقاضای بیمه درمانند.

از مطالعات اخیر گروه سوم نیز می‌توان به مطالعات وفایی‌نجر و همکاران [۱۴] (۲۰۰۷) گل‌علیزاده و همکاران [۱۵] (۲۰۱۸) موهبتی [۱۶] (۲۰۱۸) اشاره کرد که با بررسی بیمه‌های مکمل در دنیا و ایران الگو و راهکارهایی را جهت بهبود وضعیت بیمه مکمل ایران پیشنهاد کردند. از جمله مطالعه داخلی در زمینه صنعت بیمه و بیمه سلامت در حوزه ریسک صنعت بیمه صورت‌گرفته به شرح زیر می‌باشد:

ریاحی‌فر [۱۷] (۱۳۸۵) در تحقیق خود به دنبال بررسی الگوی مدیریت ریسک بیمه‌ای درمان کشور می‌باشد که برای برآورد مدل از تکنیک دلفی سه مرحله‌ای استفاده شده است. براساس نتایج، اتخاذ یک استاندارد مدیریت ریسک برای طراحی الگوی فرآیند مدیریت ریسک در هر سازمان الزامی است که بر اساس این الگو، عناصر فرآیند مدیریت ریسک، به صورتی نظم‌یافته مشخص می‌شوند. این الگو شامل زمینه‌سازی محیط داخلی، زمینه‌سازی محیط خارجی، زمینه‌سازی اجرای مدیریت ریسک، زمینه‌سازی نحوه اجرا، زمینه‌سازی تامین منابع مالی و بودجه برنامه‌ای، زمینه‌سازی تامین تخصص‌های مدیریت ریسک، تعیین وظایف کمیته مدیریت ریسک، شناسایی و تحلیل ریسک، تنظیم شاخص‌ها و معیارهای ارائه شده برای تعیین حق بیمه و تعیین سیستم مناسب پرداخت سازمان‌های بیمه‌گر به عرضه‌کنندگان خدمات می‌باشد.

با توجه به مطالعات مرتبط، این مطالعه از دو جهت به ادبیات موجود کمک می‌کند. ابتدا جامعه آماری مورد بررسی را از لحاظ ویژگی افراد بیمه- شده مورد تجزیه و تحلیل قرار می‌دهد. در مرحله دوم براساس این ویژگی‌ها، اعتبارسنجی با استفاده از هوش مصنوعی صورت می‌گیرد و افراد جامعه مورد بررسی به افراد سالم و ناسالم طبقه بندی می‌شوند.

۳. مبانی نظری

همان‌طور که اشاره شد اخیراً مطالبات بیمه سلامت دوباره مورد توجه جامعه علمی به عنوان منبعی از شواهد دنیای واقعی در جهت، بهبود و توسعه

$$u(w) = E[u(w + \pi[X]^- - x)]. \quad (3)$$

هنگامی که تابع مطلوبیت مقعر باشد این اصل، اصل مطلوبیت صفر نامیده می‌شود. فرض کنیم بیمه‌گر تابع مطلوبیت $u(x) = -\alpha e^{-\alpha x}$ و $\alpha > 0$ را دارد و هدف آن محاسبه حداقل حق بیمه طبق نظریه مطلوبیت مورد نظر $\pi[x]$ است که شرکت بیمه باید ریسک X را بپذیرد که محاسبه کرد، یعنی تابعی که به هر t ، در بازه $[-a, a]$ ، $a \in R$ نسبت می‌دهد، مقدار تابع گشتاور $M_X(t) = E[e^{xt}]$ تا زمانی که مقدار مورد انتظار محدود باشد.

$$\begin{aligned} M_X(t) &= E[e^{xt}] = \int_0^{+\infty} e^{xt} P(X=x) dx \\ &= \int_0^{+\infty} e^{xt} \frac{\beta^\theta}{\Gamma(\theta)} x^{\theta-1} dx \\ &= \int_0^{+\infty} \frac{\beta^\theta}{\Gamma(\theta)} x^{\theta-1} dx \quad (4) \end{aligned}$$

مقدار انتگرال همگرا می‌شود اگر و فقط اگر $-\beta < 0$.

تابع $\Gamma(\theta)$ که تابعی از θ است که:

$$\Gamma(\theta) = \int_0^{\theta-1} x^{\theta-1} e^{-x} dx.$$

تبدیل $y = (\beta - t)x$ را در نظر بگیرید. بنابراین:

$$\begin{aligned} M_X(t) &= \int_0^{+\infty} \frac{\beta^\theta}{\Gamma(\theta)} \left(\frac{y}{\beta-t}\right)^{\theta-1} e^{-y} \frac{1}{\beta-t} dy \\ &= \frac{\beta^\theta}{\Gamma(\theta)(\beta-t)^\theta} \int_0^{+\infty} y^{\theta-1} e^{-y} dy \\ &= \frac{\beta^\theta}{\Gamma(\theta)(\beta-t)^\theta} \Gamma(\theta) \\ &= \left(\frac{\beta}{\beta-t}\right)^\theta, t < \beta \quad (5) \end{aligned}$$

از طرف دیگر با به‌کارگیری تعریف ۱ داریم:

$$\begin{aligned} u(w) &= E[u(w + \pi[X]^- - X)] \Leftrightarrow -\alpha e^{-\alpha w} \\ &= E[-\alpha e^{-\alpha(w + \pi[X]^- - X)}] \\ &\Leftrightarrow -\alpha e^{-\alpha w} = E[-\alpha e^{-\alpha w} e^{-\alpha \pi[X]^-} e^{\alpha X}] \\ &\Leftrightarrow e^{\alpha \pi[X]^-} = M_X(\alpha) \\ &\Leftrightarrow \pi[X]^- = \frac{1}{\alpha} \ln(M_X(\alpha)) \quad (6) \end{aligned}$$

رابطه 6 را در نظر بگیرید. ارزش حق بیمه با توجه به اصل مطلوبیت مورد انتظار با استفاده از تابع مطلوبیت نهایی به سرمایه اولیه بیمه‌گر بستگی ندارد. در عوض به پارامتر α و تابع توزیع ریسک X بستگی دارد. بنابراین در موارد خاص روابط 5 و 6 ترکیب می‌شود:

$$\pi[X]^- = \frac{\theta}{\alpha} \ln\left(\frac{\beta}{\beta-\alpha}\right), \alpha < \beta \quad (7)$$

به عبارت دیگر، برای پوشش ریسک X ، بیمه‌گر به وسیله اصل مطلوبیت مورد انتظار به حداقل حق بیمه $\pi[X]^- = \frac{\theta}{\alpha} \ln\left(\frac{\beta}{\beta-\alpha}\right)$ ، $\alpha < \beta$ نیاز دارد اگر $\alpha < \beta$.

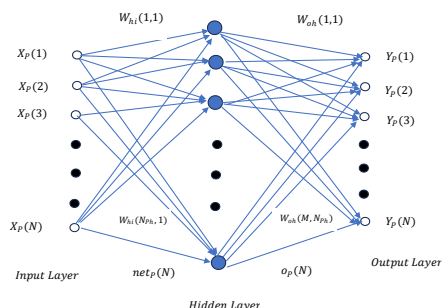
فیلیپ بوچر [22] (۲۰۱۴) در مطالعه‌ای با استفاده از داده‌های مقطعی^۴ به رتبه‌بندی خسارت‌ها در بیمه می‌پردازد. برای مدل‌سازی پارامتری تعداد خسارت‌ها یا ادعاها، مشروط به متغیرهای کمکی، اکچوئر باید توزیع شمارش را انتخاب کند. معمولاً نقطه شروع برای مدل‌سازی داده‌های شمارشی، توزیع پواسون است. پواسون دارای تابع احتمال پواسن به‌صورت زیر است:

$$Pr[N_i = n_i | X_i] = \frac{\lambda_i^{n_i} e^{-\lambda_i}}{n_i!} \quad (8)$$

ویژگی‌های بیمه‌شده که باید بر حق بیمه آن‌ها تأثیرگذار باشد به‌عنوان رگرسیون در پارامتر میانگین توزیع شمارش گنجانده شده است. این اطلاعات برون‌زا را می‌توان با متغیرهای باینری کدگذاری کرد. در بیمه معمولاً یک تابع نمایی برای داشتن $\lambda_i = t_i \exp(x_i' \beta)$ استفاده می‌شود، جایی که t_i نشان‌دهنده در معرض خطر قرارگرفتن فرد بیمه‌شده است. زیرا $E[N_i | X_i] = \lambda_i$ ، با داشتن ویژگی‌های بیمه‌شده، اکچوئر می‌تواند حق بیمه را برای یک بیمه‌شده محاسبه کند. از لحاظ تاریخی، شرکت‌های بیمه داده‌ها را به‌صورت مقطعی جمع‌آوری می‌کنند. به این معنی که پایگاه داده مورد استفاده برای تجزیه و تحلیل نرخ‌گذاری شامل یک لیست قراردادهای بیمه هر مشاهده از پایگاه داده حاوی اطلاعاتی در مورد هر قرارداد جمع‌آوری می‌شود. این اطلاعات شامل سن و جنسیت بیمه‌شونده، وضعیت سلامتی، تعداد دفعات مراجعه به سازمان‌های بهداشت و درمان و نهایتاً مبلغ پرداختی شرکت بیمه می‌باشد.

کوه و همکاران [۲۳] (۲۰۱۱) داده‌کاوی را می‌توان به‌عنوان فرآیند یافتن الگوها و روندهای ناشناخته قبلی در پایگاه‌های داده و استفاده از آن اطلاعات برای ساخت مدل‌های پیش‌بینی تعریف کرد. به‌صورت متناوب، می‌توان آن را به‌عنوان فرآیند انتخاب داده‌ها و کاوش و ساخت مدل‌ها با استفاده از ابزارهای داده وسیع برای کشف الگوهای ناشناخته قبلی تعریف کرد. با پیشرفت‌های تکنولوژی و امکان ذخیره‌سازی حجم عظیم اطلاعات، استفاده از داده‌کاوی به‌عنوان فرایند کشف الگوهای گوناگون در داده‌ها در

آستانه اضافه شده ضرب می‌کند و سپس از آن عبور می‌کند. یک تابع فعال-سازی که ممکن است خطی یا غیرخطی (واحدهای پنهان) باشد. یک شبکه سه‌لایه معمولی در شکل ۱ آمده است. فقط سه لایه MLP در این مقاله در نظر گرفته خواهد شد زیرا نشان داده شده است که این شبکه‌ها هر تابع پیوسته را محاسبه می‌کنند. برای MLP سه لایه واقعی، تمام ورودی‌ها نیز مستقیماً به همه خروجی‌ها متصل می‌شوند. والتر [28] (2005)، شبکه عصبی MLP سه لایه مطابق شکل ۱ مدل‌سازی می‌کنند:



شکل ۱: شبکه عصبی پرسپترون سه لایه

داده‌های آموزشی شامل مجموعه‌ای N_V از الگوهای آموزشی (x_p, t_p) است که p نشان‌دهنده شماره الگو است. در شکل ۱، x_p مربوط به بردار ورودی N بعدی الگوی آموزشی p ام و y_p مربوط به بردار خروجی M بعدی از شبکه آموزش دیده برای الگوی p ام است. برای سهولت علامت‌گذاری و تجزیه و تحلیل، آستانه‌های موجود در واحدهای پنهان و واحدهای خروجی با تخصیص مقدار یک به یک مدیریت می‌شوند. مؤلفه برداری افزوده شده با $x_p(N+1)$ نشان داده شده است. واحدهای خروجی و ورودی خطی هستند. فعال‌سازی‌ها ورودی‌ها را از طریق واحدهای پنهان، با $net_p(j)$ بیان می‌شود.

$$net_p(j) = \sum_{k=1}^{N+1} W_{hi}(j, k) \cdot x_p(k) \quad 1 \leq j \leq N_h \quad (9)$$

با فعال‌سازی خروجی برای P امین الگوی آموزشی، $O_p(j)$ بیان می‌شود:

$$O_p(j) = f(net_p(j)) \quad (10)$$

فعال‌سازی غیرخطی معمولاً به‌عنوان تابع سیگموئیدی انتخاب می‌شود.

$$f(net_p(j)) = \frac{1}{1 + e^{-net_p(j)}} \quad (4)$$

در رابطه (9) و (10)، واحدهای ورودی N با شاخص k نشان داده می‌شوند و $W_{hi}(j, k)$ نشان‌دهنده وزن‌هایی است که k امین واحد ورودی را به j امین واحد پنهان متصل می‌کند. عملکرد کلی MLP با میانگین مربعات خطا (MSE) که به صورت زیر بیان می‌شود.

حال افزایش است و به یک موضوع مهم در حوزه‌های مختلف از جمله صنعت بیمه تبدیل شده است [24]. با توجه به اینکه در این تحقیق از الگوریتم‌های یادگیری ماشین^۵ برای ارزیابی اعتبارسنجی مشتریان بیمه سلامت بهره می‌گیریم به بررسی مبانی نظری می‌پردازیم.

در این مطالعه از شبکه عصبی MLP به‌عنوان یکی از روش‌های داده‌کاوی استفاده شده است. شبکه عصبی MLP یک شبکه معروف عصبی است که به‌نحو بسیار مطلوبی عمل اعتبارسنجی و امتیازدهی را انجام می‌دهد که در قسمت شیوه تجزیه و تحلیل داده‌ها به تفسیر در مورد روش MLP و مراحل اجرایی اعتبارسنجی در این تحقیق با استفاده از روش شبکه عصبی MLP توضیح داده خواهد شد.

۱.۳ روش شبکه عصبی MLP

شبکه عصبی یک سیستم پردازش اطلاعات است که مشابه با شبکه‌های عصبی بیولوژیکی می‌باشد. شبکه‌های عصبی را می‌توان براساس معماری آن به انواع مختلفی طبقه‌بندی کرد. معماری الگوی ارتباط بین نورون‌ها و الگوریتم آموزشی برای تعیین مقدار وزن روی پیوند سیناپسی می‌باشد [25]. چندین شبکه عصبی در چند دهه اخیر توسعه و تحلیل شده‌اند و شبکه‌های عصبی خودسازمانده پرسپترون‌های چندلایه از جمله مهم‌ترین نوع شبکه عصبی می‌باشد. MLPها در طول سال‌ها به‌عنوان یک تکنیک بسیار قدرتمند برای حل طیف گسترده‌ای از مسائل تکامل یافته‌اند. پیشرفت‌های زیادی در بهبود عملکرد و در درک نحوه عملکرد این شبکه‌های عصبی وجود داشته است [26]. با این حال، نیاز به پیشرفت‌های بیشتر در آموزش این شبکه‌ها همچنان وجود دارد.

شبکه‌های عصبی MLP از واحدهایی تشکیل شده‌اند که در لایه‌هایی مرتب شده‌اند. هر لایه از گره‌هایی تشکیل شده است و در شبکه‌های کاملاً متصل در نظر گرفته شده در این مقاله، هر گره به هر گره در لایه‌های بعدی متصل می‌شود. هر MLP از حداقل سه لایه تشکیل شده است که از یک لایه ورودی، یک یا چند لایه پنهان و یک لایه خروجی تشکیل شده است.

لایه ورودی: برای دریافت اطلاعات خامی که وارد شبکه تغذیه شده است. لایه‌های پنهان: عملکرد این لایه‌ها به وسیله ورودی‌ها و وزن ارتباط بین آن‌ها و لایه‌های پنهان تعیین می‌شود. وزن‌ها بین واحدهای ورودی و پنهان تعیین می‌کند که چه وقت یک واحد پنهان باید فعال شود. لایه خروجی: عملکرد واحد خروجی بسته به فعالیت واحد پنهان و وزن ارتباط بین واحد پنهان و خروجی است [27].

گره‌های ورودی دارای توابع فعال‌سازی خطی و بدون آستانه هستند. هر گره واحد پنهان و هر گره خروجی علاوه بر وزن، آستانه‌هایی نیز با خود دارند [28]. گره‌های واحد پنهان دارای توابع فعال‌سازی غیرخطی و خروجی‌ها دارای توابع فعال‌سازی خطی هستند. از این رو، هر سیگنالی که به یک گره در لایه بعدی وارد می‌شود، ورودی اصلی را در یک وزن با یک

۴. مدل‌سازی

این بخش طبقه‌بندی مشتریان بیمه سلامت را بر اساس یک سری ویژگی‌های بیمه‌شده با استفاده از الگوریتم MLP و نرم‌افزار پایتون^۶ به سالم یا ناسالم نشان می‌دهد. در انجام این کار، نویسندگان توضیح می‌دهند که چگونه الگوریتم MLP باید پیاده‌سازی شود. برای شروع، مجموعه‌ای از توابع مورد استفاده در کتابخانه‌های پایتون مورد نیاز بود. برای الگوریتم MLP، بسته‌های مورد استفاده پانداها^۷ برای خواندن داده‌ها از اکسل، بسته‌های آموزشی و آزمایشی برای آموزش و آزمایش داده‌ها، با کمک بسته StandardScaler برای استانداردسازی و با کمک بسته accuracy_score برای ارزیابی دقت بودند. مراحل مدل‌سازی با استفاده از الگوریتم MLP به شرح زیر است:

مرحله ۱) آماده‌سازی داده‌ها: ابتدا داده‌ها را که به صورت زیر در نرم‌افزار پایتون تعریف می‌شود از فایل اکسل بازخوانی می‌کنیم.

$$X = \text{data} [['gender', 'age', 'specific diseases', 'visits', 'cost']] \# \text{Features} \quad (15)$$

$$y = \text{data}['result'] \# \text{Target variable} \quad (16)$$

ویژگی‌های افراد بیمه‌شده که به عنوان متغیرهای توضیحی تعریف می‌شود، شامل موارد زیر می‌شود:

gender: جنسیت که با مقیاس ۰ (جنسیت زن) و ۱ (جنسیت مرد) مشخص می‌شود.

age: سن افراد که بر اساس سال تعیین می‌شود. مثلاً سن با مقیاس از ۱ تا ۱۰۲ سال.

specific diseases: بیمه‌شدگانی که بیماری خاصی دارند با شماره ۱ مشخص می‌شوند.

visit: تعداد مراجعه به سازمان‌های بهداشتی درمانی.

cost: میزان خسارت بیمه یا پرداختی شرکت بیمه بابت بازپرداخت هزینه‌های درمانی (ریال).

متغیر هدف:

result: سالم=۱ یا ناسالم=۰ تعریف می‌شود.

مرحله ۲) تقسیم‌بندی داده‌ها به مجموعه‌های آموزشی و آزمایشی:

در الگوریتم‌های یادگیری ماشینی، داده‌ها به دو دسته آموزش و آزمایش^۸ تقسیم می‌شوند. به طور معمول، ۸۰٪ از کل داده‌ها برای آموزش استفاده می‌شوند، جایی که برای یادگیری و در نهایت پیش‌بینی استفاده می‌شود. ۲۰٪ باقی‌مانده از داده‌ها برای آزمایش و ارزیابی ذخیره شده‌است، که امکان مقایسه پیش‌بینی‌ها در برابر نتایج واقعی به دست آمده از ۸۰٪ داده‌های آموزشی را برای ارزیابی دقت مدل فراهم می‌کند. در این پژوهش جامعه آماری شامل ۲۲۴۹۴ نفر می‌باشد.

$$E = \frac{1}{N_V} \sum_{p=1}^{N_V} E_p = \frac{1}{N_V} \sum_{p=1}^{N_V} \sum_{i=1}^M [t_p(i) - y_p(i)]^2 \quad (11)$$

$$E_p = \sum_{i=1}^M [t_p(i) - y_p(i)]^2 \quad (12)$$

E_p مربوط به خطای الگوی p ام و t_p خروجی مورد نظر برای الگوی p ام است. این رابطه همچنین اجازه می‌دهد تا محاسبه خطای نگاشت واحد خروجی i به صورت رابطه زیر انجام شود.

$$E_p = \sum_{i=1}^M [t_p(i) - y_p(i)]^2 \quad (13)$$

با خروجی N ام برای الگوی آموزشی p ام با رابطه ۱۴ نشان محاسبه می‌شود.

$$y_p(i) = \sum_{k=1}^{N+1} W_{oi}(i, k) \cdot x_p(k) + \sum_{j=1}^N w_{oh}(i, j) \cdot O_p(j) \quad (14)$$

در رابطه (۱۴)، w وزن گره‌های ورودی به گره‌های خروجی و $w_{oh}(i, j)$ وزن گره‌های پنهان به گره‌های خروجی را نشان می‌دهد.

آنچه در یک مدل شبکه عصبی اهمیت دارد، آن است که وزن‌های موجود در شبکه‌های عصبی به روش بهینه‌ای برآورد شوند و بدیهی است که پس از تعیین وزن‌ها به روش بهینه با دادن بردار متغیرهای ورودی به آسانی می‌توان بردار خروجی را برآورد کرد [29] (یانگ و پلات، ۱۹۹۹).

۵. متغیرهای پژوهش

متغیرها عبارتند از عامل یا مفهومی که می‌تواند ارزش‌های عددی گوناگونی را بپذیرد و تغییر ارزش‌های پذیرفته‌شده از سوی آن موجب تغییر مقدار تابع خواهد شد. همان‌طور که گفته شد در این مطالعه اعتبارسنجی بر مبنای عوامل ریسک در حوزه سلامت انجام خواهد گرفت. لذا متغیرهای توضیحی عبارتند از خسارت یا مبلغ پرداختی شرکت بیمه بابت جبران هزینه درمان و بر حسب ریال، سن، جنسیت، تعداد دفعات مراجعه به سازمان‌های خدمات درمانی و بهداشتی و سابقه بیماری‌های خاص می‌باشد. جامعه آماری این پژوهش را ۲۲۴۹۷ نفر از بیمه‌شدگان بیمه سلامت کارکنان شرکت نفت منطقه شمال شرق ایران تشکیل می‌دهند [30]. با توجه به این که هدف از این مطالعه اعتبارسنجی افراد به دو گروه سالم (۱) و ناسالم (۰) با استفاده از شبکه عصبی MLP است، باید یک برچسب برای تشخیص سالم یا ناسالم مشخص شود. برای این منظور از شاخص سرانه هزینه جاری سلامت استفاده می‌کنیم. جاورک [31] (۲۰۲۲) در مطالعه‌ای به بررسی شاخص‌های شاخص مراقبت‌های بهداشتی می‌پردازد و شاخص هزینه‌های جاری سلامت یکی از شاخص‌های معرفی و بررسی شده‌است. شاخص هزینه سرانه سلامت برای ایران و براساس گزارش آوریل ۲۰۲۴ سازمان جهانی بهداشت [32] (برای سال ۲۰۲۱) حدود مبلغ 108,729,938 ریال است.

مرحله ۳) نرمال سازی داده‌ها: پس از تقسیم‌بندی داده‌ها به دو گروه، در مرحله ۳ داده‌ها را نرمال سازی می‌کنیم. با توجه به این که داده‌ها دارای مقیاس‌های متفاوتی هستند (مثلاً جنسیت با مقیاس ۰ و ۱، سن با مقیاس ۱ تا ۱۰۲ سال، پرداخت بیمه به ریال و تعداد مراجعه به صورت یک عدد)، برای مقایسه معنادار همه متغیرها باید در یک مقیاس باشند. بنابراین، از تابع نرمال سازی برای استانداردسازی دو مجموعه داده‌های آموزشی و آزمون به شرح زیر استفاده می‌شود:

$$scaler = StandardScaler() \quad (17)$$

$$X_{train_normalized} = scaler.fit_transform(X_{train}) \quad (18)$$

$$X_{test_normalized} = scaler.transform(X_{test}) \quad (19)$$

مرحله ۴) ایجاد یک مدل MLP طبقه‌بندی شده: هنگامی که MLP داده‌ها را دریافت می‌کند، می‌تواند ویژگی‌های مورد نیاز، تعداد لایه‌ها و میزان تکرارهای مورد نیاز را مشخص کند. الگوریتم MLP به صورت زیر تعریف شده و بر روی داده‌ها اجرایی می‌شود:

$$mlp = MLPClassifier(hidden_layer_sizes = (1, 1), max_iter = 300, random_state = 42) \quad (20)$$

$$mlp.fit(X_{train_normalized}, y_{train}) \quad (21)$$

آموزش وزن‌ها، مهمترین بخش در پیاده‌سازی یک شبکه عصبی می‌باشد. وزن متغیرها که تعیین کننده تأثیر بر متغیر وابسته سالم یا ناسالم است، در این مرحله شناسایی می‌شود. جدول ۱ وزن متغیرهای توضیحی در مدل را نشان می‌دهد.

جدول ۱: وزن متغیرهای مدل

وزن	نام متغیر
0.0009938	جنسیت
0.0360	سن
0.7979	بیماری خاص
0.000168	تعداد مراجعات
0.9477	مطالبات بیمه سلامت

منبع: یافته‌های مطالعه

همان‌طور که نتایج وزن‌ها نشان می‌دهد متغیر هزینه درمان پرداختی شرکت بیمه بیشترین وزن و بیشترین تأثیر را بر وضعیت مستقیم ناسالم بودن افراد دارد. متغیر بعدی که تأثیر منفی بر سلامت افراد دارد، متغیر «بیماری خاص» می‌باشد. باین‌حال، توجه به تأثیر غیرمستقیم متغیر «بیماری خاص» بر افراد ناسالم حائز اهمیت است، زیرا این متغیر مستقیماً بر متغیر «هزینه درمان» تأثیر می‌گذارد که خود تأثیر غیرمستقیم بر

سلامت افراد دارد. سومین متغیری که بر متغیر هدف «سلامت» تأثیر منفی دارد، سن می‌باشد. به طور طبیعی با افزایش سن، بدن دچار تغییراتی می‌شود که می‌تواند منجر به کاهش سلامت شود. سومین و چهارمین عامل موثر بر وضعیت سالم یا ناسالم به ترتیب جنسیت و تعداد مراجعه به سازمان‌های بهداشتی درمانی می‌باشد که تجزیه و تحلیل کلیه این عوامل در مرحله پیش‌بینی و پس از آن که افراد سالم و ناسالم براساس اعتبارسنجی تفکیک گردید، انجام خواهد گرفت.

مرحله ۵) پیش‌بینی: مدل در چهار مرحله گذشته ایجاد شد و مرحله ۵ جایی است که از مدل برای پیش‌بینی استفاده شد و افراد به دو دسته سالم و ناسالم تفکیک می‌شوند.

$$y_pred = mlp.predict(X_{test_normalized}) \quad (22)$$

مرحله ۶) استفاده از مدل: پس از مدل سازی، مدل را با یک مثال مورد آزمون قرار می‌دهیم. برای مثال فردی با ویژگی‌های زیر مورد ارزیابی قرار می‌گیرد: جنسیت: مرد، سن: ۳۹ سال، بیماری خاص ندارد، تعداد مراجعات ۸ بار و هزینه بابت مطالبات بیمه درمان برابر ۲۰۰۰۰۰۰۰ ریال. اجرای مدل به صورت زیر می‌باشد:

```
new_data = pd.DataFrame([[1,39, 0, 8, 20000000]],
columns=['sex', 'age', 'specific diseases ', 'visits','cost'])
```

```
new_data_normalized = scaler.transform(new_data)
```

```
predictions = mlp.predict(new_data_normalized)
```

```
print(predictions)
```

با اجرای مدل، نتیجه مدل ['saalem'] پیش‌بینی می‌شود.

اکنون می‌توان از این مدل برای بهره‌برداری و طبقه‌بندی افراد سالم و ناسالم برای هر جمعیت جدید استفاده کرد. داده‌های جدید تعداد ۲۹۷۸۷ نفر مشتمل بر تعدادی از افراد جدید و تعدادی از افراد همان جامعه آماری قبلی اما برای سال ۱۴۰۲، برای تجزیه و تحلیل وارد مدل می‌شود. نتایج اعتبارسنجی بیانگر این است که حدود ۱,۸ درصد از جمعیت جدید مورد مطالعه را افراد ناسالم تشکیل می‌دهند. نتیجه قابل ذکر این است که ۱,۸ درصد جمعیت ناسالم در کل جمعیت تقریباً ۱۷,۴۷ درصد از کل مطالبات بیمه درمانی پرداخت شده توسط شرکت بیمه را به خود اختصاص می‌دهد که با این واقعیت که شرکت بیمه مبالغ بالایی را برای هزینه‌های درمان افراد ناسالم پرداخت می‌کند، مطابقت دارد.

مرحله ۷) ارزیابی مدل: پس از فرآیند آموزش، مهمترین وظیفه ارزیابی مدل آموزش می‌باشد. در زمینه «طبقه‌بندی» از «مجموعه داده‌ها» برای دستیابی به بالاترین دقت و صحت ممکن در دسته‌بندی و تشخیص طبقات از طریق روش‌های طبقه‌بندی استفاده می‌شود. در برخی مسائل، شناسایی صحیح نمونه‌های مربوط به یک کلاس خاص از اهمیت بیشتری برخوردار است. برای مسائل مربوط به اعتبارسنجی و طبقه‌بندی،

صحت: نسبت موارد مثبت پیش‌بینی‌شده صحیح به کل موارد مثبت پیش‌بینی‌شده را نشان می‌دهد.
 کارایی: نسبت موارد واقعاً مثبت را نشان می‌دهد که به درستی مثبت پیش‌بینی شده‌اند. این پوشش موارد واقعاً مثبت را می‌سنجد.
 با توجه به این که دو معیار دقت و کارایی عکس یکدیگر هستند؛ به این صورت که افزایش یکی باعث کاهش دیگری و بالعکس می‌شود بنابراین یک معیار دیگر به نام معیار F تعریف می‌کنیم. این معیار در واقع میانگین هندسی دو معیار ذکر شده است. پاورز [34] (۲۰۱۱) معیارهای ارزیابی سیستم‌های یادگیری ماشین را مطابق با جدول 7 تعریف می‌کند:

جدول 8: معیارهای ارزیابی مدل

precision = $\frac{TP}{TP+FP}$
RECALL = $\frac{TP}{TP+FN}$
Accuracy = $\frac{TP+TN}{TP+FN+TN}$
F-measure = $\frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$

در این بخش ما در ابتدا ماتریس درهم‌ریختگی را به دست می‌آوریم. برای این امر دستور زیر را اجرا می‌کنیم:

$$cm = \text{confusion}_{matrix}(y_{test}, y_{pred}) \quad (23)$$

نهایتاً ماتریس درهم‌ریختگی به صورت زیر به دست می‌آید:

جدول 9: ماتریس درهم‌ریختگی (منبع: یافته‌های مطالعه)

TP=4409	FN=0
FP=90	TN=0

مثبت واقعی (TP): تعداد نمونه‌های مثبتی که به درستی توسط مدل پیش‌بینی شده‌است. در این مورد، ۴۴۰۹ مثبت واقعی وجود دارد.
 مثبت کاذب (FP): تعداد نمونه‌های منفی که به اشتباه توسط مدل، مثبت پیش‌بینی شده‌بودند. در این مورد، ۹۰ مثبت کاذب وجود دارد.
 منفی واقعی (TN): تعداد نمونه‌های منفی که به درستی توسط مدل پیش‌بینی شده‌است. در این حالت ۰ منفی واقعی وجود دارد.
 منفی کاذب (FN): تعداد نمونه‌های مثبتی که به اشتباه توسط مدل، منفی پیش‌بینی شده‌اند. در این حالت ۰ منفی کاذب وجود دارد. جدول ۱۰ مقادیر معیارهای ارزیابی مدل را محاسبه می‌کند.

جدول 10: مقادیر معیارهای ارزیابی مدل (منبع: یافته‌های مطالعه)

precision = $\frac{TP}{TP+FP} = 0.97995545676817$
RECALL = $\frac{TP}{TP+FN} = 1$
Accuracy = $\frac{TP+TN}{TP+FN+TN} = 0.9799955545676817$
F-measure = $\frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}} = 0.9898967220475977$

معمولاً از چندین معیار استاندارد استفاده می‌شود. ممکن است فرض شود که تنها دو مورد وجود دارد: یا الگوریتم موقعیت را به درستی شناسایی کرده‌است، یا آن را اشتباه شناسایی کرده‌است. با این حال، مشکل به این سادگی نیست. قبل از محاسبه این مقادیر، چهار پارامتر اصلی TP، FP، FN و TN باید محاسبه شود. برای یک مسئله طبقه‌بندی باینری، ماتریس درهم‌ریختگی^۹ یک ماتریس ۲×۲ بود که شامل پارامترهای ذکر شده بود. برای روشن شدن موضوع، فرض کنید مشکل طبقه‌بندی افراد سالم از افراد ناسالم است. برای این منظور، ویویچ [33] (۲۰۲۱) معیارهای استاندارد را برای طبقه‌بندی در جدول ۷ معرفی می‌کند. برای درک جدول ۷، تصور کنید که الگوریتم طبقه‌بندی پس از یادگیری و ایجاد مدل، تحت یک مجموعه آزمون قرار می‌گیرد.

جدول 7: معیار استاندارد کلاس‌بندی (ماتریس درهم‌ریختگی)

		Predicted Class	
		Positive	Negative
Actual Class	Positive	True Positive (TP)	False Negative (FN) TYPE II ERROR
	Negative	False Positive (FP) TYPE I ERROR	True Negative (TN)

برای درک جدول شماره ۷، تصور کنید که الگوریتم طبقه‌بندی پس از آموزش و ایجاد مدل، تحت یک مجموعه آزمون قرار می‌گیرد. در جدول شماره ۷، ردیف‌ها برچسب‌های واقعی را نشان می‌دهند و ستون‌ها پیش‌بینی‌های الگوریتم را نشان می‌دهند. بر این اساس، چهار مورد وجود دارد:

- مثبت واقعی (TP): برخی از افراد ناسالم به درستی ناسالم شناخته می‌شوند.
 - منفی کاذب (FN): برخی از افراد ناسالم به اشتباه سالم شناخته می‌شوند.
 - مثبت کاذب (FP): برخی از افراد سالم به اشتباه ناسالم شناخته می‌شوند.
 - منفی واقعی (TN): افراد سالم به درستی، سالم شناخته می‌شوند.
- در نهایت، هر نمونه داده در یکی از این دو "کلاس" قرار می‌گیرد. بنابراین، برای هر نمونه داده، یکی از چهار سناریوی ذکر شده در بالا ممکن است رخ دهد.

معیارهای ارزیابی:

پاورز [34] (۲۰۱۱) برای ارزیابی میزان کیفیت سیستم خودکار یادگیری ماشین، معیارهای استاندارد صحت^{۱۰}، پوشش^{۱۱}، دقت^{۱۲} و معیار F^{۱۳} را معرفی می‌کند.
 دقت: نسبت موارد واقعاً مثبت به همه پیش‌بینی‌ها را نشان می‌دهد.

جمعیت کم افراد ناسالم در جامعه مورد بررسی که حدود ۲,۸ درصد می‌باشند، هزینه‌شان بالاست و حدود ۲۲,۶ درصد کل مطالبات بیمه را به شرکت بیمه تحمیل می‌کند.

در نهایت بعد از فرآیند آموزش، مدل طراحی شده با معیارهای مختلف از جمله معیارهای استاندارد صحت، دقت، پوشش و معیار اندازه‌گیری F ارزیابی شد که هر کدام، یک سری ویژگی‌های مدل را بررسی کردند. مقادیر این معیارها به ترتیب برابر 0.979، 0.9799، ۱ و 0.989 می‌باشد که این مقادیر بیانگر دقت، صحت و کارایی بسیار بالای مدل است. این نوع مدل اعتبارسنجی یکی از کاربردی‌ترین مدل‌سازی‌های است که شرکت‌های بیمه می‌توانند همه‌ساله برای اعتبارسنجی مشتریان خود انجام دهند.

سیاسگزاری

در پایان از مرکز بهداشت و درمان شرکت نفت شمال شرق ایران که ما را در انجام مطالعه حمایت کردند، به‌ویژه آقایان دکتر مالکی، امدادی‌فر، ناصری و مهندس هدایتی صمیمانه سپاسگزاریم.

مراجع

- [1] European, G. C. A. (2011). *Use of age and disability as rating factors in insurance: why are they used and what would be the implications of restricting their use?*, position paper.
- [2] Feldman, R. (2001). *An economic explanation for fraud and abuse in public medical care programs*. The Journal of Legal Studies, 30(S2), 569-577.
- [3] Behrendt, C. A., Schwaneberg, T., Hischke, S., Müller, T., Petersen, T., Marschall, U., ... & Kriston, L. (2020). *Data privacy compliant validation of health insurance claims data: the IDOMENEO approach*. Das Gesundheitswesen, 82(S 02), S94-S100.
- [4] Lee, G., Kim, W., Oh, H., Youn, B. D., & Kim, N. H. (2019). *Review of statistical model calibration and validation—from the perspective of uncertainty structures*. Structural and Multidisciplinary Optimization, 60, 1619-1644.
- [5] Huber, L. (2023). *Validation of computerized analytical systems*. CRC Press.
- [6] Dionne, G., & Rothschild, C. (2014). *Risk classification and health insurance*. Encyclopedia of Health Economics, 3, 272-280.
- [7] Mariner, Wendy K. (2013). *The affordable care act and Health promotion: The role of insurance in defining of responsibility for health risks and costs*, Boston University school of law public law and Legal Research paper no, pp:10-13.
- [8] Hadian, M., Ghaderi, H., & Moeini, M. (2007). *Estimating Demand Function for Supplementary Health Insurance Case Study: Iran Insurance Corporation*.
- [9] DAGHIGHI, A. A. R., FAGHIH, N. M., & AGHASSI, K. (2010). *Estimating Demand Function of Supplementary Health Insurance in Iranian Insurance Industry*.

طبق جدول 10، معیار دقت (Accuracy) برابر ۰,۹۷۹۹ است یعنی با احتمال ۹۷,۹۷ درصد افراد با استفاده از این اعتبارسنجی وضعیت سالم یا ناسالم بودن افراد را درست پیش‌بینی می‌کند. معیار صحت یک نقطه‌ضعف دارد و آن این است که تفاوتی بین منفی کاذب (false negative) و مثبت کاذب (false Positive) قائل نیست. علی‌رغم نرخ خطای بسیار پایین در این اعتبارسنجی، از ۳ معیار دیگر صحت (precision)، پوشش (RECAL) و معیار F_1 (F-measure) استفاده شد. معیار صحت (precision) برابر ۰,۹۷۹۵ است این مقدار بیانگر این است که اگر الگوریتم شخص را ناسالم تشخیص دهد این تشخیص با احتمال ۹۷,۹ درصد درست است. مقدار معیار پوشش (RECAL) که تمرکز آن بر داده‌هایی است که واقعاً سالم‌اند، ۱ بوده و بیانگر این است که اگر الگوریتم، فردی را سالم تشخیص دهد ۱۰۰ درصد، درست پیش‌بینی کرده‌است. در برخی مسائل از جمله اعتبارسنجی بیمه سلامت، معیار پوشش (Recall) از معیار صحت (Precision) مهمتر است. بدین ترتیب که فرض کنیم شخصی واقعاً ناسالم است ولی به اشتباه سالم تشخیص داده‌شود، در این صورت هزینه بالایی را به شرکت بیمه تحمیل می‌کند به دلیل اینکه فرد ناسالم، حق بیمه فرد سالم را پرداخت می‌کند. معیار F_1 که در واقع ترکیب متعادلی بین معیارهای دقت و صحت است، می‌تواند در مواردی که هزینه False Negative و False Positive متفاوت است به‌کار رود (در این مقاله این مقادیر ۹۰ و ۰ است) لذا استفاده از این معیار پذیرفتنی است. مقدار معیار F_1 برابر ۰,۹۸۹۸ می‌باشد یعنی با احتمال ۹۸,۹۸ درصد پیش‌بینی درست است.

۶. نتیجه‌گیری

در شرایط فعلی، مهمترین مسئله و مشکل فقدان اعتبارسنجی مناسب مشتریان بیمه سلامت می‌باشد، طوری که ارزیابی ریسک بیمه‌گذاران در صنعت بیمه سلامت، فاقد الگوریتم‌ها و سیستم‌های خودکاری است که با حساسیت قابل‌قبولی بتوانند میزان ریسک مشتریان مختلف را بررسی و ارزیابی کنند. لذا در این مطالعه اعتبارسنجی با استفاده از روش شبکه عصبی MLP و برای افراد تحت پوشش بیمه سلامت کارکنان شرکت نفت شرق ایران انجام گرفت. مبنای اعتبارسنجی مدل براساس ارزیابی ریسک مبنی بر خسارت می‌باشد و در بخش ۵، مراحل مختلف مدل‌سازی و ارزیابی مدل به همراه کدنویسی‌های مربوطه به تفصیل ارائه‌گردیده‌است. مهمترین مرحله، آموزش وزن‌ها و سپس مدل اجراست و در نهایت پیش‌بینی انجام می‌شود. دو متغیری که به ترتیب بیشترین تأثیر در انتخاب شدن نتیجه ناسالم یا به عبارتی نتیجه اعتبارسنجی منجر به ناسالم شدن، دارند متغیر هزینه پرداختی بیمه سلامت معادل 0.9477 و متغیر بیماری خاص معادل 0.0797 می‌باشد البته این نکته حائز اهمیت است که در مورد افراد دارای بیماری خاص، غیرمستقیم با افزایش هزینه پرداختی بیمه سلامت، نهایتاً نتیجه ناسالم حاصل می‌شود. نتایج پیش‌بینی نشان می‌دهد علی‌رغم

- [25] Rizal, A., & Hartati, S. (2016). *Recurrent neural network with Extended Kalman Filter for prediction of the number of tourist arrival in Lombok*. In Proceedings of the International Conference on Informatics and Computing (ICIC). Academic Press. doi:10.1109/IAC.2016.7905712.
- [26] Manry, M. T., Chandrasekaran, H., & Hsieh, C. H. (2018). *Signal processing using the multilayer perceptron*. In Handbook of Neural Network Signal Processing (pp. 2-1). CRC Press.
- [27] Balakudehi, j., Tahghighi Sharabyan, M. (2022). *Providing a New Approach to Identify and Detect Credit Card Fraud Using ANN – ICA*. Intelligent multimedia processing and communication systems 3(2), 51-62.
- [28] Delashmit, W. H., & Manry, M. T. (2005, May). *Recent developments in multilayer perceptron neural networks*. In Proceedings of the seventh annual memphis area engineering and science conference, MAESC 7, p. 33.
- [29] Yang, Z. R., Platt, M. B., & Platt, H. D. (1999). *Probabilistic neural networks in bankruptcy prediction*. Journal of business research, 44(2), 67-74.
- [30] The East of Iran Oil Company, *Statistical report of health insurance data*.
- [31] Jaworeck, S. (2022). *A New Approach for Constructing a Health Care Index including the Subjective Level*. International Journal of Environmental Research and Public Health, 19(15), 9686.
- [32] *World Health Organization Global Health Expenditure database* (<http://apps.who.int/nha/database>). The data was retrieved on April 7, 2023.
- [33] Vujović, Ž. (2021). *Classification model evaluation metrics*. International Journal of Advanced Computer Science and Applications, 12(6), 599-606.
- [34] Powers, D. M. (2020). *Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation*. arXiv preprint arXiv:2010.16061.
- [10] Abbasi, E., Taghiabadi, M. (2011). The effect of economic factors on the demand for supplementary medical insurance. *Insurance Journal*, 104(26), 57-80.
- [11] Nakhaei Aghimuni, M., & Kamuei, M. (2010). Estimating demand function of private insurance in urban states of Iran: Tobit analysis. *Insurance Industry*, 25(3), 3-28.
- [12] Ebrahimzadeh, J., Mohammad, A. R. A. B., & Emamgholipour, S. (2015). Determinants of supplementary health insurance demand: case study of Iran. *Iranian journal of public health*, 44(11), 1563-1565.
- [13] Nosratnejad, S. H., Purreza, A., Moieni, M., & Heydari, H. (2014). Factors affecting the demand for private health insurance in Tehran. *Hospital*, 13(2), 37-44.
- [14] Vafae Najar, A., Karimi, I. R. A. J., & Seydnowzadi, M. (2007). A comparative study between complementary health assurance structure and content in selected countries and presenting a paradigm for Iran. *Journal of Health Administration*, 10(28), 57-64.
- [15] Gol-Alizadeh, E., Pirouzian, A., & Jabbari, M. R. (2018). Improving supplemental health insurance in Iran and how to achieve it. *Iran J Health Insur*, 1(1), 2-12.
- [16] Mohebi, T. (2018). Critique of health insurance configuration in developing countries. *Iranian Journal of Health Insurance*, 1(1-2), 1-2.
- [17] Riahifar, M., 2015. Designing a risk management model for medical insurance for Iran. *Insurance industry quarterly*, year 21, number 4, 119-155.
- [18] Choudhry, N. K. (2017). *Randomized, controlled trials in health insurance systems*. New England Journal of Medicine, 377(10), 957-964.
- [19] Olivieri, A., & Pitacco, E. (2015). *Introduction to insurance mathematics: technical and financial features of risk transfers*. Springer.
- [20] Ritacco, E. (2014). *Health insurance. Basic Actuarial Models*, Cham, Switzerland: Springer Verlag.
- [21] Lima Ramos, P. (2017). *Premium calculation in insurance activity*. Journal of Statistics and Management Systems, 20(1), 39-65.
- [22] Boucher, J. P., & Inoussa, R. (2014). *A posteriori ratemaking with panel data*. ASTIN Bulletin: The Journal of the IAA, 44(3), 587-612.
- [23] Koh, H. C., & Tan, G. (2011). *Data mining applications in healthcare*. Journal of healthcare information management, 19(2), 65.
- [24] Manteghipour, Mahnaz, & Rahimkhani, Parisa. (1401). *Designing a hybrid model for classification of imbalanced data in the field of Casualty Insurance*. Intelligent multimedia processing and communication systems, 3(2), ۹-۱۰.

پی نوشت

1. Multi-Layer Perceptron (MLP)
2. Objective risk factors
3. subjective risk factors
4. Cross-Section Data
5. Machine Learning
6. Python
7. Pandas
8. Train and test
9. Confusion
10. precision
11. Recall
12. Accuracy
13. F-measure