Journal of Economic Modeling Research
Print ISSN: 2228 - 6454
Online ISSN: 2538 - 4163
Homepage: https://jemr.khu.ac.ir/

Classification of Healthcare Insurance Customers Using Data-Driven Marketing Techniques

Mohammad-Reza Asefi [1] | **Abbas Khandan** [2*]

1. Master student of industrial engineering, Faculty of Economics, Kharazmi University, Tehran, Iran. E-mail: m.r.asefi8888@gmail.com
2. Corresponding Author, Assistant Professor, Faculty of Economics, Kharazmi University, Tehran, Iran. E-mail: khandan.abbas@khu.ac.ir (0000-0002-4558-6653)

Article Info

Article type:
Research Article

Article history:
Received:
05 May. 2024

Received in revised form:
18 Aug. 2024

Accepted:
14 Sep. 2024

Keywords:
Health insurance,
Data-driven marketing,
Data mining,
Machine learning,
Classification

JEL:
G32, C53, C58

ABSTRACT

The aim of this study is to identify and classify insurance customers in order to identify the target population for increasing the profitability of insurance companies, achieving a balance in premium payments, and examining the health questionnaire as an indicator of policyholders' preferences. Moreover, designing a marketing strategy to optimize advertising efficiency. In this paper, five machine learning algorithms, namely Decision Tree, Random Forest, Support Vector Machine, Naive Bayes, and Logistic Regression, are used to classify customers into two categories: profit-generating and loss-generating. Data from a private insurance company is utilized, consisting of 2,897 observations collected from December 1400 to December 1401. By utilizing machine learning methods and focusing on the target population, the chances of success can be increased. The presence of a small number of individuals who significantly reduce the profitability of insurance companies is evident. The pre-existing medical conditions of individuals have a considerable impact on their insurance usage and the damage caused to insurance companies. Machine-learning methods can provide a comprehensive understanding of insurance customers and their needs. By identifying the target population, insurance companies can increase their profitability and satisfy their customers by addressing their specific demands.

Cite this article: Asefi, Mohammad-Reza., & Khandan, Abbas. (2022). Classification of Healthcare Insurance Customers Using Data-Driven Marketing Techniques. *Journal of Economic Modeling Research*, 14 (52), 96-138. DOI: 0000000000000000000

Publisher: Kharazmi University
DOI: 00000000000000000000000000
Journal of Economic Modeling Research, Vol, 14, No. 52, 2022, pp. 96-138.

Journal of Economic Modeling Research
Print ISSN: 2228 - 6454
Online ISSN: 2538 - 4163
Homepage: https://jemr.khu.ac.ir/
Kharazmi University

دسته‌بندی مشتریان بیمه درمان با تکنیک‌های داده‌کاوی

محمدرضا آصفی[۱] | عباس خندان[*۲]

۱. دانشجوی کارشناسی ارشد مهندسی صنایع، گروه آموزشی اقتصاد امور عمومی، دانشکده اقتصاد، دانشگاه خوارزمی
رایانامه: m.r.asefi8888@gmail.com
۲. نویسنده مسئول، استادیار، گروه آموزشی اقتصاد امور عمومی، دانشکده اقتصاد، دانشگاه خوارزمی.
رایانامه: khandan.abbas@khu.ac.ir

اطلاعات مقاله

نوع مقاله:
مقاله پژوهشی

تاریخ دریافت:
۱۴۰۳/۰۲/۱۶

تاریخ ویرایش:
۱۴۰۳/۰۵/۲۸

تاریخ پذیرش :
۱۴۰۳/۰۶/۲۴

واژه‌های کلیدی:
بیمه درمان، بازاریابی داده‌محور، داده‌کاوی، یادگیری‌ماشین، طبقه‌بندی.

طبقه‌بندی JEL:
G32, C53, C58

چکیده

هدف اصلی این مطالعه، شناسایی و طبقه‌بندی مشتریان بیمه درمان به منظور شناسایی جامعه هدف و در نتیجه افزایش سودآوری شرکت‌های بیمه، ایجاد توازن در پرداختی حق‌بیمه و طراحی استراتژی بازاریابی است. در این مقاله از ۵ الگوریتم یادگیری ماشین درخت تصمیم، جنگل تصادفی، ماشین بردار پشتیبان، بیز ساده و رگرسیون لجستیک به منظور طبقه‌بندی مشتریان به دو دسته سودده و زیان‌ده استفاده شده است. به این منظور از داده‌ها و اطلاعات تعداد ۲۸۹۷ بیمه‌نامه درمان انفرادی یک شرکت بیمه خصوصی در بازه زمانی آذر ۱۴۰۰ تا آذر ۱۴۰۱ استفاده شده است. نتایج این مقاله نشان می‌دهد که با استفاده از روش‌های یادگیری ماشین و ویژگی های ثبت شده از مشتریان در پرسشنامه سلامت می‌توان سودده یا زیانده بودن آن‌ها را تا حدود مناسبی پیش‌بینی کرد. این مقاله نشان می‌دهد که تمرکز بر روی جامعه هدف معرفی شده توسط مدل شانس موفقیت و افزایش سود را به مقدار چشم گیری افزایش می‌دهد. بنابراین، با استفاده از روش‌های یادگیری ماشین می‌توان به درک مناسبی از مشخصه‌های مشتریان بیمه درمان و نیازهای آنها رسید. پیدا کردن جامعه هدف علاوه بر این‌که به افزایش سود شرکت بیمه منجر می‌شود می‌تواند با تمرکز بر خواسته‌های مشتریان به افزایش رضایتمندی آن‌ها نیز منجر شود.

استناد: آصفی، محمدرضا؛ و خندان، عباس (۱۴۰۲). دسته‌بندی مشتریان بیمه درمان با تکنیک‌های داده‌کاوی. *تحقیقات مدل‌سازی اقتصادی*، ۱۴ (۵۲)، ۹۶-۱۳۸. DOI: 00000000000000000000

ناشر: دانشگاه خوارزمی.

۱. مقدمه

با توسعه سریع صنعتی و اقتصادی، حوادث با شدت و تنوع بیشتری در حال ظهور و شکل‌گیری هستند. یکی از این خطرات مواجهه با بیماری‌ها و هزینه‌های پزشکی و درمانی و اعمال جراحی است که در اغلب موارد به علت بالا بودن هزینه‌ها، امکان رویارویی افراد و خانوار با این مشکلات و جبران خسارت آن‌ها به تنهایی وجود ندارد. از آنجائی که بسیاری از بیماری‌ها اغلب به صورت تصادفی و مستقل از یکدیگر برای افراد اتفاق می‌افتد، می‌توان با استفاده از مکانیسم بیمه و با پرداخت حق‌بیمه این خطرات را به شرکت بیمه انتقال داد. در ایران در کنار سازمان‌هایی نظیر سازمان تأمین اجتماعی، سازمان تأمین خدمات درمانی و سازمان بازنشستگی کشوری که سطح پایه خدمات درمانی را به بیمه‌شدگان ارائه می‌دهند، شرکت‌های بیمه بازرگانی نیز به طور گسترده در حوزه سلامت فعال بوده و انواع مختلفی از بیمه‌های درمان را به صورت اختیاری و بر اساس تقاضا و نیاز بیمه‌گذاران خود در چهارچوب مصوبات شورایعالی بیمه عرضه می‌نمایند. در واقع بیمه‌های درمان خصوصی یک سطح تکمیلی از خدمات درمانی را برای افرادی که مایلند خدمات بیشتر و کامل‌تری داشته باشند، ارائه می‌دهند.

در بازار سلامت به دلیل وجود مشکل عدم تقارن اطلاعات[1] سوددهی شرکت‌های بیمه کم بوده و تعیین حق‌بیمه عادلانه متناسب با ریسکی که افراد بر شرکت بیمه وارد می‌کنند، کاری بسیار دشوار است. عدم‌تقارن اطلاعات به دو صورت ویژگی‌های پنهان[2] و اقدامات پنهان[3] تجلی می‌یابد که به ترتیب می‌توانند منجر به کژگزینی[4] و کژمنشی[5] شود. کژگزینی در بازار بیمه بیانگر این نگرانی است که بیمه‌گذار اطلاعات در مورد سلامت خود را به اندازه کافی افشا نکرده باشد و در عمل ریسک بیشتری را به بیمه‌گر منتقل کند. به این ترتیب، پورتفوی بیمه‌گر متشکل از افراد پرریسک خواهد شد. کژمنشی نیز می‌تواند یک رفتار شایع در بیمار و همچنین پزشک باشد؛ بیماری که پس

[1] Assymetric information
[2] Hidden Characteristics
[3] Hidden Actions
[4] Adverse Selection
[5] Moral Hazard

از پوشش بیمه مراجعات بیشتری به پزشک دارد و پزشکی که ممکن است در جهت منافع شخصی خدمات درمانی بیشتری را به بیمار تجویز کند. این تجویز غیرضروری خدمات درمانی از جیب شرکت بیمه‌گر تحت عنوان تقاضای القائی[1] پزشکان شناخته می شود. در هر دو مورد، منفعت انتظاری ناشی از مبادله برای بیمه‌گر به خطر می‌افتد و انگیزه شرکت بیمه برای گسترش پوشش کاهش می‌یابد که به معنی ناکارایی در بازار است و علاوه بر این می‌تواند به مشکلات اجتماعی زیادی در حوزه سلامت منجر شود.

البته توسعه هوش مصنوعی و افزایش حجم داده‌ها در سال‌های اخیر به کمک شرکت‌های بیمه آمده و کاربرد فراوانی در این حوزه یافته است (بوبوتیز[2] و همکاران ۲۰۲۰). استفاده از روش‌های یادگیری ماشین[3] در دسته‌بندی مشتریان[4] بیمه درمان می‌تواند عدم تقارن اطلاعات را برطرف و در تعیین حق‌بیمه عادلانه و سودآوری شرکت مفید واقع گردد. در واقع دسته‌بندی[5] مشتریان بیمه درمان به سودده و زیان‌ده با استفاده از الگوریتم‌های یادگیری ماشین و بکارگیری تکنیک بازاریابی داده-محور[6] می‌تواند به شرکت‌های بیمه در بازار رقابتی و پویای بیمه درمان کمک کند تا استراتژی‌های مناسبی برای جذب و نگهداشت هر دسته از مشتریان که سلایق و ترجیحات آن‌ها در طول زمان تغییر می‌کند، داشته باشند (براورمن[7] ۲۰۱۵).

در این راستا، این مقاله قصد دارد نخست به این سؤال پاسخ دهد که بر اساس مشخصه‌های ثبت شده در پرسشنامه سلامت، چه پیش‌بینی می‌توان در مورد سودده یا زیانده بودن مشتریان داشت و از میان مشخصه‌ها کدام موارد اهمیت بیشتری در غربال‌گری افراد و مشتریان دارد؟ این مهم با روش یادگیری ماشین و با استفاده از اطلاعات ۲۸۹۷ بیمه‌نامه درمان انفرادی یک شرکت بیمه خصوصی در بازه زمانی آذر ۱۴۰۰ تا آذر ۱۴۰۱ انجام خواهد شد. اطلاعات ثبت شده از افراد شامل سن، شهر

[1] Induced demand
[2] Boßow-Thies
[3] Machine learning
[4] Customer Segmentation
[5] Classification
[6] Data Driven Marketing
[7] Braverman

محل سکونت، اطلاعات پرسشنامه سلامت[1] (اطلاعاتی که نشان می‌دهد افراد به چه بیماری‌های جسمی یا روحی مبتلا هستند)، حق‌بیمه پرداختی و هزینه‌های درمان می‌باشد. در این مقاله همچنین از پنج الگوریتم دسته‌بندی مختلف از جمله درخت تصمیم[2]، جنگل تصادفی[3]، بیز ساده[4]، رگرسیون لجستیک[5] و ماشین بردار پشتیبان[6] استفاده خواهد شد که نه تنها مقوم نتایج بدست آمده خواهند بود بلکه امکان مقایسه عملکردی الگوریتم‌های مختلف در پیش‌بینی را نیز فراهم خواهد کرد.

ساختار مقاله به این صورت است که در بخش دوم ابتدا به ادبیات موضوع و پیشینه پژوهش پرداخته خواهد شد. بخش سوم به ارائه آمار توصیفی از متغیرها و داده‌های پژوهش و معرفی الگوریتم‌های دسته‌بندی مورد استفاده اختصاص دارد. در بخش چهارم شش الگوریتم یادگیری ماشینی بکار گرفته شده و نتایج بدست آمده از هر الگوریتم و همچنین عملکرد الگوریتم‌های مختلف در پیش‌بینی مورد سنجش و مقایسه قرار خواهند گرفت. بخش پنجم مقاله به نتیجه‌گیری و ارائه توصیه‌های سیاستی می‌پردازد.

۲. مبانی نظری و پیشینه تحقیق

بیمه‌های تکمیلی درمان یکی از مهمترین ارکان بهبود خدمات بهداشت و درمان در کشور هستند که باعث افزایش دسترسی مردم به مراقبت‌های بهداشتی می‌شود. نکتهٔ حائز اهمیت این است که باید به شرکت بیمه بازرگانی ارائه‌دهنده بیمه درمان به عنوان یک بنگاه اقتصادی نگاه کرد که به دنبال سود بیشتر است و به منظور تحقق اهداف تجاری خود لازم است تا ریسک‌های پیش روی خود را مدیریت کند. مدیریت ریسک برای بیمه‌گر درمان به دلیل عدم تقارن اطلاعات می‌تواند بسیار دشوار باشد و بیمه‌گر در نتیجهٔ کژگزینی و کژمنشی با افزایش هزینه‌های درمان روبرو شود. کژگزینی در نتیجه پنهان بودن یک مشخصهٔ پنهان که در اینجا سلامت است اتفاق می‌افتد. طبیعتاً افراد بسته به درجات ریسک بیماری می‌بایست حق‌بیمه بالاتری بپردازند اما در شرایط عدم تقارن اطلاعات و

[1] Health Questionnaire
[2] Decision Tree
[3] Random forest
[4] Naive Bayes
[5] Logistic regression
[6] Support vector machine

پنهان بودن وضعیت سلامت، بیمه‌گر ناچار است یک حق‌بیمه متوسط از همه بیمه‌شدگان اخذ کند که به معنی متضرر شدن افراد با ریسک پایین و منتفع شدن افراد با ریسک بالاست. به عبارت دیگر، در بازار بیمه این نگرانی وجود دارد که بیمه‌گذار اطلاعات در مورد سلامت خود را به اندازه کافی افشا نکرده باشد و در عمل ریسک بیشتری را به بیمه‌گر منتقل کند. در نتیجه این عدم تقارن اطلاعات و اخذ یک حق‌بیمه متوسط بالا توسط بیمه‌گر، افراد با ریسک پایین انگیزه کمتر و افراد با ریسک بیماری بالا انگیزه بیشتری دارند که تحت پوشش قرار گیرند. به این ترتیب، پورتفوی بیمه‌گر متشکل از افراد پرریسک خواهد شد که در اصطلاح به آن کژگزینی گویند.

یک راه برای جلوگیری از کژگزینی غربال‌گری[1] مشتریان است که نخستین بار توسط اسپنس[2] (۱۹۷۳) معرفی شد. با غربالگری، طرف در موضع ضعف اطلاعاتی تلاش می‌کند تا مشخصه پنهان طرف دیگر را بر اساس دیگر اطلاعات مرتبط یا سابقه فرد کشف کند. برای این امر شرکت‌های بیمه درمان از پرسشنامه‌های پزشکی و بررسی سوابق بیماری آن‌ها استفاده می‌کنند. پرسشنامه پزشکی برای آشکار شدن مشخصه‌های پنهان و ریسک افراد کارساز است و در صورتی که افراد پرریسک شناسایی شوند یک اضافه نرخ پزشکی علاوه بر حق‌بیمه معمول پرداخت خواهند کرد. با این همه، کارآمدی این ابزار غربال‌گری برای رفع عدم تقارن اطلاعات و رفع مشکل کژگزینی به کیفیت پرسشنامه‌ها بستگی دارد. در این راستا است که داده‌کاوی از حجم بزرگ اطلاعات مشتریان در سال‌های اخیر مورد توجه زیادی قرار گرفته است. داده کاوی همان استخراج اطلاعات از داده‌ها است که بر پایهٔ روش‌ها و تکنیک‌های گوناگونی برای کشف الگوها، روابط و اطلاعات نهفته در مجموعه‌های بزرگ داده‌ها انجام می‌شود.

البته داده‌کاوی اطلاعات و آشکار شدن ریسک و دسته‌بندی مشتریان به سودده و زیان‌ده به معنی حذف مشتریان پرریسک یا عدم پوشش آن‌ها نیست بلکه شرکت بیمه می‌تواند با شناسایی ریسک مشتریان علاوه بر اخذ یک حق‌بیمه متوازن، به آن‌ها محصولاتی منطبق با نیازهایشان ارائه دهد. به

[1] Screening
[2] Spence

استفاده از اطلاعات مشتریان و بازارها با کمک تکنولوژی‌های جدید بازاریابی داده‌محور گفته می‌شود. بزرگترین چالش در بازاریابی داده‌محور عبارتند از رسیدن به بینش درست از داده‌ها، شناسایی جامعه هدف، شخصی سازی داده‌ها، طراحی و بهینه سازی سیستم با توجه به الگوریتم های موجود. در این راستاست که تکنیک‌های هوش مصنوعی، یادگیری ماشین و یادگیری عمیق مهم‌ترین ابزارهایی هستند به کمک بازاریابی داده‌محور می‌آیند (بوبو تیز و همکاران[1]، ۲۰۲۰).

در یادگیری ماشین، به جای برنامه‌نویسی همه چیز، داده‌ها به یک الگوریتم عمومی داده می‌شوند و این الگوریتم است که براساس داده‌ها منطق خود را می‌سازد. خروجی تکنیک‌های یادگیری ماشین یک مدل ریاضی است که تعدادی ورودی را می‌گیرد و پیش‌بینی‌هایی انجام می‌دهد. الگوریتم‌های یادگیری ماشین به سه دسته کلی یادگیری نظارتی[2]، یادگیری بدون نظارت و یادگیری تقویتی[3] تقسیم می‌شوند. یادگیری تحت نظارت به طور کلی به دو دسته رگرسیون (وقتی متغیر پاسخ پیوسته باشد) و دسته‌بندی (وقتی متغیر پاسخ گسسته باشد) تقسیم می‌شود. در مدل‌های دسته‌بندی مجموعه داده بر اساس برچسب مشخصی طبقه‌بندی شده و مدلی برای پیشبینی برچسب داده‌های جدید ساخته می‌شود (سن و همکاران[4]، ۲۰۲۰). این الگوریتم از آن جهت تحت نظارت نامیده می‌شود که برچسب داده‌ها باید از پیش توسط ناظر یا پژوهشگر مشخص شده باشد. در صورتی که برچسب مشخصی برای داده‌ها وجود نداشته باشد، الگوریتم در یادگیری نظارت نشده باید خود به تنهایی به‌دنبال ساختارهای جالب موجود در داده‌ها باشد. در این مقاله از الگوریتم‌های تحت نظارت استفاده خواهد شد چرا که برچسب زیان‌ده یا سودده بودن مشتریان و میزان هزینه درمان آن‌ها مشخص است و هدف پیش‌بینی مشتریان در این دسته‌ها است.

[1] Boßow-Thies et al.
[2] supervised learning
[3] Reinforcement learning
[4] Sen et al.

۳. پیشینه تحقیق

بطور کل مطالعات کم و محدودی هستند که با استفاده از داده‌کاوی به بازاریابی داده‌محور یا دسته‌بندی مشتریان بیمه درمان انفرادی در ایران پرداخته باشند از این رو در ادامه مطالعاتی که از داده‌کاوی برای کل صنعت بیمه استفاده کرده‌اند مرور خواهد شد.

تجددی نودهی و همکاران (۱۴۰۲) با تأکید بر این که پیش‌بینی هزینه‌های درمانی افراد کاری بس دشوار است، پیشنهاد می‌کند که از رویکردی مبتنی بر علم داده و یادگیری ماشین و الگوریتم‌های یادگیری جمعی برای پیش‌بینی افراد پرخطر و کم‌خطر در بیمه درمان استفاده شود. منظور از یادگیری جمعی ترکیب روش‌های گوناگون است که در این مطالعه از مدل‌های مبتنی بر رگرسیون لجستیک، شبکه‌های عصبی، ماشین‌های بردار پشتیبانی، جنگل‌های تصادفی، (LightGBM) و(XGBoost) استفاده شده است. هدف از ترکیب این روش‌ها این است که از نقاط قوت آن‌ها استفاده و نقاط ضعف آن‌ها به حداقل برسد تا دقت پیش‌بینی افزایش یابد. این مقاله در نهایت با (AUC) برابر ۰/۷۳ توانست به پیش‌بینی افراد پرخطر و کم‌خطر بپردازد که بیانگر اثربخشی مدل است.

خندان و همکاران (۱۴۰۲) با رویکرد داده‌کاوی و استفاده از الگوریتم‌های یادگیری عمیق و شبکه عصبی که دقت بالایی در پیش‌بینی دارند به پیش‌بینی بازخرید بیمه‌نامه‌های زندگی به شرط فوت می‌پردازد. در این مطالعه از داده‌های آماری ۳۵۱۷۱ خریدار بیمه‌نامه‌های عمر و مستمری یک شرکت بیمه‌ای در مقطع سال ۱۴۰۰ استفاده شد و مدل برآورد شده از دقت مطلوب ۷۴ درصد در پیش‌بینی هر دو نوع بیمه‌نامه‌های عدم بازخرید و بازخرید شده برخوردار شد. نتایج بدست آمده نشان می‌دهند که از مشخصه‌های جمعیت‌شناختی متغیرهای سن، جنسیت زن، اضافه نرخ پزشکی، نرخ خطر حادثی و از مشخصه‌های قراردادی نیز مدت بیمه‌نامه، مدت زمان سپری‌شده از شروع بیمه‌نامه، شیوه پرداخت حق‌بیمه با اقساط بلندمدت‌تر، بالاتر بودن ضرایب افزایش سالانه سرمایه و حق‌بیمه و کمتر بودن تعداد موارد پوشش و سرمایه فوت با بازخرید اثر عکس داشته و احتمال آن را کاهش می‌دهند. نسبت خویشاوندی بین بیمه‌گذار و بیمه‌شده نیز تأثیرگذار بوده و نشان داده شد

که بازخرید وقتی بیمه‌گذار بیمه‌نامه عمر را برای خود بخرد در حدأقل و با دور شدن نسبت خویشاوندی احتمال بازخرید افزایش می‌یابد.

قربانی و همکاران (۱۴۰۱) با استفاده از روش‌های داده‌کاوی از جمله جنگل تصادفی، درخت تصمیم، رگرسیون لجستیک و شبکه عصبی به طبقه‌بندی مشتریان بیمه‌های زندگی بر حسب ریزش یا عدم ریزش می‌پردازند. در این پژوهش از اطلاعات بیمه‌نامه‌های زندگی یک شرکت بیمه پایلوت در سال ۱۳۹۸ برای استان تهران استفاده شده و نتایج بدست آمده نشان می‌دهند که احتمال بازخرید بیمه‌نامه‌های عمر در سنین بالاتر، در میان زنان و افراد دارای مشاغل پر ریسک بیشتر است. از بررسی مشخصه‌های قراردادی نیز یافته‌های بدست آمده حاکی از آن است که در میان بیمه‌نامه‌های با اقساط سالانه، حق‌بیمه کمتر و درصد ضریب تغییر سرمایه بیشتر، بازخرید به احتمال کمتری اتفاق افتاده است.

پرستش (۱۳۹۹) با استفاده از داده‌ها و اطلاعات ۵۰۰ بیمه‌گذار بیمه سلامت و عمر یک شرکت بیمه‌ای منتخب طی سال‌های ۱۳۹۵ تا ۱۳۹۸ و بکارگیری الگوریتم کی-میانگین به خوشه‌بندی مشتریان می‌پردازد. طبق نتایج بدست آمده مشتریان به ۴ دسته مشتریان ویژه، برتر، میانی و ضعیف تقسیم شدند. در این تحقیق متغیرهای مورد بررسی به روش کی-میانگین وزن‌دهی شده و در نهایت سن به عنوان مهم‌ترین ویژگی، جنسیت در اولویت دوم و تحصیلات به‌عنوان یک مشخصه کم اهمیت شناخته شد. در این مقاله نتیجه گرفته می‌شود که استفاده از جایزه برای مشتریان برتر و تنبیه مشتریان زیان‌آور یک الگوریتم سودده و قابل اجرا در شرکت‌های بیمه است و به همین دلیل بازی-ورسازی در نرم افزارهای جدید که قابلیت پیاده‌سازی کارهایی مانند امتیازدهی و جایزه را فراهم می‌کنند بسیار مفید خواهد بود.

باش‌افشار و همکاران (۱۳۹۷) در مقاله ای با جامعه آماری ۱۰۰۰ نفر به خوشه‌بندی مشتریان بیمه عمر ایران برای سال ۱۳۹۲ می‌پردازند. در این مقاله با استفاده از متغیرهای مربوط به اطلاعات بیمه-نامه، اطلاعات جمعیتی و تکمیلی بیمه‌شدگان مانند سوابق بیماری، عوامل مؤثر بر ریزش مشتری شناسایی شده و الگوریتم کی-میانگین برای خوشه‌بندی مشتریان بکار گرفته می‌شود. طبق نتایج بدست آمده، متغیرهای جمعیتی همچون «جنسیت» و «سن» و متغیرهای بیمه‌ای همچون «حق‌بیمه

سالیانه» و «ضریب فوت در اثر حادثه» از جمله مهمترین عوامل تأثیرگذار در شناسایی گروه‌های مشتریان هستند. نتایج بدست آمده نشان می‌دهند که یک دسته تحت عنوان مشتریان سودآور نسبت به دریافتی کمی که از بیمه دارند حق بیمه بالایی را پرداخت می‌کنند و در مقابل دسته مشتریان ریسکی حق بیمه کمی پرداخته و ضریب فوت در اثر حادثه بالایی دارند.

از جدیدترین مطالعات خارجی در این زمینه نیز می‌توان به موارد زیر از جمله بائو و حنیف[1] (۲۰۲۴) اشاره کرد. این مقاله از یادگیری ماشین برای قیمت‌گذاری بیمه‌های سلامت استفاده می‌کند تا سودآوری شرکت‌ها در دوران همه‌گیری کرونا حفظ شود. برای این منظور از داده‌های بیمه سلامت ایالات‌متحده و چهار الگوریتم یادگیری ماشین مختلف از جمله رگرسیون خطی چندمتغیره، رگرسیون ریج[2]، الگوریتم (XGBoost) و الگوریتم جنگل تصادفی بهره گرفته شده است. نتایج بدست آمده نشان می‌دهند که الگوریتم جنگل تصادفی عملکرد بهتری داشته و عملکرد آن را می‌توان با تنظیم ابرپارامترهای آن ارتقاء داد.

سرازوات[3] و همکاران (۲۰۲۳) اشاره کرد. این مقاله به بررسی عوامل تعیین‌کننده میزان ادعای خسارت سلامت از بیمه و پیش‌بینی آن می‌پردازد و هدفش را در اختیار قراردادن ابزاری در اختیار شرکت‌ها برای ارائه بیمه سلامت اعلام می‌کند. در این مطالعه از الگوریتم‌های مختلف دسته‌بندی استفاده شده و نتیجه گرفته می‌شود که یادگیری ماشین می‌تواند با عملکرد مناسب برای بررسی ادعاهای خسارت، شخصی‌سازی بیمه‌نامه‌های سلامت، و بسیاری موارد دیگر از جمله تقلب استفاده شود.

جعفر و همکاران[4] (۲۰۲۳) اشاره کرد که با استفاده از تکنیک‌های یادگیری عمیق ترکیبی[5] (DL) به دنبال پیش‌بینی و شناسایی بیماران پرهزینه[6] (HNHC) است. برای این منظور از سه مدل یادگیری

[1] Bau and Hanif
[2] Ridge Regression
[3] Saraswat
[4] Jaffar
[5] Deep learning
[6] High-need High-cost patients

عمیق شبکه عصبی واحد بازگشتی گیت‌دار[1] (GRU)، شبکه عصبی کانولوشنی کامل[2] (FCN)، و شبکه عصبی بازگشتی وانلا[3] (VRNN) و دو مدل ترکیبی (FCN-VRNN) و (GRU-FCN) استفاده شده است. پس از ارزیابی عملکرد این مدلها با معیارهای مختلف، نتایج بدست آمده نشان می‌دهند که مدلهای ترکیبی عملکرد بهتری در پیش‌بینی دارند.

کاشیک و همکاران[4] (۲۰۲۲) اشاره کرد. در این مقاله از الگوریتم‌های یادگیری ماشین برای پیش‌بینی هزینه‌های پرداختی بیمه سلامت در هند در سال‌های ۲۰۱۹ تا ۲۰۲۰ استفاده شده است. داده‌های پژوهش اطلاعاتی مربوط به بیمه‌گذاری، مانند سن، جنسیت، شغل، شرایط بهداشتی و تاریخچه بیماری‌ها را شامل می‌شد و از الگوریتم‌های مختلف اعم از شبکه‌های عصبی، درخت تصمیم و رگرسیون خطی استفاده شد. مدل نهایی این مقاله حاکی از آن است که الگوریتم شبکه عصبی با دقت حدود ۹۲ درصدی بهتر از دیگر الگوریتم‌ها توانایی پیش‌بینی دارد.

طاها و همکاران[5] (۲۰۲۲) با استفاده از اطلاعات ۵۰۸۸۲ بیمه‌نامه درمان شامل اطلاعات ۱۵ مشخصه از مشتریان به مزایای استفاده از تکنیک‌های یادگیری ماشین به‌جای روش های آماری سنتی برای تجزیه و تحلیل اشاره می‌کنند. این مقاله در ابتدا با استفاده از تکنیک‌های انتخاب ویژگی از جمله تجزیه مولفه‌های اصلی و انتخاب ویژگی براساس همبستگی[6] بر مهمترین متغیرها و مشخصه‌ها برای پیش‌بینی درخواست‌های بیمه تمرکز کرده و سپس با الگوریتم کی-نزدیکترین همسایه و ماشین بردار پشتیبان به خوشه‌بندی و دسته‌بندی مشتریان می‌پردازد. نتایج بدست آمده نشان داد که الگوریتم ماشین بردار پشتیبان عملکرد بهتری در دسته‌بندی مشتریان بیمه درمان دارد. در این مقاله دقت پیش‌بینی در روش کی-نزدیکترین همسایه به ۷۳/۹ درصد و در روش ماشین بردار پشتیبان به ۸۰ درصد رسید.

---

[1] Gated Recurrent Unit
[2] Fully Convolutional Network
[3] Vanilla Recurrent Neural Network
[4] Kaushik et al.
[5] Taha et al.
[6] correlation

ون و همکاران[1] (۲۰۲۱) معتقدند که با داده‌کاوی که از الگوریتم‌های یادگیری ماشین و هوش مصنوعی استفاده می‌کند، می‌توان بررسی دقیق‌تری از رفتار و ترجیحات مشتریان و در نتیجه طراحی محصول بهتر و استراتژی بازاریابی موثرتری داشت. در این مقاله از دو نوع داده استفاده شده است. مجموعه داده‌های کمی از جمله اطلاعاتی مانند سابقه بیمه‌ای، سن، درآمد، تاریخ انقضای بیمه و سایر مشخصه‌های مشتریان و داده‌های کیفی شامل ملاحظاتی که توسط مشاوران بیمه در رابطه با رفتار و ترجیحات مشتریان در جلسات مشاوره ثبت کرده‌اند (نظرات مشتریان در مورد محصولات بیمه، شکایات، پیشنهادات و نیازهای خاص). در این مقاله ابتدا از الگوریتم خوشه‌بندی کی-میانگین و همچنین مدل مخلوط گوسی[2] برای تقسیم داده‌ها به دسته‌های مختلف با توجه به شباهت‌هایی که بین آن‌ها وجود دارد استفاده شد. سپس برای دسته‌بندی و تجزیه و تحلیل داده‌ها از مدل درخت تصمیم و همچنین شبکه عصبی بازگشتی (LSTM[3]) استفاده شد که امکان پیش‌بینی رفتار مشتریان در آینده را فراهم می‌کند.

راوات[4] (۲۰۲۱) با استفاده از داده‌هایی که شامل اطلاعات مربوط به پرونده‌های بیمه، اطلاعات مالی و اطلاعات درمانی بیمه‌گذاران در هند است به کاربرد یادگیری‌ماشین و مصورسازی[5] در صنعت بیمه پرداخته است. در این مقاله از الگوریتم شبکه عصبی برای پیش‌بینی و مدیریت ریسک، و از درخت تصمیم برای انتخاب بهترین پوشش بیمه و تعیین بهترین قیمت بیمه استفاده شده است. نتایج بدست آمده نشان می‌دهند که مصورسازی داده سبب درک بهتر بیمه‌گر و بیمه‌گذار شده و استفاده از یادگیری‌ماشین سبب افزایش ۲۸ درصدی سود شرکت بیمه گر می‌شود.

سانتوس و همکاران[6] (۲۰۲۱) با استفاده از داده‌های سلامت ملی پرتغال، الگوریتم‌های یادگیری ماشین را برای بهبود داده‌های موجود و پیدا کردن دسته‌بندی و تقسیم‌بندی بهتر در افرادی که بیمه خصوصی ندارند، به کار گرفتند. با استفاده از الگوریتم های خوشه‌بندی، افراد بدون بیمه خصوصی

[1] Wen et al.
[2] Gaussian Mixture Model
[3] Long-Short Term Memory
[4] Rawat
[5] Visualization
[6] Santos et al.

در این کشور بر اساس مشخصه‌های مشابه در دسته‌های مختلف تقسیم شدند و نتایج بدست آمده حاکی از این بود که افراد بدون بیمه خصوصی معمولاً از خدمات سلامت بیشتری استفاده می کنند و هزینه‌های درمانی بیشتری دارند. این مقاله نشان داد که استفاده از الگوریتم‌های یادگیری ماشین و داده‌های سلامت ملی می‌تواند در بهبود شناخت افراد بدون پوشش بیمه و محاسبه هزینه‌های درمان آنها مؤثر باشد.

نانداپالا و همکاران[1] (۲۰۲۰) به بررسی روش‌های بخش‌بندی خرد[2] در صنعت بیمه سلامت و استفاده از الگوریتم‌های پیشرفته یادگیری ماشین برای تحلیل داده‌های پیچیده در این زمینه می‌پردازد. در این مقاله از الگوریتم‌های یادگیری ماشین در تحلیل داده‌های پیچیده از جمله الگوریتم‌های خوشه‌بندی مانند کی-میانگین و دی‌بی اسکن و همچنین الگوریتم‌های شبکه‌های عصبی مانند پرسپترون چندلایه[3] و درخت تصمیم برای تحلیل داده‌ها استفاده شد و نشان داده شد که رویکرد داده‌کاوی و روش‌های یادگیری ماشین منجر به دستیابی به نتایج بهتری نسبت به روش‌های سنتی می‌شود. همچنین استفاده از بخش‌بندی خرد برای دسته‌بندی و شناسایی رفتار مشتریان باعث افزایش کیفیت سرویس‌دهی و حفظ رضایت مشتریان در صنعت بیمه سلامت می‌شود.

زاکو[4] (۲۰۱۹) در مقاله ای به بررسی کاربردهای یادگیری ماشین و دسته‌بندی مشتریان برحسب داده‌های سوابق پزشکی آن‌ها پرداخت. در این مقاله از داده‌های پایگاه داده سلامت پرتغال طی سال‌های ۲۰۱۶ تا ۲۰۱۷ استفاده شده و الگوریتم‌های خوشه‌بندی مکانی بر مبنای چگالی در کاربردهای دارای نویز[5] و کی-میانگین بکار گرفته شده است. نتایج بدست آمده بیمه‌گذاران سلامت را به ۱۱ دسته شامل ۶ دسته اصلی و ۵ دسته خرد تقسیم می‌کند که در دسته‌های خرد البته جمعیت بسیار کوچک بوده و روند ثابتی نیز مشاهده نشد. در شش دسته اصلی اما بسیاری از عوامل تأثیرگذار شناخته شدند و مدل توانست عملکرد مناسبی در پیش‌بینی داشته باشد، اگرچه همچنان

[1] Nandapala et al.
[2] Micro-Segmentation
[3] Multilayer perceptron
[4] Zaqueu
[5] Density Based Spatial of Application with Noise (DBSCAN)

دقت پایین بود و این مقاله نتیجه می‌گیرد که در خصوص بیمه سلامت نمی‌توان مشتریان را به گروه‌های مختلف تقسیم کرد و باید فردیت آن ها را در نظر گرفت.

پاندی و همکاران[1] (۲۰۱۸) به تشخیص و پیش‌بینی تقلب در بیمه‌های سلامت با استفاده از روش‌های داده‌کاوی و مدل‌سازی یادگیری ماشین پرداخته‌اند. داده‌های مورد استفاده در این مقاله شامل اطلاعات بیماران از جمله سن، جنسیت، نوع بیمه‌نامه، نوع درمان، مراکز درمانی و محل ارائه خدمات، پزشکان، بیمه‌گزاران و شرکت‌های بیمه بوده است. برای این منظور نیز از روش‌های داده‌کاوی مانند درخت تصمیم، بیز ساده، شبکه‌های عصبی و ماشین بردار پشتیبان استفاده شده است. نتایج بدست آمده نشان می‌دهد که روش‌های داده‌کاوی و مدل‌سازی پیش‌بینی می‌تواند به تشخیص تقلب در بیمه‌های سلامت کمک کند و باعث کاهش خسارت‌های بیمه‌ای شود. دقت مدل‌های پیش‌بینی برای تشخیص تقلب در بیمه‌های سلامت با استفاده از الگوریتم‌های مختلف در این مقاله بین ۸۰٪ تا ۹۵٪ بوده است. در این مقاله در نهایت پیشنهاد می‌شود که استفاده از روش‌های داده‌کاوی و مدل‌سازی پیش‌بینی برای تحلیل داده‌های بیمه‌ای به عنوان یک راهکار موثر برای مبارزه با تقلب در بیمه‌های سلامت در نظر گرفته شود.

پنگ و همکاران[2] (۲۰۱۸) بر اساس اطلاعات بیمارانی که در سامانه بیمه‌ی درمانی چین ثبت‌نام شده‌اند شامل ویژگی‌هایی همچون سن، جنسیت، شغل، شرایط بهداشتی، تاریخ تشخیص بیماری، هزینه‌ی درمانی و همچنین با کمک الگوریتم‌های یادگیری ماشین تلاش می‌کنند تا پیش‌بینی از میزان مبلغ بیمه‌ی درمانی ارائه دهند. در این مقاله از الگوریتم‌های یادگیری ماشین از جمله مدل‌های خطی و غیرخطی مانند رگرسیون لجستیکی، شبکه‌های عصبی و درخت تصمیم استفاده شده که توانایی بیشتری در تعامل با داده‌های پیچیده دارند و تخمین بهتری از مبلغ بیمه ارائه می‌دهند. نتایج تجربی این مقاله نشان می‌دهد که چارچوب پیشنهادی می‌تواند به‌صورت موثری به پیش‌بینی مبالغ بیمه‌ی درمانی بیماران کمک کند. مدل‌های ارائه شده در این مقاله توانستند با دقت ۸۶ درصد

[1] Pandey et al.
[2] Peng et al.

مشتریان زیانده را شناسایی کنند تا به این وسیله حق بیمه مناسب با ریسک وارده از آن‌ها دریافت شود.

مطالعات داخلی و خارجی بررسی شده به خوبی نشان می‌دهند که تا چه حد داده‌کاوی می‌تواند در صنعت بیمه، شناخت مشتریان و افزایش سود بیمه‌های خصوصی مهم باشد. همچنین دیدیم که با توجه به محرمانه بودن اطلاعات و دسترسی بسیار سخت به داده‌ها به ویژه در حوزه بیمه سلامت و خدمات درمانی پژوهش‌های بسیار کمی در ایران انجام شده به‌ویژه آن‌که این پژوهش‌ها در خصوص دسته‌بندی مشتریان و به تبع آن بازاریابی داده‌محور نیازمند دسترسی به داده‌های خرد مشتریان شرکت‌های خصوصی بیمه است. از این جهت مقاله حاضر اهمیت دو چندانی پیدا می‌کند. علاوه بر اهمیت موضوعی، یکی دیگر از نوآوری‌های این مقاله استفاده از چندین الگوریتم مختلف است که بر قوام نتایج می‌افزاید. روش پژوهش، الگوریتم‌های مورد استفاده و داده‌ها در بخش بعدی با جزئیات ارائه خواهند شد.

۴. مدل تحقیق و روش برآورد

شروع تجربی و آماری هر پژوهش با معرفی متغیرهای پژوهش و بررسی آمار توصیفی داده‌هاست که این مهم در زیربخش اول از این بخش انجام می‌شود. پس از درک و شناخت داده، زیربخش دوم به پیش پردازش[1] یا آماده‌سازی داده‌ها اختصاص دارد. در این مرحله است که داده‌ها برای پردازش و تجزیه و تحلیل در مدل آماده می‌شوند. مرحلهٔ بعدی پژوهش تجربی نیز مدلسازی است. در زیربخش سوم انواع الگوریتم‌های مورد استفاده معرفی خواهند شد و یافته‌های مدل‌ها به بخش بعدی موکول خواهد شد.

۴-۱. داده‌های آماری و متغیرهای پژوهش

جامعه آماری این مقاله اطلاعات ۲۸۹۷ نفر از بیمه‌گذاران بیمه درمان انفرادی یک شرکت خصوصی از تاریخ آذر ماه ۱۴۰۰ تا آذر ماه ۱۴۰۱ را شامل می‌شود. این اطلاعات مشخصه‌هایی

[1] Dart Preprocessing

مختلفی را در بر می‌گیرد که از طریق پرسشنامه سلامت یا فرم اطلاعات بیمه‌شده در هنگام خریداری پرسیده و ثبت شده است. آمار توصیفی این متغیرها در جدول ۱ ارائه شده است.

جدول ۱. معرفی متغیرهای پژوهش و آمار توصیفی

نام متغیر	نوع متغیر	بیشینه	میانگین	کمینه
سودهی مشتری	کیفی اسمی (۰ و ۱)	۱	۰/۵۶	۰
سن (سال)	کمی پیوسته	۷۷	۳۲/۸	۰
جنسیت مرد	کیفی اسمی (۰ و ۱)	۱	۰/۲۷	۰
محل زندگی	کیفی اسمی	۳	۲/۳۲	۱
کاهش وزن شدید یا از کار افتادگی	کیفی ترتیبی (۰ و ۱)	۱	۰/۰۳۴	۰
اعتیاد به سیگار یا مواد	کیفی ترتیبی (۰ و ۱)	۱	۰/۱۱۳	۰
بیماری روانی	کیفی ترتیبی (۰ و ۱)	۱	۰/۰۳	۰
بیماری قلب یا خون یا گوارش	کیفی ترتیبی	۲	۰/۱	۰
بیماری خاص یا مصرف داروی خاص	کیفی ترتیبی	۲	۰/۱۳	۰
سابقه بستری یا جراحی	کیفی ترتیبی	۴	۰/۲۶	۰
بیماری موروثی	کیفی ترتیبی (۰ و ۱)	۱	۰/۰۸۱	۰

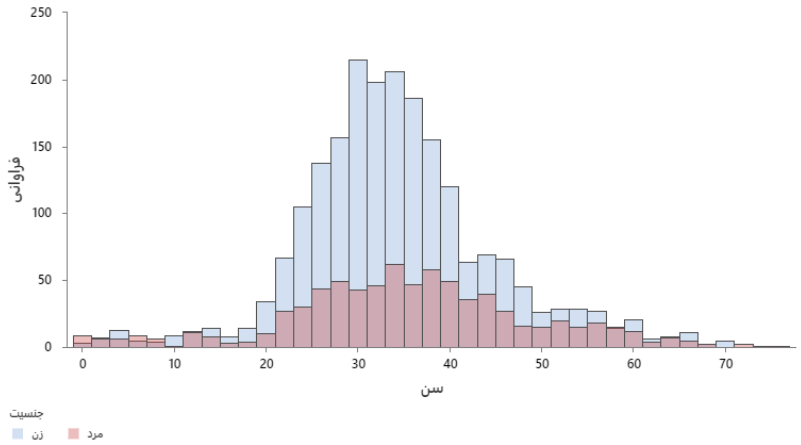
مأخذ: محاسبات تحقیق

سوددهی مشتری متغیر پاسخ این پژوهش است که به دنبال توضیح و پیش‌بینی آن هستیم. در محاسبات بیمه‌ای، مبلغ حق‌بیمه باید به گونه‌ای تعیین شود تا بتواند میزان خسارت یا هزینه درمان احتمالی را پوشش دهد. بر این مبنا می‌توان مشتریان را به دو دسته سودده و زیان‌ده تقسیم کرد. مشتریان سودده به افرادی اطلاق می‌شود که حق‌بیمه پرداختی ماهانه آن‌ها بیشتر از متوسط هزینه درمان ماهانه آن‌ها باشد. و بالعکس، مشتریان زیان‌ده افرادی هستند که اگر حق‌بیمه ماهانه آن‌ها را از متوسط ماهانه هزینه درمانشان کسر کنیم به یک عدد منفی خواهیم رسید یا به‌عبارت دیگر هزینه درمان و خسارت بیشتری در مقایسه با حق‌بیمه‌های پرداختی به شرکت بیمه وارد کرده‌اند. در جامعه آماری مورد بررسی ۵۶ درصد از مشتریان سودده و ۴۴ درصد زیان‌ده شناخته شدند. دلیل در نظر گرفتن مبالغ ماهانه از آن جهت است که افراد در مقطع مورد بررسی آذر ماه ۱۴۰۰ تا آذر ماه ۱۴۰۱

تعداد ماه‌های متفاوتی را از شروع بیمه‌نامه سپری کرده بودند و طبیعتاً امکان مقایسه هزینه درمانشان نبود مگر این‌که واحد مقایسه به ماه تعریف گردد.

برای توضیح و پیش‌بینی زیان‌ده یا سودده بودن مشتریان بیمه درمان از اطلاعات دموگرافیک و پرسشنامه سلامت مشتریان استفاده شده است که عبارتند از:

- سن مشتری: سن یکی از مشخصه‌های فردی است که می‌تواند بر تقاضا و نیازهای درمانی افراد تأثیرگذار باشد. برای مثال، با بالا رفتن سن احتمال ابتلا به بیماری‌های مزمن (دیابت و بیماری قلبی و عروقی)، نیاز به درمان طولانی‌مدت و مصرف دارو، نیاز به بستری و جراحی و ... افزایش می‌یابد. در نتیجه، سن تأثیر قابل توجهی بر هزینه بیمه درمان دارد (پناهی و همکاران ۱۳۹۳، میرزایی و همکاران، ۱۳۹۶). نمودار ۱ توزیع سنی مشتریان بیمه درمان را برای زنان و مردان نمایش می‌دهد.



شکل ۱. توزیع سنی مشتریان بیمه درمانی

مأخذ: محاسبات تحقیق

- جنسیت: در جامعه آماری مورد بررسی بیشتر افراد تحت پوشش را زنان با ۷۴/۴ درصد تشکیل می‌دهند. جنسیت نیز به عنوان یکی از مهمترین متغیرهای توضیح دهنده هزینه‌های درمان همواره مطرح بوده است (پناهی و همکاران ۱۳۹۳).

- محل زندگی: محل زندگی نیز به عنوان یکی از عوامل مهم و تاثیرگذار بر بیمه درمان مطرح است. برای مثال، هزینه‌های درمان در مناطق شهری به دلایل مختلف از جمله دسترسی بیشتر به مراکز درمانی، بالاتر بودن درآمد و در نتیجه تقاضا برای خدمات درمانی می‌تواند بالاتر از مناطق روستایی باشد. محل سکونت بر اساس کد شعبه صادرکننده در سه دسته ۱ (۲۰/۳ درصد از افراد)، ۲ (۲۷/۴ درصد از افراد) و ۳ (۵۲/۲ درصد از افراد) توسط شرکت بیمه ثبت می‌شوند. دسته ۱، ۲ و ۳ به ترتیب با شعب موجود در شهرهای کوچک، کلان‌شهرها و مناطق روستایی متناظرند.
- اطلاعات پرسشنامه سلامت: پرسشنامه سلامت ابزاری است برای غربال کردن مشتریان و رفع اطلاعات نامتقارن و بنابراین تاثیر آن بر بیمه درمانی می تواند قابل توجه باشد. در این قسمت اطلاعات مختلفی ثبت شده که از آن‌ها به عنوان متغیر توضیحی استفاده شده است.
  - کاهش وزن شدید یا ازکارافتادگی: از کارافتادگی به معنای خانه‌نشینی و عدم توانایی انجام کار است که حداقل به مدت یک ماه اتفاق افتاده باشد و کاهش وزن شدید به موارد ناخواسته کاهش وزن حداقل ۱۰ کیلو در یک ماه گفته می‌شود. در جامعه مورد بررسی ۹۶/۶ درصد از افراد چنین تجربه‌ای را نداشته و این متغیر برای آن‌ها مقدار ۰ گرفته است. متوسط خسارت درمان ماهانه برای افرادی که ازکارافتادگی یا کاهش وزن شدید داشته‌اند در مقطع زمانی مورد بررسی ۱۱۱/۸ میلیون ریال و حدود ۳/۲ برابر سایرین بوده است.
  - اعتیاد به سیگار یا مواد مخدر: در جامعه مورد بررسی بیش از ۱۱/۳ درصد از افراد به سیگار یا مواد مخدر اعتیاد داشته که این متغیر برای آن‌ها مقدار ۱ گرفته و برای سایرین یعنی ۸۸/۷ درصد از جامعه مقدار صفر دارد. متوسط خسارت درمان ماهانه برای افراد دارای اعتیاد به سیگار یا مواد در مقطع زمانی مورد بررسی ۱۰۷/۴۷ میلیون ریال و حدود ۳/۵ برابر سایرین بوده است.
  - سابقهٔ بیماری روانی: در جامعه مورد بررسی بیش از ۳ درصد از افراد و مشتریان سابقه بیماری روانی دارند که برای آن‌ها این متغیر مقدار ۱ گرفته و برای ۹۷ درصد دیگر از

جامعه بدون سابقه بیماری روانی مقدار صفر دارد. متوسط خسارت درمان ماهانه برای افراد دارای سابقه بیماری روانی در مقطع زمانی مورد بررسی حدود ۹۹ میلیون ریال و حدود ۲/۵ برابر سایرین بوده است.

- سابقه بیماری قلبی یا گوارش: در جامعه مورد بررسی حدود ۸/۶ درصد از افراد و مشتریان سابقه بیماری قلبی یا گوارش دارند که برای آن‌ها این متغیر مقدار ۱ گرفته و برای ۹۰ درصد دیگر از جامعه بدون سابقه بیماری قبلی یا گوارش مقدار صفر دارد. حدود ۱/۴ درصد از جامعه نیز سابقه هر دو بیماری را داشته‌اند که برای آنها این متغیر مقدار ۲ گرفته است. متوسط خسارت درمان ماهانه برای مردان و زنان دارای سابقه بیماری قلبی و گوارش در مقطع زمانی مورد بررسی به ترتیب بیش از ۹۹/۳ و ۷۶/۱ میلیون ریال بوده که حدود ۳/۴ برابر متوسط خسارت درمان ماهانه مردان و ۲/۲ برابر متوسط خسارت درمان ماهانه زنان بدون سابقه بیماری قلبی و گوارش بوده است. متوسط خسارت درمان ماهانه برای مردان و زنانی که سابقه هر دو بیماری را داشته‌اند بسیار بالاتر و به ترتیب حدود ۱۴۲/۹ و ۱۴۶/۳ میلیون ریال بوده است.
- سابقهٔ بیماری‌های خاص: در جامعه مورد بررسی حدود ۸۹/۱ از افراد و مشتریان بدون سابقه بیماری خاص بوده‌اند که برای آن‌ها این متغیر مقدار ۰ گرفته، حدود ۸/۹ درصد سابقه یک بیماری خاص و حدود ۲ درصد نیز سابقهٔ ۲ بیماری خاص و بیشتر داشته که این متغیر برای این گروه‌ها به ترتیب مقدار ۱ و ۲ گرفته است. متوسط خسارت درمان ماهانه برای مردان با دو بیماری خاص حتی تا ۱۶۶/۷ میلیون ریال رسیده که بیش از ۵/۵ برابر متوسط خسارت درمان ماهانه مردان بدون سابقه است. برای زنان این پراکندگی کمتر است.
- سابقهٔ بستری یا جراحی: در جامعه مورد بررسی حدود ۷۹/۱ از افراد و مشتریان هیچ سابقه جراحی یا بستری نداشته‌اند که برای آن‌ها این متغیر مقدار ۰ گرفته است. در میان سایر افراد جامعه حدود ۱۶/۳ درصد سابقه یکبار جراحی یا بستری، حدود ۳/۸ درصد سابقهٔ دوبار جراحی و بستری، حدود ۰/۶ درصد از افراد تجربه سه‌بار و حدود ۰/۲

درصد از مشتریان نیز سابقه چهار بار و بیشتر بستری و جراحی داشته‌اند که برای آنها این متغیر به ترتیب مقادیر ۱، ۲، ۳ و ۴ گرفته است. متوسط خسارت درمان ماهانه برای زنان با سه بار سابقه بستری و جراحی حتی به ۲۰۰/۸ میلیون ریال رسیده که بیش از ۶ برابر متوسط خسارت درمان ماهانه زنان بدون سابقه بستری و جراحی است.

- سابقهٔ بیماری موروثی: در جامعه مورد بررسی حدود ۹۱/۹ از افراد و مشتریان هیچ سابقه بیماری موروثی نداشته‌اند که برای آن‌ها این متغیر مقدار ۰ گرفته و در مقابل حدود ۸/۱ درصد سابقه بیماری موروثی داشته‌اند که برای آنها این متغیر مقدار ۱ گرفته است. متوسط خسارت درمان ماهانه برای مردان و زنان دارای سابقه موروثی به ترتیب به ۹۸/۱ و ۶۴ میلیون ریال می‌رسد که به ترتیب حدود ۳/۳ برابر و ۱/۷ برابر متوسط خسارت درمان ماهانه مردان و زنان بدون سابقه بیماری موروثی است.

۲-۴. آماده‌سازی و پیش‌پردازش داده‌ها

در فرآیند آماده‌سازی داده‌ها، نخستین کار پاکسازی مجموعه داده‌ها از مقادیر گمشده و تکراری است. داده‌های مورد استفاده در این مقاله در سه پایگاه داده مجزا یکی اطلاعات خسارت، دیگری داده‌های پرسشنامه سلامت، و سومی مشخصات افراد تحت پوشش قرار داشت که البته با کلید اصلی شماره پرونده به یک دیگر مرتبط بودند. برای آماده‌سازی داده‌ها، ابتدا پایگاه داده مربوط به پرسشنامه سلامت که به صورت سطری ثبت شده بود با استفاده از زبان برنامه نویسی پایتون به ستونی تبدیل شد و سطرهای با مقادیر گمشده حذف شدند. پس از حذف مقادیر گمشده و تکراری، سه پایگاه داده مجزا به کمک کلید اصلی یعنی شماره پرونده به یک پایگاه داده تبدیل شد.

یکی دیگر از کارهایی که در فرآیند آماده‌سازی داده‌ها انجام می‌شود عبارت است از آزمون نرمال بودن توزیع داده‌ها و همچنین شناسایی مقادیر پرت و نویز در مجموعه داده. نرمال بودن توزیع داده‌ها برای متغیرهایی که اطلاعات آن‌ها گسسته است که قاعدتاً بی‌معنی است و می‌توان این ویژگی را تنها در خصوص دو متغیر سودآوری مشتریان (پیش از تبدیل آن به یک متغیر دودویی)

و سن انجام داد. نتایج آزمون اندرسون دارلینگ[1] در سطح ۵ درصد فرضیه نرمال بودن داده‌های متغیرهای سن و سودآوری را نیز رد می‌کند و بنابراین می‌توان نتیجه گرفت که هیچ یک از متغیرها بصورت نرمال توزیع نشده‌اند. چون توزیع ناهمگن می‌تواند نتایج را تحت تأثیر قرار دهد و همچنین از آنجائی‌که پایه و اساس برخی از روش‌های آماری بر نرمال بودن توزیع‌ها بنا شده، متغیر سن با استفاده از تبدیل Box-Cox نرمالسازی شدند. برای متغیر سودآوری که مقادیر منفی دارد نمی‌توان از این تبدیل استفاده کرد و به همین دلیل از تبدیل جانسون[2] استفاده شد. در ادامه مقاله از این داده‌های نرمال شده در مدلسازی استفاده شده است.

برای شناسایی و حذف داده‌های پرت نیز بر متغیر پاسخ یعنی سودآوری مشتریان متمرکز شده و از نمودارهای جعبه‌ای[3] و فاصله ژاکارد[4] استفاده شد. شیوه کار به این صورت است که ابتدا با نمودار جعبه‌ای داده‌هایی که از چارک اول و سوم بیش از ۱/۵ برابر دامنه میان چارکی فاصله دارند را به عنوان نقاط کاندیدا شناسایی می‌کنیم و سپس در خصوص مقادیر خارج از این محدوده بر اساس فاصله ژاکارد تصمیم گیری نهایی گرفته خواهد شد. فاصله ژاکارد یکی از روش‌های محاسبه فاصله بین دو مجموعه یا محاسبه میزان شباهت و تفاوت دو مجموعه است. سپس با استفاده از فاصله ژاکارد، ماتریس گاور[5] تشکیل داده می‌شود که یک ماتریس مربعی است و فاصله بین هر جفت رکورد در یک مجموعه داده را نمایش می‌دهد. در نهایت، از الگوریتم کی-مدوید[6] برای شناسایی داده پرت استفاده شد که در واقع یک الگوریتم خوشه‌بندی داده‌های یک مجموعه است. نتیجهٔ خوشه‌بندی این شد که ۱۲۷ داده به عنوان داده پرت شناسایی و حذف شدند.

یکی دیگر از کارهایی که برای آماده‌سازی داده‌ها انجام می‌شود عبارت است از هم‌مقیاس-سازی[7] که به عملیات تغییر مقیاس داده‌های یک مشخصه گفته می‌شود. دلیل تغییر مقیاس این است که برخی از الگوریتم‌های یادگیری ماشین به واحد اندازه‌گیری وابسته‌اند و مقیاس داده‌های تاثیر

[1] Anderson-Darling
[2] Johnson Transformation
[3] Box-Plot
[4] Jaccard Distance
[5] gower matrix
[6] K-Medoid
[7] Feature Scaling

مهمی بر خروجی آن‌ها می‌گذارد. به این ترتیب به جهت بهبود کارایی الگوریتم‌های یادگیری ماشین است که داده‌ها پیش از استفاده در مدل نهایی در اصطلاح هم‌مقیاس سازی می‌شوند. در این مقاله لزومی به استانداردسازی متغیرهای ترتیبی نیست و تنها متغیر سن است که با کسر میانگین و تقسیم بر انحراف‌معیار در بازه [1,1-] استاندارد شد.

۳-۴. الگوریتم‌های مدلسازی

این مقاله به دنبال پیش‌بینی زیان‌ده یا سودده بودن مشتریان بیمه درمان و همچنین شناسایی مهمترین عوامل یا مشخصه‌های توضیح دهنده آن است. از آنجائی‌که برچسب زیان‌دهی یا سوددهی مشتریان مشخص است، در نتیجه از الگوریتم‌های تحت نظارت یادگیری ماشین استفاده خواهد شد. بطور مشخص در این مقاله از پنج الگوریتم طبقه‌بندی بیز ساده، رگرسیون لجستیک، جنگل تصادفی، درخت تصمیم و ماشین بردار پشتیبان استفاده می‌شود که در ادامه با جزئیات بیشتر معرفی می‌شوند.

۱-۳-۴. الگوریتم درخت تصمیم

در این الگوریتم بر اساس مشخصه‌ها سوالاتی مطرح شده و دسته بندی رخ می‌دهد و برای نمایش این دسته‌بندی نیز یک درخت تصمیم‌گیری متشکل از گره‌ها[1] و شاخه‌ها[2] ایجاد می‌شود. گره ها نشان دهنده مشخصه‌هایی هستند که بر اساس آن‌ها تفکیک و دسته‌بندی صورت گرفته است. شاخه‌ها نیز از تقسیم داده ها ظاهر می‌شوند. معیارهای متعددی برای تصمیم‌گیری در مورد محل تقسیم و میزان تقسیم استفاده می‌شوند که از آن جمله می‌توان به آنتروپی[3]، شاخص جینی[4]، مربع-کای و شاخص حدأکثر عمق درخت[5] اشاره کرد (السـاقیر و همکاران[6]، ۲۰۱۷). آنتروپی ناخالصی مقادیر نمونه را اندازه می‌گیرد و می‌توان آن را از معادله ۱ محاسبه کرد جایی که S مجموعه داده، c کلاس‌ها و $p(c)$ نسبت داده‌های متعلق به کلاس c به کل داده‌ها هستند. مقادیر آنتروپی بین ۰ و ۱ قرار می‌گیرند. اگر تمام داده‌ها به یک کلاس تعلق بگیرند آنتروپی صفر و اگر

[1] Nodes
[2] Branch
[3] Entropy
[4] Gini index
[5] Maximum Depth of Tree
[6] Alsagheer et al.

نیمی از نمونه‌ها به یک کلاس و نیمی دیگر در کلاس دیگری طبقه‌بندی شـــوند، آنتروپی به بالاترین حد خود یعنی ۱ خواهد رسید.

$$Entropy(S) = -\sum_{c \in C} p(c) \log_2(p(c)) \quad (۱)$$

برای انتخاب این‌که کدام مشخصه برای تقسیم انتخاب شود باید از مشخصه با کمترین مقدار آنتروپی استفاده شود. به تفاوت آنتروپی قبل و بعد از تقسیم در یک مشخصه معین نیز اطلاعات بدست آمده گفته می‌شود. مشخصه با بالاترین بهره اطلاعات بهترین تقسیم را ایجاد می‌کند.

در شاخص جینی، دو مشاهده که به طور تصادفی انتخاب شده‌اند ابتدا در یک کلاس طبقه‌بندی می‌شوند. ناخالصی جینی احتمال طبقه‌بندی نادرست داده‌های تصادفی بر اساس برچسب کلاس را اندازه می‌گیرد. اگر داده‌ها همه به یک کلاس تعلق داشته باشد، ناخالصی جینی صفر است. در واقع خالصی و ناخالصی همان احتمال موفقیت و شکست در دسته‌بندی است. مجموع مربع احتمال موفقیت و شکست $(p^2 + q^2)$ برای محاسبه جینی زیرگره‌ها استفاده می‌شود. هرچه ارزش جینی بیشتر باشد، ارزش همگنی بیشتر است. به طریق مشابه، شاخص مربع-کای طبق معادله ۲ از تفاوت بین مجموع مربعات برچسب دسته‌بندی شده با برچسب واقعی به دست می‌آید که در واقع همان «موفقیت» یا «شکست» است. هرچه مقدار مربع کای بیشتر باشد، مقدار اهمیت آماری انشعاب شاخه بیشتر است.

$$Chi - square = \sqrt{\frac{(Actual - Expected)^2}{Expected}} \quad (۲)$$

۲-۳-۴. الگوریتم جنگل تصادفی

الگوریتم جنگل تصادفی از ترکیب تعدادی درخت تصادفی[1] مستقل از هم تشکیل شده که هر درخت تصمیم با استفاده از یک مجموعه داده‌ها و مشخصه‌هایی که با تکنیک نمونه‌برداری تصادفی انتخاب شده‌اند، آموزش داده می‌شود. مکانیزم عمل الگوریتم جنگل تصادفی به این صورت است که نخست، تعداد درخت تصمیم در جنگل تصادفی تعیین می‌شود. دوم، برای هر درخت تصمیم با

[1] Random Trees

استفاده از تکنیک‌های نمونه‌برداری (معمولاً با جایگزینی تصادفی نمونه‌ها) یک مجموعه داده آموزشی ایجاد می‌شود. سوم، با توجه به مشخصه‌ها و داده‌ها، هر درخت ساخته و بهترین تقسیم‌بندی برای هر گره از درخت انتخاب می‌شود. چهارم، تمام درختان تصمیم ساخته شده به عنوان جنگل تصادفی ترکیب می‌شوند. با ورود یک مشاهده جدید برای طبقه‌بندی یا پیش‌بینی، تمام درختان تصمیم به طور مستقل پاسخی تولید می‌کنند. و در پایان برای تصمیم نهایی در مورد طبقه‌بندی یا پیش‌بینی مشاهده جدید، اکثریت پاسخ‌های تولید شده توسط درختان تصمیم در جنگل تصادفی یعنی کلاس با بیشترین تعداد رأی انتخاب می‌شود. جنگل تصادفی به دلیل مزایایی مانند قدرت پیش‌بینی، مقاومت در برابر بیش‌برازش و توانایی استفاده از مشخصه‌های زیاد، به یکی از روش‌های محبوب در حوزه یادگیری ماشین تبدیل شده است (هان و همکاران[1]، ۲۰۲۲).

۳-۳-۴. الگوریتم بیز ساده

دسته‌بندی‌های بیزی[2] یک نوع دسته بندی‌های آماری هستند که می‌توانند احتمال عضویت در کلاس‌های مختلف (احتمال تعلق یک مشاهده به یک کلاس خاص) را پیش‌بینی کنند. دلیل نام‌گذاری این الگوریتم به بیز ساده این است که بر فرض استقلال‌شرطی کلاس[3] استوار است که باعث ساده‌سازی محاسبات می‌شود (هان و همکاران[4]، ۲۰۱۲). الگوریتم بیز ساده چند مرحله را شامل می‌شود. نخست، هر مشاهده با بردار n بعدی $\mathbf{X}=(x_1,x_2,x_3,...x_n)$ نمایش داده می‌شود که x_i مقدار مشخصه خاصه A_i است. دوم، تعداد m برچسب کلاس با نام های $C_1,C_2,C_3,...,C_m$ موجود است و الگوریتم قرار است مشاهده X را به کلاسی تعلق دهد که بیشترین احتمال پسین به شرط $\mathbf{X}$ را داشته باشد. فرضیه بیز در معادله ۳ نشان داده شده است.

$$P(C_i|\mathbf{X}) = \frac{\left(P(\mathbf{X}|C_i)\right)P(C_i)}{P(\mathbf{X})} \qquad (۳)$$

[1] Han et al.
[2] Bayesian Classifires
[3] Class Conditional Independence
[4] Han et al.

اگر مقدار $P(\mathbf{X})$ برای کلیه کلاس‌ها ثابت باشد و علاوه بر این، وقتی احتمالات پیشین ناشناخته باشد فرض احتمال یکسان $P(C_1)=P(C_2)=...=P(C_m)$ را به کار بگیریم، این امر به معنی دسته‌بندی مشاهدات در کلاس‌ها به صورتی است که هدف تنها بیشینه کردن $P(C_i|\mathbf{X})$ باشد. اینجاست که فرض استقلال‌شرطی کلاس‌ها به ویژه وقتی تعداد مشخصه‌ها زیاد باشد منجر به ساده‌سازی شده و می‌توان معادله ۴ را بیشینه کرد.

$$P(\mathbf{X}|C_i) = \prod_{i=1}^{n} P(x_k|C_i) \quad , i = 1,2,\dots,m \quad , k = 1,2,\dots,n \qquad (۴)$$

دسته‌بند بر چسب کلاس $\mathbf{X}$ را C_i می‌داند اگر و تنها اگر $P(\mathbf{X}|C_i) > P(\mathbf{X}|C_j)\ for\ 1 \le j \le m, j \ne i$، یعنی بیشترین احتمال پسین به شرط $\mathbf{X}$ را داشته باشد. در دسته‌بندی بیز ساده مراحل بالا تکرار می‌شود. سادگی و دقت بالا مهمترین مزیت بیز ساده هستند اما نقطه ضعف اصلی این الگوریتم به فرض استقلال‌شرطی کلاس‌ها باز می‌گردد که ممکن است در مواردی درست نباشد (سن و همکاران[1]، ۲۰۲۰؛ جادها و شان[2]، ۲۰۱۶).

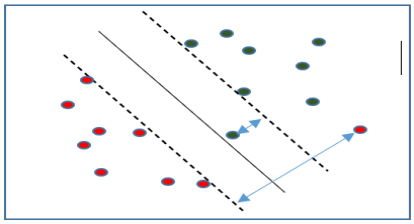
۴-۳-۴. الگوریتم ماشین بردار پشتیبان (SVM)

هدف از الگوریتم بردار پشتیبان (SVM) یافتن یک ابر صفحه در یک فضای N بعدی است که به طور مشخص نقاط داده را دسته‌بندی می‌کند. به آخرین داده هر دسته، بردار پشتیبان و به معادله خط جداکننده‌ی دسته‌ها، ابرصفحه گفته می‌شود. ابعاد ابرصفحه به تعداد مشخصه‌ها بستگی دارد. به عنوان مثال، اگر تعداد مشخصه ورودی دو باشد آنگاه ابرصفحه فقط یک خط و اگر تعداد مشخصه ورودی سه باشد آنگاه ابرصفحه به یک صفحه دو بعدی تبدیل می‌شود. این الگوریتم به دنبال یافتن بهترین معادله جداکننده دسته‌های مساله (کلاس‌های تابع هدف) از یکدیگر بوده و مکانیزم این جداسازی بر اساس بیشترین فاصله نزدیک‌ترین داده با خط معادله جداکننده است. در واقع، الگوریتم بردار پشتیبان به دنبال کمینه کردن رابطه ۵ است که در آن عبارت اول فاصله نزدیکترین نقطه تا خط جداکننده را اندازه می‌گیرد و عبارت دوم خطای در پیش‌بینی است. پارامتر C وزنی

[1] Sen et al.
[2] Jadhav & Channe

است که به خطا داده می‌شود و مقادیر پایین آن به یک حاشیهٔ نرم بزرگتر برای جداکننده مطابق نمودار ۲ منجر می‌شود.

$$\frac{1}{2}\|w\|^2 + c\sum_{i=1}^{m}\varepsilon_i \qquad (۵)$$



شکل ۲. خط جداکننده بردار پشتیبان با حاضیه‌های نرم

مأخذ: محاسبات تحقیق

البته بسیاری از داده‌ها به صورت خطی قابلیت جداسازی ندارند و به همین دلیل ازمفهومی به نام هسته یا کرنل[1] استفاده می‌شود که داده‌ها را به ابعاد بزرگتر منتقل می‌کند تا بتوان از دسته‌بندی خطی برای آن‌ها استفاده کرد (چاوهان[2] و همکاران، ۲۰۱۹).

۵-۳-۴. رگرسیون لجستیک

رگرسیون لجستیک یک روش دسته‌بند دوکلاسه است و اساس آن بر روش بیشینه درست‌نمایی[3] بنا شده است. علت نام‌گذاری این است که در این مدل برای دسته‌بندی داده‌ها از یک تابع لجستیکی[4] استفاده می‌شود. شیوه کار به این صورت است که برای هر مشاهده یک آزمایش برنولی با احتمال موفقیت P و عدم موفقیت برابر ۱-P تعریف می‌کنیم. چون احتمالات مقداری بین ۰ و ۱ می‌گیرند، به‌جای آن از شانس[5] استفاده می‌کنیم که برابر $\frac{P}{1-P}$ است. الگوریتم شانس‌ها در بازه $+\infty$ و $-\infty$ گسترده شده و وقوع آن مطابق رابطه ۶ به متغیرهایی مستقل بستگی دارد.

---

[1] Kernel
[2] Chauhan et al.
[3] Maximum Likelihood
[4] Logistics Function
[5] Odds

$$\ln\left(\frac{\mathrm{p}}{1-\mathrm{p}}\right) = b_0 + b_1x_1 + b_2x_2 + \cdots + b_kx_k = \beta X \qquad (۶)$$

تابع بیشینه درستنمائی تشکیل و با برآورد رگرسیون ضرایب بدست خواهند آمد (هیلی[1]، ۲۰۰۶). با در دست داشتن ضرایب می‌توان شانس‌ها و احتمالات (موفقیت و عدم موفقیت) را برآورد کرد که هر کدام متناظر با تعلق مشاهده به یک کلاس هستند. از مهمترین نقاط قوت رگرسیون لجستیک می‌توان به سرعت بالا و قابلیت تفسیرپذیری آن اشاره کرد (مالوف[2]، ۲۰۱۱).

۵. نتایج تجربی

هدف این مقاله بررسی مشتریان بیمه بر اساس خصوصیات فردی و سابقه سلامتشان و دسته‌بندی آن‌ها در دو گروه سودده و زیان‌ده است. این مهم با استفاده از ۵ الگوریتم یادگیری ماشین که در بخش قبلی معرفی شد انجام می‌شود. این بخش ابتدا بصورت جداگانه به پیاده‌سازی مدل‌ها و ارزیابی عملکرد آن‌ها در پیش‌بینی پرداخته و سپس با کنار هم قرار دادن آن‌ها به مقایسه عملکردی این الگوریتم‌ها و دیگر یافته‌ها متمرکز خواهد شد.

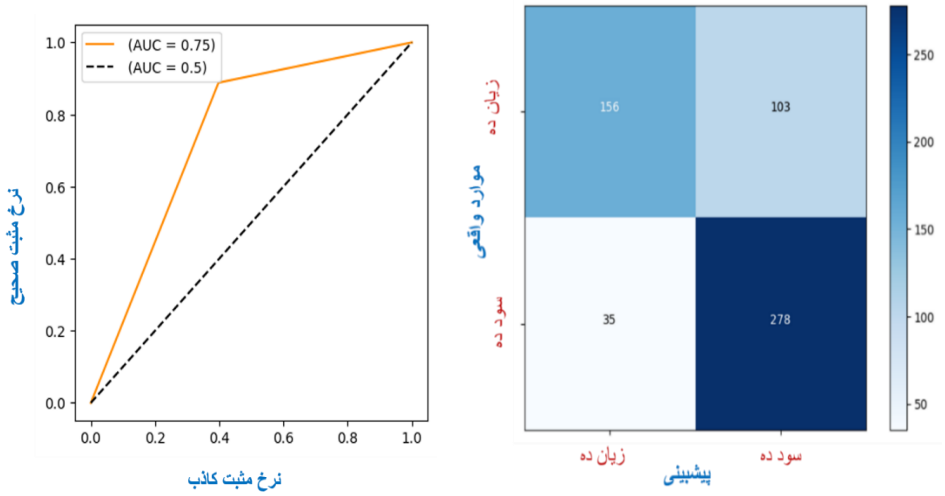
۵-۱. پیاده‌سازی و ارزیابی عملکرد مدل‌ها

۵-۱-۱. پیاده‌سازی و ارزیابی عملکرد الگوریتم درخت تصمیم

در پیاده‌سازی الگوریتم درخت تصمیم باید ابتدا دو پارامتر اصلی آن یعنی حداکثر عمق و شاخص انشعاب را مشخص کرد. برای این منظور، در ابتدا الگوریتم درخت تصمیم با استفاده از داده‌های آموزشی یک بار بر اساس شاخص انشعاب جینی و بار دیگر بر اساس شاخص آنتروپی در عمق‌های مختلف اجرا شد. در هر عمق یک مدل با استفاده از داده‌های آموزشی ساخته شده و سپس از داده‌های تست برای سنجش دقت آن مدل استفاده شد. در نهایت، پارامترهایی که با بالاترین دقت در مدل درخت تصمیم همراه شده‌اند به عنوان بهترین مقادیر پارامترها برای ساخت مدل نهایی درخت تصمیم انتخاب شدند. با بررسی انجام شده بهترین دقت و امتیاز-F در عمق ۶ بدست آمد و بنابراین، یک مدل نهایی درخت تصمیم با عمق ۶ طراحی و بکار گرفته شد.

[1] Healy
[2] Maalouf

عملکرد الگوریتم درخت تصمیم در دسته‌بندی داده‌های تست در نمودار ۳ تحت عنوان ماتریس اغتشاش[1] نمایش داده شده است. اگر دو طبقه سودده و زیانده را در نظر بگیریم، ماتریس اغتشاش از چهار قسمت مختلف سودده واقعی، زیانده واقعی، سودده کاذب و زیانده کاذب تشکیل شده است. سودده واقعی تعداد مشاهدات واقعاً سودده هستند که به درستی توسط مدل سودده پیش‌بینی شده‌اند؛ قسمت پایین و راست ماتریس. زیانده واقعی نیز به تعداد مشاهدات واقعاً زیانده گفته می‌شود که به درستی توسط مدل به عنوان زیانده پیش‌بینی شده‌اند؛ سمت چپ و بالای ماتریس. سودده کاذب در سمت راست و بالای ماتریس به مشاهدات واقعاً زیانده که به اشتباه توسط مدل به عنوان سودده پیش‌بینی شده‌اند گفته می‌شود. و در نهایت، زیانده کاذب در سمت چپ و پایین ماتریس مشاهدات واقعاً سودده هستند که به اشتباه توسط مدل به عنوان زیانده پیش‌بینی شده‌اند. در واقع در ماتریس اغتشاش، عناصر روی قطر اصلی مشاهداتی هستند که به درستی و عناصر روی قطر فرعی مشاهداتی هستند که توسط مدل به اشتباه پیش‌بینی شده‌اند.



شکل ۳. ماتریس اغتشاش و منحنی ROC الگوریتم درخت تصمیم

مأخذ: محاسبات تحقیق

[1] Confusion Matrix

از ماتریس اغتشاش می‌توان برای محاسبه معیارهای مختلفی مانند دقت[1]، صحت[2]، بازخوانی[3] و امتیاز F[4] استفاده کرد که هر کدام از جهتی عملکرد مدل را مورد ارزیابی قرار می‌دهند. دقت به نسبت تعداد کل مشاهداتی که به درستی پیش‌بینی شده‌اند به تعداد کل مشاهدات در مجموعه داده گفته می‌شود. به عنوان مثال در ماتریس اغتشاش الگوریتم درخت تصادفی، عناصر قطر اصلی که به درستی پیش‌بینی شده‌اند ۴۳۴ مورد بوده که ۷۵/۸۷ درصد از کل ۵۷۲ مشاهده را تشکیل می‌دهند. به‌عبارت دیگر، دقت الگوریتم درخت تصمیم برابر ۷۵/۸۷ درصد بوده است.

شاخص صحت نه بر کل مشاهدات که بر کلاس یا طبقه‌ای که هدف پیش‌بینی بوده متمرکز است. در مورد موضوع این مقاله به عنوان مثال، ما به دنبال پیش‌بینی مشتریان سودده هستیم و بر این کلاس تمرکز داریم. بر این اساس، شاخص صحت تعداد مشاهداتی که مدل سودده تشخیص می‌دهد را به نسبت کل مشاهدات به واقع سودده اندازه می‌گیرد. به‌عبارت دیگر، در ماتریس اغتشاش بر سطر پایین مشاهدات به واقع سودده تمرکز کرده و درصدی از آنها که مدل به درستی سودده پیش‌بینی کرده را اندازه می‌گیریم که برابر ۸۸/۸۲ درصد است. شاخص بازخوانی بر مشاهداتی که توسط الگوریتم سودده پیش‌بینی شده‌اند یعنی ستون دوم ماتریس اغتشاش تمرکز کرده و اندازه می‌گیرد که چه نسبتی از آن‌ها به واقع سودده بوده‌اند. بر اساس اطلاعات ماتریس اغتشاش، شاخص بازخوانی از تقسیم تعداد ۲۷۸ بر مجموع ۳۸۱ مشاهده ستون دوم بدست می‌آید که برابر ۷۲/۹۶ است. بین معیارهای صحت و بازخوانی یک بده‌بستان وجود دارد، به این صورت که تلاش برای بهبود یکی اغلب منجر به بدتر شدن معیار دوم می‌گردد. امتیاز F به عنوان یک معیار ارزیابی ترکیبی میانگین هارمونیک صحت و بازخوانی است که وزن یکسانی به هر دو معیار می‌دهد. طبق اطلاعات بالا، برای الگوریتم درخت تصمیم امتیاز F برابر ۰/۸ است.

دو معیار دیگر برای ارزیابی عملکرد مدل‌ها عبارتند از بالابری[5] و نرخ اشتباه طبقه‌بندی. بالابری توانایی مدل در پیش‌بینی یا دسته‌بندی را نسبت به یک برازش تصادفی می‌سنجد. برازش تصادفی از

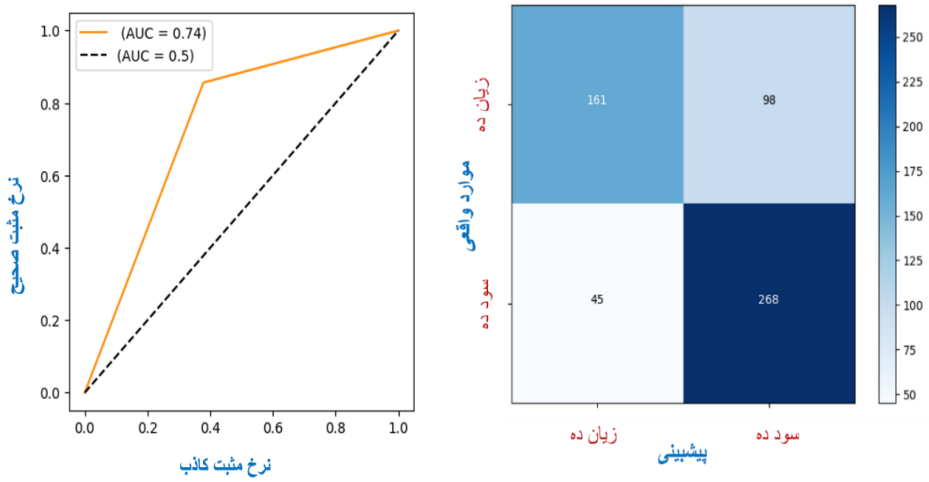
[1] Accuracy
[2] Precision
[3] Recall
[4] F-Score
[5] Lift score

تقسیم سطر دوم ماتریس اغتشاش (مشاهدات سودده) بر کل مشاهدات بدست می‌آید که برابر ۵۴/۷۲ درصد است. حالا توانایی مدل در تشخیص درست که همان شاخص بازخوانی است را اگر بر برازش تصادفی تقسیم کنیم، معیار بالابری مدل برابر ۱/۳۳ بدست می‌آید. معیار بالابری بالاتر از ۱ نشان‌دهنده عملکرد بهتر مدل نسبت به برازش تصادفی است. معیار نرخ اشتباه طبقه‌بندی[1] میزان اشتباه مدل در طبقه‌بندی را نسبت به کل داده‌ها بیان می‌کند. به عبارت دیگر و برعکس معیار دقت، نرخ اشتباه طبقه‌بندی از تقسیم عناصر روی قطر فرعی بر کل مشاهدات بدست می‌آید که بر اساس ماتریس اغتشاش الگوریتم درخت تصمیم برابر ۲۴/۱۲ درصد است. می‌توان از منحنی مشخصه عملکرد[2] (ROC) نیز برای ارزیابی عملکرد مدل استفاده کرد که در شکل ۳ نمایش داده شده است. سطح زیر منحنی یا شاخص AUC[3] بدست آمده برابر ۰/۷۵ است که نشان می‌دهد مدل با دقت ۷۵ درصد توانسته بین دو دسته‌بندی متفاوت (دسته مثبت سودده و دسته منفی زیان‌ده) تفاوت‌های موجود را تشخیص دهد.

۲-۱-۵. پیاده‌سازی و ارزیابی عملکرد الگوریتم جنگل تصادفی

در پیاده‌سازی الگوریتم جنگل تصادفی نیز باید ابتدا ابرپارامترها از جمله عمق درخت‌های جنگل بهینه شود. برای این منظور مشابه درخت تصمیم عمل کرده و با داده‌های آموزشی در عمق‌های مختلف مدل‌های مختلفی ساخته و سپس مورد ارزیابی قرار گرفت. در نهایت، بهترین دقت و امتیاز-F در عمق ۷ بدست آمد که به عنوان مدل نهایی انتخاب گردید. عملکرد الگوریتم جنگل تصادفی در دسته‌بندی داده‌های تست در ماتریس اغتشاش نمودار ۴ نمایش داده شده است.

[1] Misclassification Rate
[2] Receiver operating characteristic curve
[3] Area under the ROC Curve



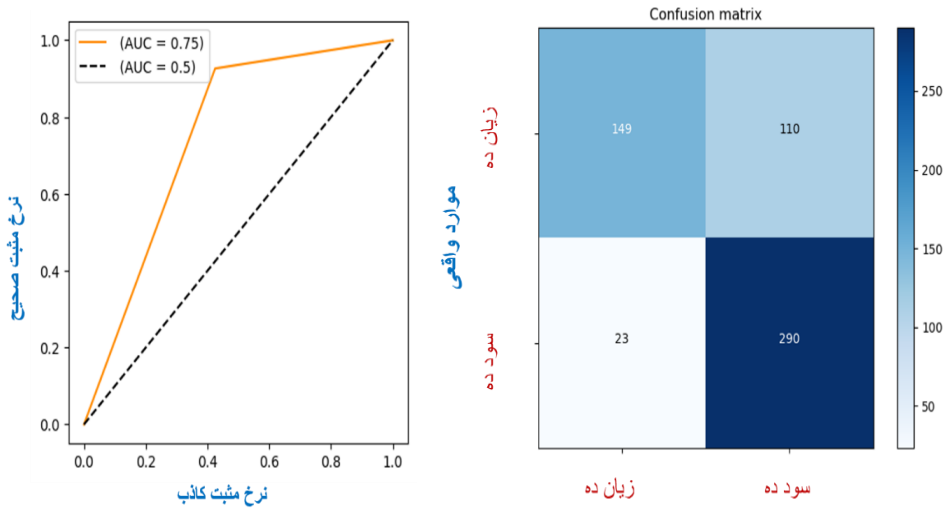
شکل ۴. ماتریس اغتشاش و منحنی ROC الگوریتم جنگل تصادفی

مأخذ: محاسبات تحقیق

معیارهای مختلف ارزیابی مدل جنگل تصادفی از ماتریس اغتشاش قابل محاسبه است. معیار دقت الگوریتم جنگل تصادفی ۷۷/۶ درصد بدست آمده که نشان می‌دهد این الگوریتم در دسته‌بندی مشتریان به این نسبت پیش‌بینی درستی داشته است. شاخص بازخوانی برابر ۸۶ درصد بدست آمده که به این معنی است که اگر یک مشتری واقعاً سودده باشد، ۸۶ درصد احتمال دارد که الگوریتم او را به درستی تشخیص دهد. در مقابل، شاخص صحت برابر ۷۳/۵ درصد بدست آمده که به این معنی است که از بین تمامی مواردی که توسط مدل سودده پیش‌بینی شده تنها ۷۳/۵ درصد آن‌ها واقعاً سودده بوده‌اند و مدل به درستی پیش‌بینی کرده است. شاخص ترکیبی امتیاز-F برای مدل جنگل تصادفی برابر ۸۱/۹ درصد است. به طور کلی، بالاتر بودن مقدار امتیاز-F نشان دهنده عملکرد مناسب الگوریتم است. شاخص بالابری مدل نیز برابر ۱/۳۸۶ بدست آمده که نشان می‌دهد عملکرد مدل در دسته‌بندی مؤثر بوده است. نرخ اشتباه نیز برای الگوریتم جنگل تصادفی برابر ۲۲/۴ درصد است. منحنی مشخصه عملکرد (ROC) مدل نیز در شکل ۴ نمایش داده شده است. سطح زیر منحنی یا شاخص AUC بدست آمده برابر ۰/۷۴ است.

۳-۱-۵. پیاده‌سازی و ارزیابی عملکرد الگوریتم بیز ساده

عملکرد الگوریتم بیز ساده در دسته‌بندی داده‌های تست در ماتریس اغتشاش نمودار ۵ نمایش داده شده است. معیار دقت الگوریتم بیز ساده ۷۶ درصد بدست آمده که به این معناست که الگوریتم بیز ساده در تقسیم مشتریان به دو دسته سودده و زیانده در تقریباً ۷۶ درصد از موارد پیش‌بینی درستی داشته است. به لحاظ معیار بازخوانی، مقدار شاخص برابر ۹۳ درصد بدست آمده که نشان می‌دهد مدل توانسته ۹۳ درصد افرادی که واقعا سودآور بوده‌اند را به درستی شناسایی کند. این مقدار قابل توجهی است اما از آن طرف اگر به معیار صحت دقت شود، مقدار بدست آمده به این معنی است که از میان مشتریانی که توسط مدل سودده پیش‌بینی شده، تنها ۷۲/۵ درصد واقعا سودده بوده‌اند. شاخص ترکیبی امتیاز-F نیز مقداری برابر ۸۱/۳ درصد است که نشان دهنده تعادل نسبتا خوب میان دو معیار صحت و بازخوانی است. نرخ بالابری ۱/۳۵ و نرخ اشتباه طبقه‌بندی نیز برابر ۲۳/۳ درصد بدست آمده است. منحنی مشخصه عملکرد (ROC) الگوریتم بیز ساده در نمودار ۵ نمایش داده شده و می‌توان دید که سطح زیر منحنی یا شاخص AUC بدست آمده برابر ۰/۷۵ است.

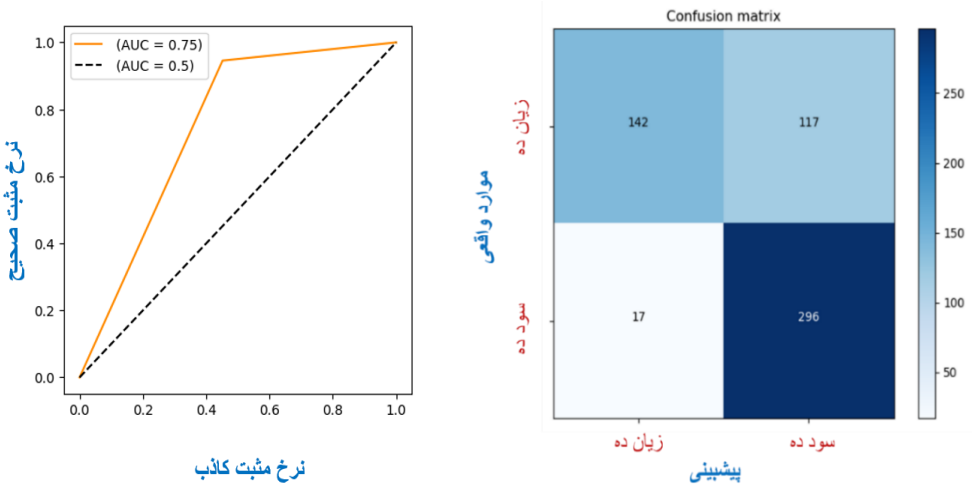


شکل ۵. ماتریس اغتشاش و منحنی ROC الگوریتم بیز ساده

مأخذ: محاسبات تحقیق

۴-۱-۵. پیاده‌سازی و ارزیابی عملکرد الگوریتم ماشین بردار پشتیبان (SVM)

عملکرد الگوریتم ماشین بردار پشتیبان (SVM) در دسته‌بندی داده‌های تست در ماتریس اغتشاش نمودار ۶ نمایش داده شده است. دقت مدل بردار پشتیبان برابر ۷۶/۵ بدست آمده که به این معناست که این الگوریتم در دسته‌بندی مشتریان سودده و زیان‌ده در تقریباً ۷۶/۵ درصد از موارد پیش‌بینی درستی داشته است. مقدار شاخص بازخوانی مدل برابر ۹۵ بدست آمده که به این معناست که مدل می‌تواند این نسبت از مواردی که به‌واقع سودده هستند را به درستی شناسایی کند. مقدار شاخص صحت نیز برابر ۷۲ بدست آمده که نشان می‌دهد از مواردی که توسط مدل سودده پیش‌بینی شده این نسبت در واقع نیز سودده بوده‌اند. شاخص ترکیبی امتیاز-F برابر با ۸۱/۵ بدست آمده که نشان دهنده تعادل میان صحت و بازخوانی و عملکرد مطلوب مدل از هر دو جنبه است. شاخص بالابری ۱/۳۳ نیز نشان می‌دهد مدل می‌تواند شانس شناسایی مشتریان را تا ۱/۳۳ برابر در مقایسه با یک دسته‌بندی تصادفی افزایش دهد. نرخ اشتباه در دسته‌بندی نیز بابر ۲۳/۵ درصد بوده است. منحنی مشخصه عملکرد (ROC) الگوریتم بردار پشتیبان (SVM) نیز در نمودار ۶ نمایش داده شده که سطح زیر منحنی یا شاخص (AUC) بدست آمده برابر ۰/۷۵ است.

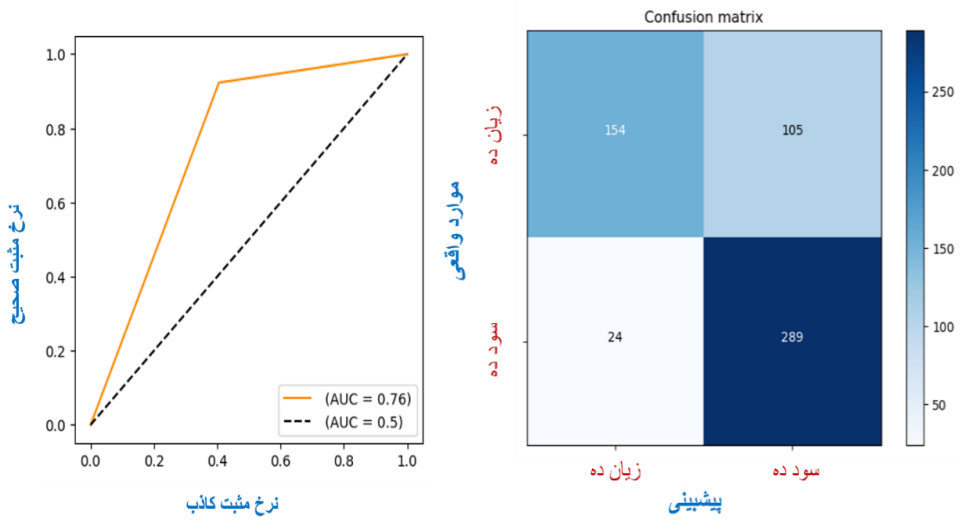


شکل ۶. ماتریس اغتشاش و منحنی ROC الگوریتم ماشین بردار پشتیبان (SVM)

مأخذ: محاسبات تحقیق

۵-۱-۵. پیاده‌سازی و ارزیابی عملکرد رگرسیون لجستیک

عملکرد رگرسیون لجستیک در دسته‌بندی داده‌های تست در ماتریس اغتشاش نمودار ۷ نمایش داده شده است. دقت مدل رگرسیون لجستیک برابر ۷۹ درصد بدست آمده که نشان می‌دهد الگوریتم رگرسیون لجستیک در دسته‌بندی مشتریان سودده و زیان‌ده در این درصد از موارد به درستی عمل کرده است. مقدار شاخص بازخوانی مدل ۹۳ درصد است که نشان می‌دهد مدل در شناسایی این درصد از مواردی که واقعاً مثبت و سودآور بوده‌اند، عملکرد درستی داشته است. شاخص صحت ۷۴ درصد است که نشان می‌دهد از میان مواردی که سودآور پیش‌بینی شده‌اند نیز ۷۴ درصد واقعاً سودده بوده‌اند. شاخص ترکیبی امتیاز-F برابر با ۸۱/۷ نشان دهنده تعادل میان این دو معیار است. مقدار شاخص بالابری برای رگرسیون لجستیک ۱/۳۷۴ است که نشان می‌دهد این مدل در شناسایی مشتریان سودده شانس را به این نسبت افزایش می‌دهد. نرخ اشتباه در دسته‌بندی نیز در رگرسیون لجستیک ۲۱/۶ درصد بوده است. منحنی مشخصه عملکرد (ROC) رگرسیون لجستیک در نمودار ۷ نمایش داده شده و می‌توان دید که سطح زیر منحنی یا شاخص (AUC) بدست آمده برابر ۰/۷۶ است.



شکل ۷. ماتریس اغتشاش و منحنی ROC رگرسیون لجستیک

مأخذ: محاسبات تحقیق

۲-۵. مقایسه عملکرد الگوریتم‌ها در دسته‌بندی مشتریان

در این مقاله به منظور دسته‌بندی و شناسایی مشتریان سودده الگوریتم‌های مختلفی بکار گرفته شد که عملکرد آن‌ها به صورت مجزا و البته با جزئیات بررسی گردید. جدول ۲ با کنار هم قراردادن شاخص‌های عملکردی الگوریتم‌های بکارگرفته شده، امکان مقایسه بین این الگوریتم‌ها را فراهم می‌کند. الگوریتم‌هایی که بر اساس هر معیار بهترین عملکرد را داشته‌اند به ترتیب با نشان‌های *، ** مشخص شده‌اند. البته باید توجه داشت که اهداف شرکت‌های بیمه‌ها و رویکردشان به مساله است که مشخص می‌کند کدامیک از معیارها برایشان اولویت بیشتری دارد. اگر هدف شرکت بیمه بالا بردن شانسش در شناسایی مشتریان سودده باشد پس باید معیار بالابری را هدف قرار داده و از رگرسیون لجستیک و الگوریتم جنگل تصادفی بهره بگیرد. اگر میزان دقت در شناسایی هر دو گروه مشتریان سودده و زیانده برای شرکت بیمه مهم است و میخواهیم پاسخ های اشتباه را کم کنیم، پس باید بر معیار دقت یا نرخ اشتباه تمرکز کرده و الگوریتم جنگل تصادفی و بیز ساده را به کار بگیریم. اگر شرکت بیمه می‌خواهد توازنی بین صحت و پوشش با وزن برابر برقرار باشد، پس باید بر امتیاز F تمرکز کرده و از الگوریتم جنگل تصادفی و رگرسیون لجستیک استفاده کند. اگر اولویت شرکت‌های بیمه شناسایی مشتریان سودده باشد پس باید به سراغ معیار بازخوانی رفته و الگوریتم ماشین بردار پشتیبان و بیز ساده را مد نظر قرار دهد. اگر برای شرکت بیمه این مهم باشد که از بین افرادی که سودده پیش بینی کرده چند درصد واقعا سودده بوده‌اند، پس باید به الگوریتم درخت تصمیم و رگرسیون لجستیک توجه کند.

با این وجود و صرفنظر از اهداف شرکت بیمه می‌توان دید که الگوریتم جنگل تصادفی بنظر و بطور کل عملکرد بهتری در مقایسه با سایر الگوریتم‌ها داشته است. الگوریتم جنگل تصادفی در ۳ معیار دقت بالا، نرخ اشتباه پایین و امتیاز F جایگاه بهترین عملکرد و در معیار بالابری دومین جایگاه را داشته است.

جدول ۲. مقایسه عملکردی الگوریتم‌های یادگیری ماشین

الگوریتم	بالابری	نرخ اشتباه طبقه‌بندی	امتیاز-F	صحت	بازخوانی	دقت
لجستیک	*۱/۳۷۴	۰/۲۱۶	**۰/۸۱۷۵	**۰/۷۴۲	۰/۹۳۱	۰/۷۴۸
ماشین بردار پشتیبان	۱/۳۳۴	۰/۲۳۵	۰/۸۱۵	۰/۷۲۳	*۰/۹۵	۰/۷۶۵
بیز ساده	۱/۳۵	**۰/۲۳۳	۰/۸۱۳	۰/۷۲۵	**۰/۹۳۳	**۰/۷۶۷
جنگل تصادفی	**۱/۳۶۸	*۰/۲۲۴	*۰/۸۱۹	۰/۷۳۵	۰/۸۶۵	*۰/۷۷۶
درخت تصمیم	۱/۳۶۴	۰/۲۴۲	۰/۸۰	*۰/۷۴۸	۰/۸۹	۰/۷۵۸

مأخذ: محاسبات تحقیق

۳-۵. تفسیرپذیری نتایج

از میان الگوریتم‌های پنجگانه بکارگرفته شده متأسفانه اغلب امکان تفسیرپذیری ندارند و ما نمی‌دانیم در نهایت این الگوریتم‌ها چگونه پیش‌بینی می‌کنند و به کدام مشخصه‌ها اهمیت بیشتری برای سودده بودن یا زیانده بودن مشتریان می‌دهند. از میان الگوریتم‌ها بکار رفته تنها دو الگوریتم درخت تصمیم و رگرسیون لجستیک این قابلیت را دارند که در ادامه به تفسیر نتایج این دو مدل پرداخته می‌شود.

طبق نتایج بدست آمده از الگوریتم درخت تصمیم، مهم ترین عامل یا فاکتور دسته‌بندی اعتیاد افراد به سیگار یا مواد شناسایی شده به گونه‌ای که الگوریتم وجود این عامل را قطعاً زیانده شناسایی کرده و سودده بودن مشتریان را به عدم اعتیاد آنها به سیگار یا مواد به عنوان شرط لازم منوط کرده است. اگر افراد اعتیاد به سیگار یا مواد نداشته باشند، در آن صورت باید به عنوان شرط کافی برای سودده بودن به سایر فاکتورها توجه کرد. عامل دوم از نظر الگوریتم درخت تصمیم یا به‌عبارتی دومین فاکتور دسته‌بندی سن افراد است. اگر سن افراد بالا باشد، الگوریتم آن‌ها را زیانده شناسایی کرده است. اما اگر افراد ضمن عدم اعتیاد به سیگار یا مواد به لحاظ سنی جوان هم باشند، در آن

صورت عوامل دیگر باید ملاک قرار گیرند که از آن جمله می‌توان به ترتیب اهمیت به سابقهٔ بیماری‌های خاص و سابقهٔ بستری و جراحی اشاره کرد. جنسیت در مرتبهٔ بعدی اهمیت از نظر الگوریتم درخت تصمیم قرار دارد. به عبارت دیگر مردان جوان بدون سابقهٔ اعتیاد به سیگار و مواد یا بدون سابقهٔ بیماری خاص یا بستری و جراحی سودآورترین مشتریان به‌لحاظ هزینه‌های درمان خواهند بود. زنان جوان بدون سابقهٔ اعتیاد به سیگار و مواد یا بدون سابقهٔ بیماری خاص یا بستری و جراحی همچنان ممکن است هزینه‌های بالایی به شرکت بیمه وارد کنند. الگوریتم درخت تصمیم دو عامل محل زندگی و همچنین سابقهٔ بیماری قلبی یا گوارش را حائز اهمیت ندانسته است.

نتایج حاصل از برآورد رگرسیون لجستیک نیز در جدول ۳ نمایش داده شده است. از آنجائی که همه متغیرها استانداردسازی شده و یک مقیاس دارند، می‌توان آن‌ها را بر حسب بزرگی ضرایب مطابق جدول مرتب کرد. طبق نتایج بدست آمده اعتیاد به سیگار یا مواد مهمترین عامل شناسایی شده است که می‌تواند احتمال زیانده بودن را بیش از ۵۷ درصد افزایش دهد. وجود بیماری خاص و همچنین سابقهٔ بیشتری و جراحی نیز در مراتب بعدی اهمیت قرار دارند و می‌توانند به ترتیب احتمال زیاندهی مشتریان را حدود ۱۹ درصد و بیش از ۱۰ درصد افزایش دهند. سن متغیر بعدی تأثیرگذار است که در بازه [-1,1] استاندارد سازی شد. طبق نتیجه بدست آمده ۱ واحد افزایش در متغیر سن استاندارد شده (به عنوان مثال اختلاف مسن ترین فرد در نمونه با افراد در سن میانه) می‌تواند احتمال زیانده بودن مشتریان را حدود ۹ درصد افزایش دهد. جنسیت نیز اهمیت دارد و زن بودن احتمال زیانده بودن مشتریان را ۸/۷ درصد بالاتر می‌برد.

جدول ۳. عوامل تأثیرگذار بر احتمال زیانده بودن مشتریان بیمه درمان

منبع	LogWorth	P-value
اعتیاد به سیگار یا مواد	۵۷/۲۳	۰/۰۰۰۰
بیماری خاص یا مصرف دارو خاص	۱۸/۸	۰/۰۰۰۰
سابقه بستری یا جراحی	۱۰/۴	۰/۰۰۰۰
سن	۸/۹	۰/۰۰۰۰
جنسیت (زن)	۸/۷	۰/۰۰۰۰
کاهش وزن شدید یا ازکارافتادگی	۴/۵	۰/۰۰۰۴
بیماری روانی	۴/۳	۰/۰۰۰۵
بیماری موروثی	۱/۶	۰/۰۲۶
کد شعبه صادر کننده حواله	۱/۳	۰/۰۴۵
بیماری گوارش	۰/۴	۰/۳۹

مأخذ: محاسبات تحقیق

تمامی موارد بالا توسط الگوریتم درخت تصمیم نیز به عنوان عوامل مهم و تأثیرگذار شناخته شده بودند. طبق نتایج رگرسیون لجستیک می‌توان دید که کاهش وزن شدید یا ازکارافتادگی، سابقهٔ بیماری روانی و سابقهٔ بیماری موروثی نیز به لحاظ آماری بر احتمال زیانده بودن مشتریان تأثیرگذارند اگرچه میزان اثرگذاری آن‌ها به نسبت سایر عوامل کمتر است. محل زندگی که با کد شعبه صادرکننده مشخص شده معنی‌داری آماری ضعیفی دارد و بیماری قلبی و گوارش نیز تأثیرگذار شناخته نشده است.

۶. نتیجه‌گیری

در صنعت بیمه درمان دو مفهوم بسیار مهم وجود دارد که از سود شرکت‌ها کاسته و موجب ناکارایی در بازار می‌شوند؛ کژگزینی و کژمنشی. کژگزینی به آن معنی است که بیمه‌نامه‌گذارانی که احتمال بیشتری برای نیاز به خدمات درمانی دارند بیشتر از سایر افراد بیمه بخرند که می‌تواند برای بیمه‌گران به معنای هزینه بیشتر و افزایش خطرات مالی در پرداخت خسارت باشد. غربالگری یک روش برای رفع این مشکل است به این معنی که شرکت بیمه می‌تواند بر اساس سابقهٔ پزشکی مشتریان سودده را از زیانده متمایز کند. متأسفانه به دلیل نبود پایگاه جامع سلامت، شرکت‌های بیمه مجبورند برای بررسی سلامت افراد به پرسشنامه سلامتی اکتفا کنند که بیمه‌گذار بصورت خوداظهاری در اختیار آن‌ها قرار می‌دهد.

این مقاله تلاش کرد تا با داده‌کاوی از اطلاعات شخصی و پرسشنامه سلامت و بکارگیری الگوریتم‌های یادگیری ماشین به شرکت بیمه مورد بررسی برای شناسایی مشتریان سودده از زیانده کمک کند. نتایج بدست آمده طبیعتاً می‌تواند توسط دیگر شرکت‌های بیمه درمان نیز مورد استفاده قرار گیرد. در بررسی انجام شده مشخص شد که الگوریتم جنگل تصادفی بهترین عملکرد را در پیش‌بینی داشته و توانسته به دقت حدود ۷۸ درصد در دسته‌بندی مشتریان سودده و زیانده داشته باشد. همچنین مشخص شد که از میان فاکتورهای مختلف اعتیاد به مصرف سیگار یا مواد، سابقهٔ بیماری خاص یا مصرف دارو خاص، سابقه بستری یا جراحی، سن، جنسیت (زن بودن) از مهمترین عوامل تأثیرگذار بر هزینه‌های درمان بوده و احتمال زیانده بودن مشتریان را به شدت افزایش می‌دهند. این نتایج همراستا و منطبق با نتایج بررسی‌های مطالعات دیگری است که در بخش پیشینه پژوهش و معرفی متغیرهای پژوهش به بحث گذاشته شده بودند. نتایج بدست آمده همچنین نشان دادند که عوامل دیگر از جمله کاهش وزن شدید یا ازکارافتادگی، سابقهٔ بیماری روانی و سابقهٔ بیماری موروثی نیز در مرتبهٔ بعدی تأثیرگذاری بر زیانده بودن مشتریان قرار دارند. نتایج این مقاله همچنین همراستا با سایر مطالعات در این زمینه نشان داد که یادگیری ماشین و هوش مصنوعی می‌تواند به خوبی در شناسایی و دسته‌بندی مشتریان مورد استفاده قرار گیرد.

استفاده از نتایج این مقاله از جهت مختلف برای شرکت‌های بیمه درمان مفید خواهد بود. نخست این‌که دسته‌ای از مشتریان سودده را شناسایی کردیم که تمرکز بر آن‌ها می‌تواند سود شرکت بیمه را افزایش دهد. نتایج بدست آمده نشان داد که تمرکز بر مردان جوان بدون سابقهٔ اعتیاد به سیگار و مواد، بدون سابقهٔ بیماری خاص، بستری و جراحی، کاهش وزن شدید یا ازکارافتادگی، بیماری روانی و بیماری موروثی می‌تواند احتمال سودده بودن مشتریان را نسبت به فراوانی آن‌ها در جامعه (برازش تصادفی) تا ۱/۳۷۴ برابر افزایش دهد. نتایج بدست آمده نشان دادند که تمرکز مدل بر این گروه به معنی شناسایی موفق بیش از ۹۵ درصد مشتریان سودده است. این همه نشان می‌دهد که شرکت بیمه باید به این دسته از مشتریان توجه ویژه داشته و با اعمال تخفیفات یا برنامه‌های وفاداری نسبت به جذب و حفظ و نگهداشت آن‌ها اهتمام داشته باشد. دوم این‌که الگوریتم‌های مورد استفاده توانستند دسته‌ای از مشتریان را زیانده شناسایی کنند که ۹۰ درصد آن‌ها در واقع نیز زیانده هستند. با این عملکرد مناسب در شناسایی مشتریان زیانده، حال شرکت بیمه با تعیین حق‌بیمه مناسب متناسب با ریسک می‌تواند عدالت در اعمال حق‌بقیه را محقق کند. سوم این‌که نتایج بدست آمده نشان دادند که پرسشنامه سلامت در برخی از موارد با ضعف همراه است. به‌عنوان نمونه سابقهٔ بیماری قلبی یا گوارش تأثیرگذار نبود اما این نتیجه خلاف انتظار است و این امر نشانگر لزوم اعتبارسنجی یا تصحیح سوالات مربوط به این حوزه در پرسشنامه سلامت است.

این مقاله از معدود مطالعاتی است که در داخل کشور در این ارتباط انجام شده و دلیل آن محدودیت‌های داده است. با توجه به دسترسی سخت به داده‌های خرد بیمه تحقیقات زیادی در حوزه بررسی مشتریان بیمه صورت نگرفته و به این ترتیب فرصت‌های مطالعاتی زیادی برای تحقیقات پیش‌رو وجود دارد. مقولات زیادی وجود دارد از جمله اضافه کردن ویژگی‌های مانند قد و وزن که تحقیقات جهانی نشان داده‌اند در هزینه‌های درمانی تاثیر بسزایی دارند، و بررسی نکاتی مانند تحصیلات فرد، درآمد ماهیانه فرد، نوع شغل فرد و..... که بررسی آن‌ها و تاثیر آن‌ها می‌تواند بسیار جذاب باشد و شرکت‌های بیمه را در شناسایی مشتریان خود کمک کند.

References

Bash Afshar, M., SaedPanah, M., & Tireh Eidouzehi, F. (2018). Clustering model of life insurance customers (Case study: An insurance company). *Iranian Journal of Insurance Research, 7*(2), 108-119. (in Persian)

Ghorbani, H., Ghanbarzadeh, M., & Ofoghi, R. (2022). Investigating the churn of life insurance customers using data mining methods (A case Study: One of the Iran's insurance companies). *Iranian Journal of Insurance Research, 11*(4), 305-320. (in Persian)

Khandan, A., Niakan, L., & Fakharinezhad, Z. (2023). Predicting term life insurance surrender using deep neural networks. *Iranian Journal of Insurance Research, 12*(4), 265-282. (in Persian)

Mirzaie, M., Darabi, S., & Babapoor, M. (2017). Population Aging in Iran and Rising Health Care Costs. *Salmand: Iranian Journal of Ageing, 12*(2), 156-169. (in Persian)

Panahi, H., Janati, A., Narimani, M., Assadzadeh, A., Mohammadzadeh, P., & Naderi, A. (2014). Catastrophic expenditures for hospitalized patients in Tabriz, Iran. *Payesh, 13*(6), 655-663. (in Persian)

Parastesh, M. (2020). Clustering insurance customers based on data mining techniques for use in gamification techniques. *Iranian Journal of Insurance Research, 9*(4), 426-443. (in Persian)

Tajaddodi Nodehi, M., Hosseini Khatibani, S., Yazdinejad, M., & Zolfi, S. (2023). Predicting people's health insurance costs using machine learning and ensemble learning methods. *Iranian Journal of Insurance Research, 13*(1), 1-14. (in Persian)

Alsagheer, R. H., Alharan, A. F., & Al-Haboobi, A. S. (2017). Popular decision tree algorithms of data mining techniques: a review. *International Journal of Computer Science and Mobile Computing, 6*(6), 133-142.

Bau, Y. T., & Hanif, S. A. M. (2024). Comparative Analysis of Machine Learning Algorithms for Health Insurance Pricing. *International Journal on Informatics Visualization,* 8 (1).

Boßow-Thies, S., Hofmann-Stölting, C., & Jochims, H. (2020). *Data-driven Marketing*: Springer.

Braverman, S. (2015). Global review of data-driven marketing and advertising. *Journal of Direct, Data and Digital Marketing Practice, 16*, 181-183.

Chauhan, V. K., Dahiya, K., & Sharma, A. (2019). Problem formulations and solvers in linear SVM: a review. *Artificial Intelligence Review, 52*(2), 803-855.

Han, J., Pei, J., & Tong, H. (2022). *Data mining: concepts and techniques*: Morgan kaufmann.

Han, J., Kamber, M., & Pei, J. (2012). Data mining concepts and techniques third edition. *University of Illinois at Urbana-Champaign Micheline Kamber Jian Pei Simon Fraser University*.

Healy, L. M. (2006). Logistic regression: An overview. *Eastern Michighan College of Technology*.

Jadhav, S. D., & Channe, H. (2016). Comparative study of K-NN, naive Bayes and decision tree classification techniques. *International Journal of Science and Research (IJSR), 5*(1), 1842-1845.

Jaffar, M., Shafiq, S., Shahzadi, N., Alrajeh, N., Jamil, M., & Javaid, N. (2023). Efficient Deep Learning Models for Predicting Super-Utilizers in Smart Hospitals. IEEE, 11.

Kaushik, K., Bhardwaj, A., Dwivedi, A. D., & Singh, R. (2022). Machine learning-based regression framework to predict health insurance premiums. *International Journal of Environmental Research and Public Health, 19*(13), 7898.

Maalouf, M. (2011). Logistic regression in data analysis: an overview. *International Journal of Data Analysis Techniques and Strategies, 3*(3), 281-299.

Nandapala, E., Jayasena, K., & Rathnayaka, R. (2020). *Behavior segmentation based micro-segmentation approach for health insurance industry.* Paper presented at the 2020 2nd International Conference on Advancements in Computing (ICAC).

Pandey, P., Saroliya, A., & Kumar, R. (2018). *Analyses and detection of health insurance fraud using data mining and predictive modeling techniques.* Paper presented at the Soft Computing: Theories and Applications: Proceedings of SoCTA 2016, Volume 2.

Peng, F., Wang, D., Zhang, D., Cao, H., & Liu, X. (2018). The prospect of layered double hydroxide as bone implants: A study of mechanical properties, cytocompatibility and antibacterial activity. *Applied Clay Science, 165*, 179-187.

Rawat, S., Rawat, A., Kumar, D., & Sabitha, A. S. (2021). Application of machine learning and data visualization techniques for decision support in the insurance sector. *International Journal of Information Management Data Insights, 1*(2), 100012.

Santos, F. P., Pacheco, J. M., Santos, F. C., & Levin, S. A. (2021). Dynamics of informal risk sharing in collective index insurance. Nature Sustainability, 4(5), 426-432.

Saraswat, B. K., Singhal, A., Agarwal, S., & Singh, A. (2023). Insurance Claim Analysis Using Traditional Machine Learning Algorithms. 2023 International Conference on Disruptive Technologies (ICDT). DOI: 10.1109/ICDT57929.2023.10150491

Sen, P. C., Hajra, M., & Ghosh, M. (2020). *Supervised classification algorithms in machine learning: A* survey *and review.* Paper presented at the Emerging Technology in Modelling and Graphics: Proceedings of IEM Graph 2018.

Spence, A. M. (1973). Job Market Signaling. *Quarterly Journal of Economics. 87*(3), 355–374.

Taha, A., Cosgrave, B., & Mckeever, S. (2022). Using feature selection with machine learning for generation of insurance insights. *Applied Sciences*, *12*(6), 3209.

Wen, C., Gao, K., & Xiao, Y. (2021). Data-Driven Market Segmentation in Insurance Industry and Other Related Sectors. *Journal of Finance and Accounting*, *9*(6), 268-272.

Zaqueu, J. R. (2019). Customer clustering in the health insurance industry by means of unsupervised machine learning. *Nova University De Lisboa, Lisbon.* Available at: https://run.unl.pt/bitstream/10362/89468/1/TAA0043.pdf