سامانه‌های پردازشی و ارتباطی چندرسانه‌ای هوشمند
Intelligent Multimedia Processing and Communication Systems (IMPCS)

Summer 2024, 5 (2), 11-21
DOR:

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 Cheshomi [۲], Ahmad Safi [3], Rasoul Ramezani [4], Taghi Ebrahimi Salari [5]

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 ۱،۸% of the population as "unhealthy." This seemingly small group accounts for a disproportionately high ۱۷،٤۷% 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

سامانه‌های پردازشی و ارتباطی چندرسانه‌ای هوشمند
Intelligent Multimedia Processing and Communication Systems (IMPCS)

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

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

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

سعید شعوری[۱*]، علی چشمی[۲]، احمد سیفی[۳]، رسول رمضانی[۴]، تقی ابراهیمی سالاری[۵]

۱. دانشجوی دکتری، گروه اقتصاد، دانشگاه فردوسی مشهد، مشهد، ایران. (نویسندهٔ 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. lspring05@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 طرح های بیمه سلامت با تمرکز ویژه بر کارکنان شرکت نفت شرق ایران پیشنهادمی‌کند. در این مطالعه، یک شبکه عصبی پرسپترون چندلایه (MLP) مشتمل بر شش مرحله برای اعتبارسنجی مشتری استفاده‌می‌شود. آموزش وزن‌ها به عنوان گامی حیاتی در پیاده‌سازی شبکه عصبی، تأثیر متغیرهای تو ضیحی را بر خروجی مدل تعیین‌می‌کند. سپس مدل آموزش‌دیده برای اعتبار سنجی استفاده‌می شود. نتایج اعتبار سنجی نشان‌می‌دهد که متغیرهای هزینه پرداختی بیمه سلامت و متغیرهای بیماری خاص بیشترین تأثیر را بر طبقه‌بندی مشتریان ناسالم دارند. نتیجه حائز اهمیت پس از اعتبار سنجی بیانگر این است که حدود ۱٫۸٪ از جمعیت به‌عنوان "نا سالم" شناسایی شده‌است. این گروه به‌ظاهر کوچک، ۱۷٫۴۷ درصد کل پرداختی شرکت بیمه بابت مطالبات بیمه درمان را تشکیل‌می‌دهند این درحالی است که اکنون حق بیمه این افراد برابر افراد سالم می‌باشد. مدل اعتبارسنجی پیشنهادی یک رویکرد عملی برای شرکت‌های بیمه ارائه‌می‌کند تا پروفایل ریسک مشتری را ارزیابی‌کرده تا حق بیمه‌ها را بر این اساس تنظیم‌کنند. این رویکرد سیستم بیمه‌ای عادلانه‌تر و پایدارتر را ترویج‌می‌کند.

طبقه‌بندی JEL: C52, I13, C45

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

۱. مقدمه

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 تأمین‌می‌شود 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 قرارگرفته‌است. تا کنون، مطالعات بسیار کمی در دسترس است که اعتبار مطالبات بیمه سلامت را مورد اعتبارسنجی قرارداده‌باشد[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 علاوه‌بر 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 داشته‌باشند و در مقابل، متقاضیان بیمه سلامت نیز به تناسب ریسکی که ایجادمی‌نمایند هزینه پرداخت‌می‌کنند و در صورتی که فردی کم‌ریسک باشد هزینه گزاف پرداخت‌نخواهدکرد.

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 ارزیابی‌می‌کند، بلکه کمک‌می‌کند تا فرآیند مدل بر اساس نتایج اعتبارسنجی، بهبودیابد[۴]. به‌عبارتی، اعتبارسنجی چیزی جز سازماندهی مناسب نیست و مهمترین ویژگی در این تعریف درجه اطمینان بالا، فرآیند خاص، ثابت و مشخصات از پیش تعیین‌شده است[۵]. امتیاز بیمه براساس نتایج اعتبارسنجی افراد است و بر حق بیمه‌ای که برای برخورداری از خدمات بیمه سلامت پرداخت‌می‌کنند، تأثیرمی‌گذارد. امتیاز پایین نشان‌دهنده ریسک بالاتر افراد برای شرکت‌های بیمه است بنابراین منجر به تقاضای حق بیمه بالاتر خواهدشد. امتیاز بالا به دلیل ریسک کمتر، باعث کاهش حق بیمه می‌شود. به‌طورکلی در این مطالعه سعی‌داریم اعتبارسنجی مشتریان بیمه سلامت را برای کارکنان شرکت نفت شرق ایران از طریق ارزیابی مبنی بر ویژگی‌های افراد بیمه‌شده و با استفاده از شبکه عصبی پرسپترن چندلایه یا MLP[1] انجام‌دهیم.

مقاله به شرح زیر تنظیم شده‌است: در بخش اول، اعتبار سنجی بیمه سلامت تجزیه و تحلیل شده‌است. در قسمت دوم پیشینه تحقیق بررسی و در قسمت سوم مبانی نظری تحقیق ارائه‌شده‌است. بخش چهارم به معرفی داده های تحقیق می‌پردازد و بخش پنجم به جزئیات طراحی و مدل‌سازی فرآیند اعتبارسنجی می‌پردازد. بخش پایانی یافته‌ها را خلاصه‌می‌کند و توصیه‌هایی نیز ارائه‌می‌دهد.

۲. پیشینه تحقیق

مطالعات زیادی در مورد اعتبارسنجی در حوزه‌های مختلف انجام‌شده‌است که در این بخش مطالعات انجام‌شده در زمینه اعتبارسنجی، طبقه‌بندی و شناسایی عوامل ریسک مربوط به بخش بیمه سلامت اشاره‌می‌شود: الکساندر بهرنت و همکاران [۳] (۲۰۲۰) به بررسی مبانی اعتبارسنجی مطالبات بیمه سلامت می‌پردازند. آن‌ها دو رویکرد را برای اعتبارسنجی بیمه سلامت معرفی‌می‌کنند. رویکرد اول اعتبارسنجی مبتنی بر مدل و رویکرد دوم اعتبارسنجی مبتنی بر طبقه‌بندی می‌باشد. پیش‌فرض اصلی رویکرد اول این است که منظور از اعتبار داده‌ها نیستند بلکه ویژگی و تفسیر داده‌ها و پیامدهای تحلیل‌هایی است که روی داده‌ها انجام‌می‌شود. 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 قابل‌مشاهده 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 اشاره‌دارد. در این مطالعه از یک مدل غربالگری متعارف قرارداد بیمه استفاده‌شده تا این مبادلات را در طیف وسیعی از محیط‌های اطلاعاتی مطالعه‌شود.

مرینر، وندی [۷] (2013) در تحقیقی به نقش بیمه درتعریف 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 مشاهده‌شد 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 هستنند.

در مطالعات اخیر گروه اول هادیان و همکاران [۸] (۲۰۰۷)، دقیقی و همکاران [۹] (۲۰۱۰) با استفاده از روش پنل دیتا، بیان‌کردند درآمد سرانه، مخارج بهداشتی انتظاری و نرخ تورم مهمترین عوامل تعیین‌کننده تابع تقاضای بیمه سلامت هستند .عباسی و تقی‌آبادی [۱۰] (2011) نشان-دادند که یک ارتباط بلندمدتی بین درآمد سرانه و نرخ باسوادی با تقاضای بیمه مکمل وجوددارد 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 نشان‌دادند 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 وفایی‌نجار و همکارن [۱۴] (۲۰۰۷) گل‌علیزاده و همکاران [۱۵] (۲۰۱۸) موهبتی [۱۶] (۲۰۱۸) اشاره‌کرد که با بررسی بیمه‌های مکمل در دنیا و ایران الگو و راهکارهایی را جهت بهبود وضعیت بیمه مکمل ایران پیشنهادکردند. از جمله مطالعه داخلی در زمینه صنعت بیمه و بیمه سلامت در حوزه ریسک صنعت بیمه صورت‌گرفته به شرح زیر می‌باشد:

ریاحی‌فر [17] (۱۳۸۵) در تحقیق خود به دنبال بررسی الگوی مدیریت ریسک بیمه‌ای درمان کشور می‌باشد که برای برآورد مدل از تکنیک دلفی سه مرحله‌ای استفاده‌شده‌است. براساس نتایج، اتخاذ یک استاندارد مدیریت ریسک برای طراحی الگوی فرآیند مدیریت ریسک در هر سازمان الزامی است که بر اساس این الگو، عناصر فرایند مدیریت ریسک، به صورتی نظم‌یافته مشخص‌می‌شوند. 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 ارائه‌شده 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 کمک‌می‌کند. ابتدا جامعه آماری مورد بررسی را از لحاظ ویژگی افراد بیمه-شده مورد تجزیه و تحلیل قرارمی‌دهد. در مرحله دوم براساس این ویژگی‌ها، اعتبارسنجی با استفاده از هوش مصنوعی صورت‌می‌گیرد و افراد جامعه مورد بررسی به افراد سالم و ناسالم طبقه بندی‌می‌شوند.

۳. مبانی نظری

همان‌طور که اشاره‌شد اخیراً مطالبات بیمه سلامت دوباره مورد توجه جامعه علمی به عنوان منبعی از شواهد دنیای واقعی در جهت، بهبود و توسعه بیشتر مراقبت‌های پزشکی و به‌اصطلاح آزمایش‌های عمل‌گرایانه قرارگرفته-است[18]. یکی از مهمترین معیارهای اعتبارسنجی، تعیین درجه ریسک مشتریان می‌باشد. درواقع، یک ریسک را می‌توان به‌عنوان یک عدد تصادفی X تعریف‌کرد که نتیجه (یا تحقق) واقعی آن ناشناخته است[19]. افرادی که به مجموعه ای از بیمه‌نامه‌های بیماری تعلق‌دارند، جمعیت نسبتاً ناهمگونی را تشکیل‌می‌دهند، 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 طبقه‌بندی‌کرد:

- ✓ عوامل ریسک عینی[2] ویژگی های جسمانی بیمه‌شده به ویژه: سن، جنسیت، سوابق سلامت، شغل می‌باشد.
- ✓ در میان عوامل ریسک ذهنی[3]، نگرش شخصی نسبت به سلامت را تشخیص‌می‌دهیم که تقاضای فردی برای درمان‌های پزشکی و در نتیجه درخواست مزایای بیمه را تعیین‌می‌کند.

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 پوشش‌می‌دهند. متغیر ریسک را با Xنشان‌می‌دهیم که غیرمنفی است و کلیه خسارت‌های احتمالی که می‌تواند متوجه شرکت بیمه شود را شامل‌می‌شود و حق بیمه را تابعی از X و به‌صورت π(X)نشان‌می‌دهیم. X یک متغیر تصادفی است که با RV نشان-داده‌می‌شود. یک RV نتیجه جمع تعداد محدود n تا از RVs است که $X = S_n = \sum_{i=1}^{n} X_i$,$\{X_i\}_{i\geq 1}$ یک دنباله از RVs مستقل و غیرمنفی و دارای توزیع یکسان است که در آن X_i مطالبات iام را نشان‌می‌دهد و در نهایت X_i دارای توزیع پواسن می‌باشد :

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

که S جمع X_i ها ا ست طوری که X_i در N زمان ت صادفی رخداده‌ا ست و X_i ها م ستقل از N ه ستند بدین معنی که شمار مطالبات بیمه رخداده و مقدار پولی که مرتبط نیستند.

پترو لیما رومز [۲۱](۲۰۱۷) اصل مطلوبیت مورد انتظار برای بیمه‌گر را بیان‌می‌کند و تصریح‌می‌کند 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:

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

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.

تعریف۱: طبق مطلوبیت مورد انتظار حداقل ارزش $\pi(X)$ که یک بیمه‌گر با سرمایه اولیه w باید هزینه‌کند برای پوشش ریسک X حل معادله زیر است:

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

هنگامی که تابع مطلوبیت مقعر باشد این اصل، اصل مطلوبیت صفر نامیده‌می‌شود. فرض‌کنیم 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 $X \sim Gama(\theta, \beta)$. $\beta > 0 , \theta > 0$. ابتدا باید تابع مولد گشتاور X را محاسبه‌کرد، 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.

$$M_X(t) = E[e^{Xt}] = \int_0^{+\infty} e^{Xt} P(X = x)dx$$

$$= \int_0^{+\infty} e^{xt} \frac{\beta^\theta}{\Gamma(\theta)} x^{\theta-1} dx$$

$$= \int_0^{+\infty} \frac{\beta^\theta}{\Gamma(\theta)} x^{\theta-1} dx \quad (4)$$

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:

$$M_X(t) = \int_0^{+\infty} \frac{\beta^\theta}{\Gamma(\theta)} \left(\frac{y}{\beta - t}\right)^{\theta-1} e^{-y} \frac{1}{\beta - t} dy$$

$$= \frac{\beta^\theta}{\Gamma(\theta)(\beta - t)^\theta} \int_0^{+\infty} y^{\theta-1} e^{-y} dy$$

$$= \frac{\beta^\theta}{\Gamma(\theta)(\beta - t)^\theta} \Gamma(\theta)$$

$$= \left(\frac{\beta}{\beta - t}\right)^\theta , t < \beta \quad (5)$$

از طرف دیگر با به‌کارگیری تعریف ۱ داریم:

$$u(w) = E[u(w + \pi[X]^- - X)] \Leftrightarrow -\alpha e^{-\alpha w}$$
$$= E\left[-\alpha e^{-\alpha(w + \pi[X]^- - X}\right]$$

$$\Leftrightarrow -\alpha e^{-\alpha w} = E\left[-\alpha e^{-\alpha w} e^{-\alpha \pi[X]} e^{\alpha X}\right]$$

$$\Leftrightarrow e^{\alpha \pi[X]^- = M_X(\alpha)}$$

$$\Leftrightarrow \pi[X]^- = \frac{1}{\alpha} \ln\left(M_X(\alpha)\right) \quad (6)$$

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 ۵ و ۶ ترکیب‌می‌شود:

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

به عبارت دیگر، برای پوشش ریسک X، بیمه‌گر به‌وسیله اصل مطلوبیت مورد انتظار به حداقل حق بیمه $\pi[X]^- = \frac{\theta}{\alpha} \ln\left(\frac{\beta}{\beta - \alpha}\right), \alpha < \beta$ 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 انتخاب‌کند. معمولاً نقطه شروع برای مدل سازی 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 تأثیربگذارد به‌عنوان رگرسیون در پارامتر میانگین توزیع شمارش گنجانده‌شده‌است. این اطلاعات برون زا را می‌توان با متغیرهای باینری کدگذاری‌کرد. در بیمه معمولاً یک تابع نمایی برای داشتن $\lambda_i = t_i \exp(x_i' \beta)$ استفاده‌می‌شود، جایی که t_i نشان‌دهنده در معرض خطر قرارگرفتن فرد بیمه‌شده i است. زیرا $E[N_i | X_i] = \lambda_i$ ، با داشتن ویژگی‌های بیمه‌شده، اکچوئری می‌تواند حق بیمه را برای یک بیمه‌شده محاسبه‌کند. از لحاظ تاریخی، شرکت‌های بیمه داده‌ها را به‌صورت مقطعی جمع‌آوری‌می‌کنند. 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 جمع‌آوری‌می‌شود. 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 تعریف‌کرد. به‌صورت متناوب، می‌توان آن را به‌عنوان فرآیند انتخاب داده‌ها و کاوش و ساخت مدل‌ها با استفاده از انبارهای 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

حال افزایش است و به یک موضوع مهم در حوزه های مختلف از جمله صنعت بیمه تبدیل‌شده‌است[24]. با توجه به اینکه در این تحقیق از الگوریتم‌های یادگیری ماشین[5] برای ارزیابی اعتبارسنجی مشتریان بیمه سلامت بهره‌می‌گیریم به بررسی مبانی نظری می‌پردازیم.

در این مطالعه از شـــبکه عصـــبی MLP به‌عنوان یکی از روش‌های داده‌کاوی استفاده‌شده‌است. شبکه عصبی MLP یک شبکه معروف عصبی است که بـــه‌نحـــو بـــسیار مطلـــوبی عمـــل اعتبارسنجی و امتیازدهی را انجـام مـی‌دهـد که در قسمت شیوه تجزیه و تحلیل داده‌ها به تفسیر در مورد روش MLP و مراحل اجرایی اعتبار سنجی در این تحقیق با ا ستفاده از روش شبکه عصبی MLP توضیح داده‌خواهدشد.

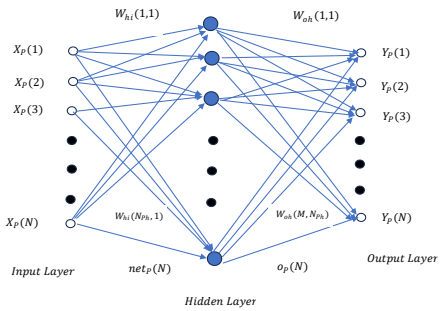
۱،۳. روش شبکه عصبی MLP

شبکه عصبی یک سیستم پردازش اطلاعات است که مشابه با شبکه‌های عصبی بیولوژیکی می‌باشد. شبکه‌های عصبی را می‌توان براساس معماری آن به انواع مختلفی طبقه‌بندی‌کرد. 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 تکامل‌یافته‌اند. پیشرفت‌های زیادی در بهبود عملکرد و در درک نحوه عملکرد این شبکه‌های عصبی وجودداشته‌است[26]. بااین‌حال، نیاز به پیشرفت‌های بیشتر در آموزش این شبکه‌ها همچنان وجوددارد.

شبکه‌های عصبی MLP از واحدهایی تشکیل‌شده‌اند که در لایه‌هایی مرتب‌شده‌اند. هر لایه از گره‌هایی تشکیل‌شده‌است و در شبکه‌های کاملاً متصل در نظرگرفته‌شده در این مقاله، هر گره به هر گره در لایه‌های بعدی متصل‌می‌شود. هر MLP از حداقل سه لایه تشکیل‌شده‌است که از یک لایه ورودی، یک یا چند لایه پنهان و یک لایه خروجی تشکیل‌شده‌است.

لایه ورودی: برای دریافت اطلاعات خامی که وارد شبکه تغذیه شده است. لایه‌های پنهان: عملکرد این لایه‌ها به‌وسیله ورودی‌ها و وزن ارتباط بین آن‌ها و لایه‌های پنهان تعیین می‌شود. وزن‌ها بین واحدهای ورودی و پنهان تعیین‌می‌کند که چه وقت یک واحد پنهان باید فعال‌شود. 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 واردمی‌شود، ورودی اصلی را در یک وزن با یک آستانه اضافه‌شده ضرب‌می‌کند و سپس از آن عبورمی‌کند. یک تابع فعال-سازی که ممکن است خطی یا غیرخطی (واحدهای پنهان) باشد. یک شبکه سه‌لایه معمولی در شکل ۱ آمده‌است. فقط سه لایه MLP در این مقاله در نظر گرفته‌خواهدشد زیرا نشان‌داده شده‌است که این شبکه‌ها هر تابع پیوسته را محاسبه‌می‌کنند. 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 مدل‌سازی‌می‌کنند:



شکل۱:.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_pمربوط به بردار ورودی N بعدی الگوی آموزشیp ام و y_p مربوط به بردار خروجی M بعدی از شبکه آموزش‌دیده برای الگویP 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 افزوده‌شده با $x_p(N+1)$ نشان‌داده‌شده‌است. واحدهای خروجی و ورودی خطی هستند. فعال‌سازی‌ها ورودیjامین واحد پنهان، با netp(j) بیان‌می‌شود.

$$net_p(j) = \sum_{k=1}^{N+1} W_{hi}(j,k).x_p(k) \qquad 1 \le j \le N_h \qquad (9)$$

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

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

فعال‌سازی غیرخطی معمولاً به‌عنوان تابع سیگموئیدی انتخاب‌می‌شود.

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

در رابطه (9) و (10)، واحدهای ورودی N با شاخص k نشان‌داده‌می‌شوند و $W_{hi}(j,k)$نشان‌دهنده وزن‌هایی است که k امین واحد ورودی را به jامین واحد پنهان متصل‌می‌کند. عملکرد کلی MLP با میانگین مربعات خطا (MSE) که به‌صورت زیر بیان‌می‌شود.

$$E = \frac{1}{N_V}\sum_{P=1}^{N_V} E_P = \frac{1}{N_V}\sum_{P=1}^{N_V}\sum_{i=1}^{M}\left[t_p(i) - y_p(i)\right]^2 \quad (11)$$

$$E_p = \sum_{P}^{N}\left[t_p(i) - y_p(i)\right]^2 \quad (12)$$

Epمربوط به خطای الگویp ام و t_p خروجی مورد نظر برای الگویp ام است. این رابطه همچنین اجازه‌می‌دهد تا محاسبه خطای نگاشت واحد خروجی i به‌صورت رابطه زیر انجام‌شود.

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

با خروجی iام برای الگوی آموزشی pام با رابطه ۱۴ نشان محاسبه‌می‌شود.

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

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

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 برآوردکرد [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 خواهدشد. همان‌طور که گفته‌شد در این مطالعه اعتبارسنجی بر مبنای عوامل ریسک در حوزه سلامت انجام خواهدگرفت. 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 تشکیل‌می‌دهند[30]. با توجه به این‌که 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 مشخص‌شود. 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.

۴. مدلسازی

این بخش طبقه‌بندی مشتریان بیمه سلامت را بر اساس یک سری ویژگی‌های بیمه‌شده با استفاده از الگوریتم MLP و نرم‌افزار پایتون[6] به سالم یا ناسالم نشان‌می‌دهد. در انجام این کار، نویسندگان توضیح می‌دهند که چگونه الگوریتم MLP باید پیاده‌سازی‌شود. برای شروع، مجموعه‌ای از توابع مورد استفاده در کتابخانه‌های پایتون مورد نیاز بود. برای الگوریتم MLP، بسته‌های مورد استفاده پانداهاس[7] 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 به‌صورت زیر در نرم‌افزار پایتون تعریف‌می‌شود از فایل اکسل بازخوانی‌می‌کنیم.

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

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

ویژگی‌های افراد بیمه‌شده که به‌عنوان متغیرهای توضیحی تعریف‌می‌شود، شامل موارد زیر می‌شود:

gender. جنسیت که با مقیاس ۰ (جنسیت زن) و ۱ (جنسیت مرد) مشخص‌می‌شود.

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

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

visit: تعداد مراجعه به سازمان‌های بهداشتی درمانی.

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

متغیر هدف:

result : سالم=۱ یا ناسالم=۰ تعریف‌می‌شود.

مرحله۲) تقسیم‌بندی داده‌ها به مجموعه‌های آموزشی و آزمایشی: در الگوریتم‌های یادگیری ماشینی، داده‌ها به دو دسته آموزش و آزمایش[8] تقسیم‌می‌شوند. به‌طور معمول، ۸۰٪ از کل داده‌ها برای آموزش استفاده می‌شوند، جایی که برای یادگیری و درنهایت پیش‌بینی استفاده‌می‌شود. ۲۰٪ باقی‌مانده‌ازداده‌هابرای آزمایش و ارزیابی ذخیره شده‌است، که امکان مقایسه پیش‌بینی‌ها در برابر نتایج واقعی به‌دست‌آمده از ۸۰٪ داده‌های آموزشی را برای ارزیابی دقت مدل فراهم‌می‌کند. در این پژوهش جامعه آماری شامل ۲۲۴۹۴ نفر می‌باشد.

مرحله ۳) نرمال‌سازی داده‌ها: پس از تقسیم‌بندی داده‌ها به دو گروه، در مرحله ۳ داده‌ها را نرمال‌سازی‌می‌کنیم. با توجه به این‌که 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 به‌صورت یک عدد)، برای مقایسه معنادار همه متغیرها باید در یک مقیاس باشند. بنابراین، از تابع نرمال‌سازی برای استانداردسازی دو مجموعه داده‌های آموزشی و آزمون به شرح زیر استفاده‌می‌شود:

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

$$X\_train\_normalized = scaler.fit\_transform(X\_train) \quad (18)$$

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

مرحله ۴) ایجاد یک مدل MLP طبقه‌بندی‌شده: هنگامی که MLP داده‌ها را دریافت‌می‌کند، می‌تواند ویژگی‌های مورد نیاز، تعداد لایه‌ها و میزان تکرارهای مورد نیاز را مشخص‌کند. الگوریتم MLP به صورت زیر تعریف‌شده و بر روی داده‌ها اجرامی‌شود:

$$mlp = MLPClassifier(hidden\_layer\_sizes = (1,1), max\_iter = 300, random\_state = 42) \quad (20)$$

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

آموزش وزن‌ها، مهمترین بخش در پیاده‌سازی یک شبکه عصبی می‌باشد. وزن متغیرها که تعیین‌کننده تأثیر بر متغیر وابسته سالم یا ناسالم است، در این مرحله شناسایی‌می‌شود. جدول ۱ وزن متغیرهای توضیحی در مدل را نشان‌می‌دهد.

جدول۱: وزن متغیرهای مدل

نام متغیر	وزن
جنسیت	0.0009938
سن	0.0360
بیماری خاص	0. 7979
تعداد مراجعات	0.000168
مطالبات بیمه سلامت	0.9477

منبع: یافته‌های مطالعه

همان‌طورکه نتایج وزن‌ها نشان‌می‌دهد 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. بااین‌حال، 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 که افراد سالم و ناسالم براساس اعتبارسنجی تفکیک گردید، انجام‌خواهدگرفت.

مرحله ۵) پیش‌بینی :مدل در چهار مرحله گذشته ایجادشد و مرحله ۵ جایی است که از مدل برای پیش‌بینی استفاده‌شد و افراد به دو دسته سالم و ناسالم تفکیک‌می‌شوند.

$$y\_pred = mlp.predict(X\_test\_normalized) \quad (22)$$

مرحله ۶) استفاده از مدل: پس از مدل‌سازی، مدل را با یک مثال مورد آزمون قرارمی‌دهیم. برای مثال فردی با ویژگی های زیر مورد ارز یابی قرارمی‌گیرد: جنسیت: مرد، سن:۳۹ سال، بیماری خاص ندارد، تعداد مراجعات ۸ بار و هزینه بابت مطالبات بیمه درمان برابر ۲۰۰۰۰۰۰۰ ریال. اجرای مدل به‌صورت 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 ['salem'] پیش‌بین می‌شود.

اکنون می‌توان از این مدل برای بهره‌برداری و طبقه‌بندی افراد سالم و ناسالم برای هر جمعیت جدید استفاده‌کرد. 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 مدل‌می‌شود. نتایج اعتبارسنجی بیانگر این است که حدود ۱,۸ درصد از جمعیت جدید مورد مطالعه را افراد ناسالم تشکیل‌می‌دهند. نتیجه قابل‌ذکر این است که ۱,۸ درصد جمعیت ناسالم در کل جمعیت تقریباً ۱۷,۴۷ درصد از کل مطالبات بیمه درمانی پرداخت‌شده توسط شرکت بیمه را به خود اختصاص‌می‌دهد که با این واقعیت که شرکت بیمه مبالغ بالایی را برای هزینه‌های درمان افراد ناسالم پرداخت‌می‌کند، مطابقت‌دارد.

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

معمولاً از چندین معیار استاندارد استفاده‌می شود. ممکن است فرض شود که تنها دو مورد وجود دارد: یا الگوریتم موقعیت را به‌درستی 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، TN و FN باید محاسبه شود. برای یک مسئله طبقه‌بندی باینری، ماتریس درهم‌ریختگی[9] یک ماتریس ۲×۲ بود که شامل پارامترهای ذکرشده‌بود. برای روشن شدن موضوع، فرض کنید مشکل طبقه‌بندی افراد سالم از افراد ناسالم است. برای این منظور، ویوویچ [33] (۲۰۲۱) معیارهای استانداردی را برای طبقه بندی در جدول ۷ معرفی‌می‌کند. برای درک جدول ۷، تصور کنید که الگوریتم طبقه‌بندی پس از یادگیری و ایجاد مدل، تحت یک مجموعه آزمون قرارمی‌گیرد.

جدول7: معیار استاندارد کلاس‌بندی (ماتریس درهم‌ریختگی)

		Predicted Class: Positive	Predicted Class: 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، تصورکنید که الگوریتم طبقه‌بندی پس از آموزش و ایجاد مدل، تحت یک مجموعه آزمون قرارمی‌گیرد. در جدول شماره ۷، ردیف‌ها برچسب‌های واقعی را نشان‌می‌دهند و ستون‌ها پیش‌بینی‌های الگوریتم را نشان‌می‌دهند. بر این اساس، چهار مورد وجوددارد:

۱ .مثبت واقعی (TP): برخی از افراد ناسالم به‌درستی ناسالم شناخته‌می‌شوند.

۲ .منفی کاذب (FN): برخی از افراد ناسالم به اشتباه سالم شناخته‌می‌شوند.

۳ .مثبت کاذب (FP): :برخی از افراد سالم به اشتباه ناسالم شناخته-می‌شوند.

۴ .منفی واقعی (TN): افراد سالم به‌درستی، سالم شناخته‌می‌شوند.

در نهایت، هر نمونه داده در یکی از این دو "کلاس" قرارمی‌گیرد. بنابراین، برای هر نمونه داده، یکی از چهار سناریوی ذکرشده در بالا ممکن است رخ‌دهد.

معیارهای ارزیابی:

پاورز[34] (۲۰۱۱) برای ارزیابی میزان کیفیت سیستم خودکار یادگیری ماشین، معیارهای استاندارد صحت[10] ، پوشش[11] ، دقت[12] و معیارF [13]را معرفی‌می‌کند.

دقت: نسبت موارد واقعاً مثبت به همه پیش‌بینی‌ها را نشان‌می‌دهد.

صحت: نسبت موارد مثبت پیش‌بینی‌شده صحیح به کل موارد مثبت پیش‌بینی‌شده را نشان‌می‌دهد.

کارایی: نسبت موارد واقعاً مثبت را نشان‌می‌دهد که به‌درستی مثبت پیش‌بینی شده‌اند. این پوشش موارد واقعاً مثبت را می‌سنجد.

با توجه به این‌که دو معیار دقت و کارایی عکس یکدیگر هستند؛ به این‌صورت که افزایش یکی باعث کاهش دیگری و بالعکس می‌شود بنابراین یک معیار دیگر به نام معیار F تعریف‌می‌کنیم. این معیار درواقع میانگین هند سی دو معیار ذکر شده‌است. پاورز [34] (۲۰۱۱) معیارهای ارزیابی سیستم‌های یادگیری ماشین را مطابق با جدول 7 تعریف‌می‌کند:

جدول8: معیارهای ارزیابی مدل

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

در این بخش ما در ابتدا ماتریس درهم‌ریختگی را به‌دست‌می‌آوریم. برای این امر دستور زیر را اجرامی‌کنیم:

$$cm = confusion_{matrix(y_{test}, y_{pred})} \quad (23)$$

نهایتاً ماتریس درهم‌ریختگی به‌صورت زیر به‌دست‌می‌آید:

جدول۹: ماتریس درهم ریختگی(منبع: یافته‌های مطالعه)

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

مثبت واقعی (TP): تعداد نمونه‌های مثبتی که به‌درستی 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 به‌درستی 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 محاسبه‌می‌کند.

جدول10: مقادیرمعیارهای ارزیابی مدل(منبع: یافته‌های مطالعه)

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

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 پیش‌بینی‌می‌کند. 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 F۱ (F- measure) استفاده‌شد. معیارصحت(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 بودهو بیانگر این است که اگر الگوریتم، فردی را سالم تشخیص‌دهد ۱۰۰ درصد، درست پیش‌بینی کرده‌است. در برخی مسائل از جمله اعتبارسنجی بیمه سلامت، معیار پوشش (Recall) از معیار صحت (Precision) مهمتر است. بدین‌ترتیب که فرض‌کنیم شخصی واقعاً ناسالم است ولی به اشتباه سالم تشخیص داده‌شود، در این صورت هزینه بالایی را به شرکت بیمه تحمیل‌می‌کند به‌دلیل اینکه فرد ناسالم، حق بیمه فرد سالم را پرداخت-می‌کند. معیار F1 که درواقع ترکیب متعادلی بین معیارهای دقت و صحت است، می‌تواند در مواردی که هزینه False Positive و False Negative متفاوت است به‌کاررود (در این مقاله این مقادیر ۹۰ و ۰ است) لذا استفاده از این معیار پذیرفتنی است. مقدار معیارF1 برابر۰,۹۸۹۸ 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 ارزیابی‌کنند. لذا دراین مطالعه اعتبارسنجی با استفاده از روش شبکه عصبی MLP و برای افراد تحت پوشش بیمه سلامت کارکنان شرکت نفت شرق ایران انجام‌گرفت. 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 به‌ترتیب بیشترین تأثیر در انتخاب‌شدن نتیجه ناسالم یا به‌عبارتی 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 حاصل‌می‌شود. نتایج پیش‌بینی نشان‌می‌دهد علی‌رغم 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 تحمیل‌می‌کند.

درنهایت بعد از فرآیند آموزش، مدل طراحی‌شده با معیارهای مختلف از جمله معیارهای استاندارد صحت، دقت، پوشش و معیار اندازه‌گیری F ارزیابی‌شد که هر کدام، یک سری ویژگی‌های مدل را بررسی‌کردند. مقادیر این معیارها به‌ترتیب 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 انجام‌دهند.

سپاسگزاری

در پایان از مرکز بهداشت و درمان شرکت نفت شمال شرق ایران که ما را در انجام مطالعه حمایت‌کردند، 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 th 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 andMultidisciplinary 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.

[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] Vafaee 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.

[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.

پی‌نوشت

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