Iran. J. Insur. Res., 13(1): Pages: 1-14, Winter 2024, Serial #47

Iranian Journal of Insurance Research
(IJIR)

Homepage: https://ijir.irc.ac.ir/?lang=en

Iranian Insurance Research Center (IIRC)

ORIGINAL RESEARCH PAPER

Predicting people's health insurance costs using machine learning and ensemble learning methods

M. Tajaddodi Nodehi[1], S. Hosseini Khatibani[1], M. Yazdinejad[2,*], S .Zolfi[1]

[1] *Department of Computer, Faculty of Computer Engineering, Al Taha Institute of Higher Education, Tehran, Iran*

[2] *Department of Artificial Intelligence, Faculty of Computer Engineering, Isfahan University, Isfahan, Iran*

ARTICLE INFO

Article History:
Received 10 October 2023
Revised 31 October 2023
Accepted 18 November 2023

Keywords:
Data mining
Ensemble learning
Healthcare insurance cost
Machin learning
Risk

*Corresponding Author:
Email: *mohsen.yazdinejad@eng.ui.ac.ir*
Phone: +9831 37934500
ORCID: 0000-0001-7805-6344

ABSTRACT

BACKGROUND AND OBJECTIVES: The healthcare insurance industry faces a significant challenge predicting individuals' insurance costs, which are based on complex parameters such as age and physical characteristics. Insurance companies categorize policyholders into high-risk and low-risk groups to manage risks and avoid potential losses. However, the accurate estimation of costs for each individual can be a daunting task. By leveraging data science and machine learning techniques, insurance companies can improve their cost estimation accuracy and better manage risks. This approach can help insurance companies to provide more accurate insurance coverage and pricing for individuals leading to higher customer satisfaction and lower financial losses.

METHODS: To address this challenge, a data science and machine learning-based approach that uses ensemble learning to predict high-risk and low-risk individuals is used. The method involves several steps including data preprocessing, feature engineering, and cross-validation to evaluate the model's performance. The first step involves preprocessing the data by cleaning it, handling missing values, and encoding categorical variables. The second step generates new features using feature engineering techniques such as scaling, normalization, and dimensionality reduction. Next, ensemble learning is used to combine multiple regression methods such as logistic regression, neural networks, support vector machines, random forests, LightGBM, and XGBoost. By combining these methods, the aim is to leverage their strengths and minimize their weaknesses to achieve better prediction accuracy. Finally, the model's performance is evaluated using cross-validation techniques such as k-fold cross-validation. These techniques help to validate the model's accuracy and prevent overfitting.

FINDINGS: The proposed approach achieves an AUC of 0.73 demonstrating its effectiveness in predicting high-risk and low-risk individuals.

CONCLUSION: In conclusion, the healthcare insurance industry can benefit greatly from data science and machine learning-based approaches. By accurately predicting high-risk and low-risk individuals, insurance companies can better manage risks and provide more accurate coverage and pricing for their customers. This can lead to the improvement of customer satisfaction and the reduction of financial losses for insurance companies.

DOI: 10.22056/ijir.2024.01.01

نشریه علمی پژوهشنامه بیمه

سایت نشریه: https://ijir.irc.ac.ir/?lang=fa

مقاله علمی

پیش‌بینی هزینه‌های بیمهٔ درمانی افراد با استفاده از یادگیری ماشین و روش یادگیری جمعی

مهسا تجددی نودهی[۱]، سمانه حسینی خطیبانی[۱]، محسن یزدی نژاد[۲،*]، سمیه زلفی[۱]

[۱] گروه کامپیوتر، دانشکده مهندسی کامپیوتر، موسسه آموزش عالی آل طه، تهران، ایران

[۲] گروه هوش مصنوعی، دانشکده مهندسی کامپیوتر، دانشگاه اصفهان، اصفهان، ایران

اطلاعات مقاله

تاریخ های مقاله:

تاریخ دریافت: ۱۸ مهر ۱۴۰۲

تاریخ داوری: ۰۹ آبان ۱۴۰۲

تاریخ پذیرش: ۲۷ آبان ۱۴۰۲

کلمات کلیدی:

داده کاوی

ریسک

هزینه بیمه درمان

یادگیری جمعی

یادگیری ماشین

*نویسنده مسئول:

ایمیل: mohsen.yazdinejad@eng.ui.ac.ir

تلفن: ۳۷۹۳۴۵۰۰ ۹۸۳۱+

ORCID: 0000-0001-7805-6344

DOI: 10.22056/ijir.2024.01.01

چکیده:

پیشینه و اهداف: صنعت بیمهٔ درمانی در پیش‌بینی هزینه‌های بیمه افراد که براساس پارامترهای پیچیده‌ای مانند سن و ویژگی‌های فیزیکی است، با چالش مهمی مواجه است. شرکت‌های بیمه برای مدیریت ریسک و جلوگیری از زیان احتمالی، بیمه‌گذاران را به دو گروه پرخطر و کم‌خطر دسته‌بندی می‌کنند. بااین‌حال، برآورد دقیق هزینه‌ها برای هر فرد می‌تواند کار سختی باشد. برای مقابله با این چالش، ما رویکردی مبتنی بر علم داده و یادگیری ماشین را پیشنهاد می‌کنیم که از یادگیری جمعی برای پیش‌بینی افراد پرخطر و کم‌خطر استفاده می‌کند.

روش‌شناسی: روش پیشنهادی شامل مراحل مختلفی از جمله پیش‌پردازش داده‌ها، مهندسی ویژگی‌ها و اعتبارسنجی متقابل برای ارزیابی عملکرد مدل است. در مرحلهٔ اول، داده‌ها را با پاک کردن، مدیریت مقادیر ازدست‌رفته و رمزگذاری متغیرهای طبقه‌بندی، پیش‌پردازش می‌کنیم. در مرحلهٔ دوم، ما ویژگی‌های جدیدی را با استفاده از روش‌های مهندسی ویژگی‌ها مانند مقیاس‌بندی، نرمال‌سازی و کاهش ابعاد تولید می‌کنیم. این روش‌ها به استخراج اطلاعات معنادار از داده‌ها و بهبود عملکرد مدل کمک می‌کند. در مرحلهٔ بعد، ما از یادگیری جمعی برای ترکیب روش‌های رگرسیون متعدد، مانند رگرسیون لجستیک، شبکه‌های عصبی، ماشین‌های بردار پشتیبانی، جنگل‌های تصادفی، LightGBM و XGBoost استفاده می‌کنیم. هدف از ترکیب این روش‌ها این است که از نقاط قوت آن‌ها استفاده کنیم و نقاط ضعف آن‌ها را به حداقل برسانیم تا به دقت پیش‌بینی بهتری دست یابیم. در نهایت، عملکرد مدل را با استفاده از روش اعتبارسنجی متقاطع k-fold ارزیابی می‌کنیم. این روش به اعتبارسنجی دقت مدل و جلوگیری از برازش بیش از حد کمک می‌کند.

یافته‌ها: رویکرد پیشنهادی ما به AUC برابر با ۰/۷۳ دست می‌یابد که اثربخشی آن را در پیش‌بینی افراد پرخطر و کم‌خطر نشان می‌دهد.

نتیجه‌گیری: با استفاده از علم داده و روش‌های یادگیری ماشین، شرکت‌های بیمه می‌توانند دقت برآورد هزینهٔ خود را بهبود بخشند و ریسک را بهتر مدیریت کنند. این رویکرد می‌تواند به شرکت‌های بیمه کمک کند تا پوشش بیمه‌ای و قیمت‌گذاری دقیق‌تری را برای افراد ارائه دهند که به رضایت بیشتر مشتریان و کاهش زیان‌های مالی منجر می‌شود.

توجه: مدت‌زمان بحث و انتقاد برای این مقاله تا ۱ آوریل ۲۰۲۴ در وب‌سایت IJIR در «نمایش مقاله» باز است.

مقدمه

برای بهبود پاسخگویی به سلامت افراد در صنعت بیمه، پیش‌بینی هزینهٔ بیمهٔ درمانی مورد نیاز است (Sommers, 2020). بیمه یکی از ارکان توسعهٔ نظام مراقبت‌های بهداشتی در جهان است. مطالعه و بررسی در حوزهٔ خسارات و بیماری‌ها کمک می‌کند تا ذی‌نفعان به‌راحتی بتوانند در این خصوص سیاست‌گذاری کنند (Ahmadlou *et al.*, 2023). ازاین‌رو با پیش‌بینی هزینه‌های درمانی هم بیمه‌شونده و هم بیمه‌گذار می‌توانند تاحدودی آینده را پیش‌بینی کنند و گزینه‌های بهتری برای تصمیم‌گیری داشته باشند . (Tkachenko *et al.*, 2018) نرخ بیمه تحت تأثیر برخی مسائل پزشکی است. برآورد دقیق هزینه‌های مراقبت‌های بهداشتی فردی و درمانی برای طیفی از ذی‌نفعان و آژانس‌های بهداشتی مهم است (Morid *et al.*, 2017).

کاهش هزینه‌ها با استفاده از مدل‌های پیش‌بینی بسیار مهم است. با توجه به آیین‌نامهٔ شمارهٔ ۸۱ بیمهٔ مرکزی ج.ا. (مقررات تعیین حق‌بیمه) مادهٔ شمارهٔ ۵ هریک از مؤسسات بیمه موظف‌اند تعرفهٔ حق‌بیمهٔ رشته‌های بیمه‌ای خود را به‌نحوی تعیین نمایند که در هر سال ضریب خسارت در یک محدودهٔ معین قرار بگیرد. به‌طور مثال ضریب خسارت در رشتهٔ درمان می‌بایست بیشتر از ۵۰٪ و کمتر از ۸۵٪ باشد. اگر مقدار ضریب خسارت کمتر از مقدار نرمال مقدارگذاری شود، این محصول برای شرکت‌های بیمه رشتهٔ سودآوری محسوب نمی‌شود و به سمت زیان‌دهی حرکت می‌کند (Sepahvand *et al.*, 2022). دپارتمان‌های بهداشتی حجم عظیمی از داده‌های مربوط به بیماری، بیماران و فرایندهای تشخیصی را ارائه می‌کنند، اما این داده‌ها به‌درستی تجزیه‌وتحلیل نمی‌شوند (Milovic and Milovic, 2012). ارائه‌دهندگان بیمه و سازمان‌هایی که خدمات درمانی ارائه می‌کنند، باید با برنامه‌ریزی صحیح منابع محدود سازمان را اولویت‌بندی کنند. علاوه‌براین، اطلاع از هزینه‌های احتمالی آینده می‌تواند به بیماران کمک کند تا گزینه‌های تصمیم‌گیری برای بیمه با فرانشیزها و قیمت‌های معقول را انتخاب کنند. این عوامل به رشد مقررات بیمه‌ای نیز کمک می‌کند (Kumar *et al.*, 2012). یکی از مهم‌ترین چالش‌های صنعت بیمه، پیش‌بینی هزینه‌های بیمهٔ درمانی افراد است. هزینه‌های درمانی با ویژگی‌های داده‌های فردی پیچیده سنجیده می‌شود و به همین دلیل است که پیش‌بینی در این مورد، باید مبتنی بر رویکرد داده‌محور شخصی‌سازی شده باشد که عوامل زیادی را در نظر می‌گیرد (Perova and Pliss, 2017). حجم زیاد داده (Shakhovska *et al.*, 2015)، تأثیر عوامل گوناگون (Chyrun *et al.*, 2018) و وابستگی پارامترهای بین متغیرها (Babichev *et al.*, 2018) از چالش‌های دیگری است که در پژوهش‌ها به‌طور کامل مطالعه نشده است. عوامل یادشده، استفاده از ابزارهای هوش مصنوعی را برای حل این چالش ضروری می‌کند (Bodyanskiy *et al.*, 2017).

الگوریتم‌های یادگیری ماشین نقش مهمی در پیش‌بینی دقیق هزینهٔ بیمهٔ درمانی دارند (Yang *et al.*, 2018). داده‌کاوی نقش مهمی در جنبه‌های مختلف صنعت بیمه ایفا می‌کند، مانند ارزیابی ریسک، کشف تقلب، تجزیه‌وتحلیل پذیره‌نویسی، تجزیه‌وتحلیل ادعا، تجزیه‌وتحلیل بازاریابی، توسعهٔ محصول، پروفایل مشتری و مانند آن. همراه با داده‌کاوی، صنعت در حال تغییر به‌سمت الگوریتم‌های یادگیری ماشین است تا با استفاده از تجزیه‌وتحلیل انجام‌شده روی مجموعه داده‌های بزرگ برای تشخیص بهتر تقلب، تأیید KYC، ارزیابی خط‌مشی رفتاری و تسویهٔ ادعای سفارشی، پیش‌بینی کند. استفاده از یادگیری ماشین، به‌شدت در صنعت بیمه در حال افزایش است. یادگیری ماشین در تجزیه‌وتحلیل ادعا و پردازش برای جداسازی ادعاها، شناسایی ادعاهای دور از دسترس و حتی کلاهبرداری استفاده می‌شود، و موجب کاهش دخالت انسان در پردازش ادعا می‌شود. استفاده از الگوریتم‌های یادگیری ماشین در الگوی ثبت ادعای ذی‌نفع و همچنین الگوی پذیرش ادعای خودکه می‌تواند برای بهینه‌سازی کل جریان فرایند برای ثبت‌نام خط‌مشی کاری استفاده کند، کمک می کند.

رویکردهای زیادی برای حل مسئلهٔ پیش‌بینی هزینه‌های بیمهٔ سلامت وجود دارد. روش‌های رگرسیون (Marmolejo-Ramos *et al.*, 2023)، که اغلب برای این اهداف استفاده می‌شوند، همیشه دقت کافی را ارائه نمی‌دهند. این امر به‌دلیل استفاده از ویژگی‌های واقعی داده‌هاست و واقعیت این است که در بیشتر موارد آن‌ها توزیع نرمال ندارند و فرض همسویی را برآورده نمی‌کنند. علاوه‌بر این، استفاده از برخی روش‌های بهینه‌سازی برای طبقه‌بندی کاملاً زمان‌برند. شبکهٔ عصبی (Du *et al.*, 2023) و مدل‌های فازی عصبی (Park *et al.*, 2023)، همیشه دقت بالایی ندارند، همچنین، الگوریتم‌های آموزشی که تکرار اساس کار آن‌هاست، به زمان طولانی برای کار نیاز دارند که ایرادی قابل توجه برای داده‌های با حجم بالاست. یکی از بزرگ‌ترین مزیت‌های هوش محاسباتی و روش‌های استخراج داده‌ها، قابلیت تطبیق آن‌ها با حوزهٔ مسئله است، و همین موضوع می‌تواند منشأ چالش‌هایی برای آن‌ها باشد (Benedek *et al.*, 2022).

در این مقاله روشی برای پیش‌بینی هزینهٔ بیمهٔ درمانی و شناسایی افراد پرهزینه و پرخطر برای بیمهٔ درمانی با استفاده از علم داده‌کاوی و الگوریتم‌های یادگیری ماشین پیشنهاد می‌شود. در روش پیشنهادی از مهندسی ویژگی‌ها، همچون ساخت ویژگی‌های جدید، نرمال‌سازی ویژگی‌ها و روش‌های انتخاب ویژگی برای بهبود نتایج پیش‌بینی استفاده می‌شود. روش پیشنهادی با یادگیری جمعی و با ترکیب روش‌های رگرسیون لجستیک، شبکهٔ عصبی پرسپترون چندلایه، ماشین بردار پشتیبان، جنگل تصادفی، LightGBM و XGBoost، پیش‌بینی را انجام می‌دهد.

در ادامه در بخش ۲ مروری بر روش‌های پیش‌بینی هزینه‌ها در صنعت بیمه ذکر می‌شود، در بخش ۳ روش پیشنهادی شرح داده می‌شود، و در بخش ۴ به ارزیابی و بررسی روش پیشنهادی می‌پردازیم و در نهایت در بخش ۵ جمع‌بندی و نتیجه‌گیری و پیشنهادهایی برای کارهای آتی ارائه می‌شود.

مروری بر پیشینهٔ پژوهش

پیش‌بینی هزینهٔ بیمهٔ درمانی کار بسیار مهمی در بخش مراقبت‌های بهداشتی است. برای برآورد هزینه‌های درمانی بیمه‌گذار، هزینهٔ پزشکی را بسیاری از پژوهشگران با استفاده از الگوریتم‌های مختلف یادگیری ماشین پیش‌بینی کرده‌اند. نویسندگان در Vijayalakshmi *et al.* (2023) از مجموعه داده با ۲۴ ویژگی شامل تمامی ویژگی‌های مرتبط مورد نیاز برای پیش‌بینی هزینهٔ بیمه استفاده کردند. برای پیاده‌سازی از الگوریتم‌های رگرسیون مانند رگرسیون خطی، رگرسیون درخت تصمیم، رگرسیون کمند، رگرسیون ریج، رگرسیون جنگل تصادفی، رگرسیون شبکهٔ الاستیک، رگرسیون بردار پشتیبان، رگرسیون K نزدیک‌ترین همسایه و رگرسیون شبکهٔ عصبی، استفاده شد و رگرسیون جنگل تصادفی با ۰,۹۵۳۳ به‌عنوان مقدار R-Squared عملکرد بهتری نشان داد.

Albalawi *et al.* (2023) دو رویکرد ارائه کردند که اولی از هوش محاسباتی برای پیش‌بینی هزینه‌های بیمهٔ مراقبت‌های بهداشتی با استفاده از الگوریتم‌های یادگیری ماشین استفاده می‌کند. و دومی اسپارک است که یک ابزار کلان‌داده در نظر گرفته می‌شود. در میان رویکرد اول، الگوریتم‌های رگرسیون خطی و رگرسیون چندجمله‌ای معروف‌اند که براساس ویژگی‌های داده‌های ورودی است. رگرسیون خطی روشی است که رابطهٔ بین دو یا چند متغیر را نشان می‌دهد. اما در تحلیل چندجمله‌ای، رابطهٔ بین متغیرهای وابسته و مستقل با استفاده از چندجمله‌ای درجهٔ n مدل‌سازی می‌شود. نتایج به‌دست‌آمده نشان می‌دهد که عملکرد مدل رگرسیون درختی تقویت‌شده با گرادیان بسیار بهتر از یک جنگل چندمتغیره و تصادفی R2 برابر با ۰۰/۹۰۶۷ است.

Anwar ul Hassan *et al.* (2021) هزینهٔ بیمهٔ درمانی را با استفاده از رویکرد هوشمند محاسباتی پیش‌بینی کرد. آن‌ها رگرسیون خطی، XGBoost، SGB، درخت تصمیم، جنگل تصادفی، رگرسیون خطی چندگانه و KNN را در داده‌های Kaggle با RMSE برابر با ۰/۳۴۰ و دقت ۸۶ درصد اعمال کردند.

Christobel and Subramanian (2022) هزینهٔ بیمهٔ درمانی را با استفاده از روش‌های خطی، lasso، ridge و رگرسیون چندجمله‌ای در همان مجموعه داده‌ها با دقت ۸۸٪ پیش‌بینی کرد. Shakhovska *et al.* (2022)، از روش‌های گروهی مانند KNN، ماشین بردار پشتیبان، درخت رگرسیون، رگرسیون خطی و SGB در همان مجموعه داده‌ها استفاده کرد. الگوریتم Boruta برای انتخاب ویژگی استفاده شد و ۱۷۳/۲۱۳ را به‌عنوان RMSE تولید کرد. Drewe-Boss *et al.* (2022) در داده‌هایی که از InGef جمع آوری شده به R2 برابر با ۰/۲۶۶و MAPE برابر با ۲۰۰۴/۳۳ و CPM (اندازهٔ پیش‌بینی کامینگ) ۰/۳۲۶ رسیدند، که برای آن از شبکهٔ عصبی عمیق و رگرسیون استفاده شده است. Shyamala Devi *et al.* (2021) از رگرسیون چندجمله‌ای، جنگل تصادفی و آزمون ANOVA روی داده‌های مخزن ماشین UCI استفاده کرد و با رگرسیون چندجمله‌ای R2 معادل ۰/۸۸ رسیدند. Pfutzenreuter and Lima (2021) از رگرسیون خطی، SVR، KNR، درخت ساده، جنگل تصادفی و XGBoost در داده‌های Kaggle استفاده کرد، R2 را با ۰/۸۸ تولید کرد. Kaushik *et al.* (2022) شبکهٔ عصبی مصنوعی را در همان مجموعه داده با دقت ۹۲/۷۲ درصد اعمال کرد. Tkachenko *et al.* (2018) روش خطی تکه‌ای را با استفاده از ساختار عصبی SGTM برای پیش‌بینی هزینهٔ بیمهٔ درمانی در همان مجموعه داده ایجاد کرد. آن‌ها ۶۰ /۳۰ به‌عنوان MAPE و ۲۹۳۶۳۴/۳۴۵۳ به‌عنوان MAE دریافت کردند.

Hanafy and Omar (2021) قیمت بیمهٔ درمانی را با استفاده از الگوریتم‌های یادگیری ماشین و مدل‌های رگرسیون DNN در همان مجموعه داده پیش‌بینی کردند. آن‌ها از MAE، RMSE و R-squared به‌عنوان متریک استفاده کردند و به‌ترتیب مقادیر ۰/۱۷۴۴۸، ۰/۳۸۰۱۸ و ۸۵/۸۲۹۵ را به‌دست آوردند.

Lakshmanarao *et al.* (2020) مدلی برای پیش‌بینی هزینهٔ درمانی با استفاده از الگوریتم‌های رگرسیون مانند MLR، SVR، DTR و جنگل تصادفی ساختند که در آن رگرسیون جنگل تصادفی نسبت به سایر الگوریتم‌ها عملکرد بهتری داشت. آن‌ها معیارهایی مانندR2 ۰/۸۵، MSE 23294452، MAE 2760 و RMSE 4826 را در همان مجموعه داده به‌دست آوردند.

Kafuria (2022) یک مدل پیش‌بینی با استفاده از الگوریتم‌های یادگیری ماشین برای محاسبهٔ قیمت بیمهٔ درمانی ایجاد کرد. او از MLR، KNN، LASSO، XGBoost و RFR استفاده کرد. الگوریتم XFBoost معیار ارزیابی R2 را ۸۵ درصد، MAE را۲۶۸۸، RMSE را ۴۷۴۸ تولید کرد. Bhardwaj and Anand (2020) مدلی را با استفاده از MLR، DTR و رگرسیون درخت تصمیم تقویت‌کنندهٔ گرادیان در همان مجموعه داده برای پیش‌بینی مبلغ بیمهٔ درمانی ایجاد کردند که دقت گرادیان بوست برابر با ۹۹/۵٪ بود. در پژوهشی دیگر Goundar *et al.* (2020) مدل پیش‌بینی را با استفاده از شبکهٔ عصبی مصنوعی پیش‌خور و شبکهٔ عصبی مکرر در داده‌های BSP LIFE Limited (Fuji) و RNN، ۹۳ درصد برای پیش‌بینی بیمهٔ درمانی ایجاد کردند. همچنین Fauzan and Murfi (2018) پیش‌بینی خسارت بیمه را با استفاده از XGBoost، AdaBoost، Stochastic GB، جنگل تصادفی و شبکهٔ عصبی انجام دادند که در آن XGBoost عدد ۰/۲۸ را به‌عنوان ضریب جینی نرمال‌شده به‌دست آورد. در ادامه می‌توان به پژوهش Kumar Sharma and Sharma (2020) اشاره داشت که سیستم پیش‌بینی بیمهٔ سلامت را با استفاده از رگرسیون خطی چندگانه در همان مجموعه داده با MAPE برابر با ۳٪ و R2 با مقدار ۰/۷۶ ایجاد کردند. موضوع حائز اهمیت این است که همهٔ سیستم‌های موجود از مجموعه داده تنها با هفت ویژگی استفاده می‌کردند.

در Eriksson *et al.* (2004) رویکرد خطی تکه‌ای برای حل یک مشکل رگرسیون پیشنهاد شد. مزایای آن در حجم زیادی از پردازش داده‌ها، که در آن دقت کار روش بسیار مهم است، مشهود است. در Tkachenko *et al.* (2018) و Doroshenko (2018)، برای حل

مسئله طبقه‌بندی با سرعت بالا براساس تقریب پله‌ای تکه‌ای، استفاده از ساختار عصبی مانند SGTM پیشنهاد شده است. نویسندگان نتایج رضایت‌بخشی از طبقه‌بندی ارائه می‌کنند، و علاوه‌براین، الگوریتم آموزشی ساختار عصبی غیرتکراری SGTM سرعت بالای روش پیشنهادی را فراهم می‌کند.

با وجود تلاش‌های صورت‌گرفته، هنوز پیش‌بینی هزینه‌های بیمهٔ درمانی افراد از دقت کافی برخوردار نیست. ازطرفی اکثر کارهای پیشین بر روی تعداد مشخصی از ویژگی‌ها پیش‌بینی را انجام می‌دهند، بسیاری از پژوهشگران از الگوریتم‌های یادگیری ماشینی مختلف برای پیش‌بینی حق‌بیمه با ویژگی‌های محدود مانند سن، جنس، فرزند، فرد سیگاری، منطقه و هزینه‌ها استفاده کردند. اما تنها این ویژگی‌ها برای پیش‌بینی هزینه‌ها و میزان خطر افراد بیمه‌شونده برای بیمهٔ سلامت مناسب نیستند و تعداد ویژگی‌های بیشتری برای پیش‌بینی دقیق‌تر مورد نیاز است. بنابر آنچه ذکر شد در روش پیشنهادی با استفاده از الگوریتم‌های یادگیری ماشین و افزایش ویژگی‌ها و سایر روش‌های پیشنهادشده، راه‌حلی برای پیش‌بینی دقیق‌تر هزینه‌های بیمهٔ سلامت افراد پیشنهاد می‌شود.

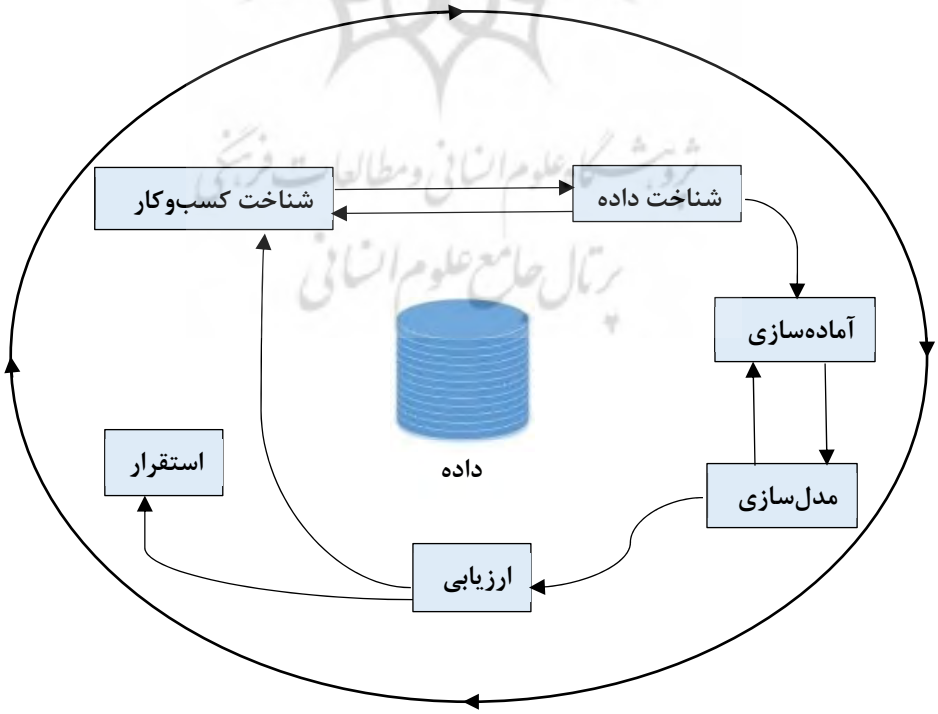
روش‌شناسی پژوهش

صنعت بیمهٔ درمانی با چالش مهمی در پیش‌بینی هزینه‌های بیمهٔ افراد روبه‌رو است. برای مدیریت ریسک و جلوگیری از زیان احتمالی، شرکت‌های بیمهٔ بیمه‌گذاران را به دو گروه پرخطر و کم‌خطر دسته‌بندی می‌کنند. بااین‌حال، برآورد دقیق هزینه‌ها برای هر فرد کار سختی است. در این پژوهش، یک رویکرد مبتنی بر علم داده و یادگیری ماشین پیشنهاد شده است که از یادگیری جمعی برای پیش‌بینی هزینه‌های سلامت افراد بیمه‌شونده استفاده می‌کند و افراد پرخطر و کم‌خطر برای بیمه‌گذار را تعیین می‌کند. پرسش پژوهشی مطرح در این پژوهش این است که چگونه می‌توان هزینه‌های نزدیک به واقعیت را برای افراد بیمه‌شونده محاسبه کرد.

این پژوهش در چارچوب یک مدل فرایندی به نام CRISP مطابق شکل ۱ انجام شده است.

پس از شناخت کسب‌وکار و بررسی عوامل مؤثر بر ریسک مشتریان بیمهٔ سلامت و شناخت داده‌های در دسترس، سایر مراحل به‌ترتیب انجام شده که در ادامه شرح داده شده است.

در پیش‌پردازش، ویژگی‌هایی که مقادیر ناقص داشتند، حذف شد، مهندسی ویژگی‌ها انجام گرفت و ویژگی‌های جدید تولید شد. در ادامهٔ روند پیش‌پردازش داده‌ها، داده‌ها نرمال شد و ویژگی‌های دسته‌ای مدیریت شد. در مرحلهٔ بعدی همان‌طور که در شکل ۲ نشان داده شده است، با استفاده از شش الگوریتم یادگیری ماشین، ریسک درمانی هر فرد پیش‌بینی شد. این شش الگوریتم یادگیری ماشین از گروه‌های مختلف یادگیری ماشین انتخاب شدند. از گروه شبکهٔ عصبی، شبکهٔ عصبی چندلایه در روش پیشنهادی برای پیش‌بینی استفاده شد. شبکهٔ عصبی چندلایه از کلاس‌بندی‌کننده‌های پرکاربرد است (Zhang *et al.*, 2023). از گروه الگوریتم‌های استنتاج آماری با ناظر، رگرسیون لجستیک و ماشین بردار پشتیبان و جنگل تصادفی انتخاب شد. همچنین از روش‌های مبتنی بر درخت، درخت



شکل ۱: مراحل متودولوژی CRISP
Fig. 1: Steps of CRISP methodology

تصمیم، جنگل تصادفی، و ماشین تقویت گرادیان که از روش‌های پرکاربرد برای پیش‌بینی است در مدل پیشنهادی استفاده شد. جنگل تصادفی و ماشین تقویت گرادیان، که توسط مجموعه‌ای از درختان تصمیم‌گیری رگرسیون/طبقه‌بندی تشکیل شده است، دقت پیش‌بینی قابل قبولی را در بسیاری از کاربردها نشان می‌دهد (Bogaert *et al.*, 2021) و نسبت به مقادیر ازدست‌رفته مقاوم‌اند، همچنین به مقیاس‌های ورودی حساس نیستند و مانند الگوریتم‌های محبوب از جمله ماشین بردار پشتیبان و شبکه‌های عصبی، از نظر تنظیم/ آموزش آسان‌ترند. درخت تصمیم، به‌رغم قابلیت تفسیر بالا، معمولاً به اندازهٔ ماشین‌های تقویت گرادیان و جنگل‌های تصادفی برای کارهای پیش‌بینی با ورودی‌های با ابعاد بالا کارا نیست. جنگل تصادفی علاوه‌براینکه یک رقیب قوی برای مدل‌سازی پیش‌بینی‌کننده است، به‌دلیل پیشرفت در مبانی نظری و کاربردهای تجربی آن، نقش محوری در یادگیری ماشین آماری و اقتصادسنجی ایفا می‌کند. استفاده از روش‌های استنتاج مبتنی بر جنگل تصادفی برای تعمیم الگو در آزمون‌های آماری با اهمیت متغیر و خطاهای استاندارد و فواصل اطمینان ارتباط‌های کشف‌نشده استفاده می‌شود. استنتاج آماری مبتنی بر جنگل تصادفی متفاوت با استنتاج نمونهٔ معمولی بوده و معمولاً برای اطمینان از استحکام و تعمیم‌پذیری نمونه‌ها استفاده می‌شود (Chou *et al.*, 2023). از مدل‌های ترکیبی یادگیری ماشین، XGBoost و LightGBM در مدل پیشنهادی استفاده شد.

پارامترهای این شش الگوریتم با روش Grid Search تنظیم شد و سپس نتایج آن‌ها با راهبرد ارزیابی اعتبارسنجی متقاطع K-fold ارزیابی شد. در پایان، نتایج پیش‌بینی این شش الگوریتم با روش ترکیبی میانگین‌گیری وزن‌دار ترکیب شد و به‌عنوان نتیجهٔ نهایی اعلام شد.

آماده‌سازی داده‌ها

ابتدا داده‌های موثق مربوط به هزینهٔ بیمهٔ درمانی تکمیلی برای ده هزار کد ملی تهیه شد. برای استخراج این داده‌ها، از پرونده‌های هزینه از سال ۱۳۹۷ تا ۱۴۰۰ استفاده شد و در ادامه با حواله‌های خسارت تجمیع یافت و با اطلاعات ثبت احوال تطبیق داده شد. همچنین، اطلاعات مربوط به جنسیت، سن و محیط جغرافیایی هر شخص نیز تجمیع شد. علاوه‌براین، متغیرهای دیگر نظیر تعداد کل خسارات هر شخص به تفکیک هر سال، و جمع، میانگین، کمترین و بیشترین مبلغ خسارت هر شخص در هر سال نیز محاسبه شد. با توجه به اینکه داده‌های چهار سال را در اختیار داشتیم، داده‌های سال‌های ۱۳۹۷ تا ۱۳۹۹ را برای ساخت ویژگی‌های مجموعه داده انتخاب کردیم و هدف پیش‌بینی مجموع خسارت بیمه‌شده در سال ۱۴۰۰ است. شکل ۳ نمایشی از مجموعه دادهٔ اولیه را نشان می‌دهد.

پیش‌پردازش

تعداد اولیهٔ رکوردها ۵۶۰۴ بود، اما پس از پیش‌پردازش و حذف رکوردهای دارای دادهٔ مفقوده، تعداد ۵۳۵۶ رکورد برای پردازش انتخاب شد. در روند استخراج اطلاعات، عملیات پیش‌پردازش انجام

شکل ۲: مراحل روش پیشنهادی
Fig. 2: Steps of the proposed method

	age	gender	sum 1400	city name	count 1399	count 1398	count 1397	min 1399	min 1398	min 1397	max 1399	max 1398	max 1397	sum 1399	sum 1398	sum 1397	avg 1399	avg 1398	avg 1397
1	35	1	2496514	تهران	1	7	2	5962640	468000	345600	5962640	6658200	7250582	5962640	23664750	7596182	5962640	3380679	3798091
2	35	1	12762960	تهران	15	8	3	200000	191000	200000	4000000	4000000	5780630	14745260	7186400	6380630	983017.3	898300	2126877
3	34	1	10091474	تهران	29	42	35	243000	24300	30000	3768212	6300000	2624400	41759346	33169062	14222255	1439977	789739.6	406350.1
4	34	1	14466656	تهران	23	19	23	408240	563400	47016	12600000	4099500	6196665	62002351	36733630	28930020	2695754	1933349	1257827
5	34	0	13498545	تهران	40	43	23	102204	46800	41730	5400000	4000000	4000000	42007813	36179521	16343699	1050195	841384.2	710595.6
6	34	1	8332054	تهران	4	8	2	450484	308316	362455	7130650	5764000	2177940	12817566	21156714	2540395	3204392	2644589	1270198

شکل ۳: مجموعه دادهٔ اولیه
Fig. 3: Initial dataset

شد. این عملیات شامل حذف رکوردهای بی‌ارزش، و همچنین اصلاح رکوردهای دارای داده‌های ناقص یا اشتباه به‌دلیل خطای کاربری بود.

نرمال‌سازی: همان‌طور که در شکل ۳ دیده می‌شود برخی ویژگی‌ها حاوی داده‌های پرت هستند، بنابراین استفاده از روش‌های نرمال‌سازی مثل Min Max Normalization که ویژگی‌ها را در بازه‌های مشخص مثل ۰ تا ۱ نرمال می‌کند مناسب نیستند، چراکه این روش‌ها به داده‌های پرت حساس‌اند و عملکرد مدل‌ها را تحت تأثیر قرار می‌دهند. لذا در روش پیشنهادی برای نرمال‌سازی داده‌های عددی از استانداردسازی یا Z transformation استفاده شده است. فرمول ۱ استانداردسازی را نشان می‌دهد به‌طوری‌که z_i نشان‌دهندهٔ نمونهٔ استانداردشده، x_i مقدار اولیهٔ نمونه، x' میانگین نمونه و s انحراف معیار نمونه است.

$$Z_i = \frac{x_i - x'}{s} \quad (۱)$$

ساخت ویژگی‌ها: ساخت ویژگی به‌معنای ایجاد ورودی‌ها یا ویژگی‌های جدید براساس ویژگی‌های موجود است. هدف از ایجاد ویژگی‌های جدید، بهبود عملکرد مدل در حل یک مسئلهٔ یادگیری ماشین است. با استفاده از این روش‌ها می‌توان دقت مدل را افزایش داد. در اینجا، برای بررسی اهمیت و تأثیر ترند، نسبت مراجعات بیماران در سال‌های مختلف نسبت به سال ۱۳۹۹، به‌عنوان ویژگی‌های جدید ساخته شد، که در شکل ۴ مشاهده می‌شوند.

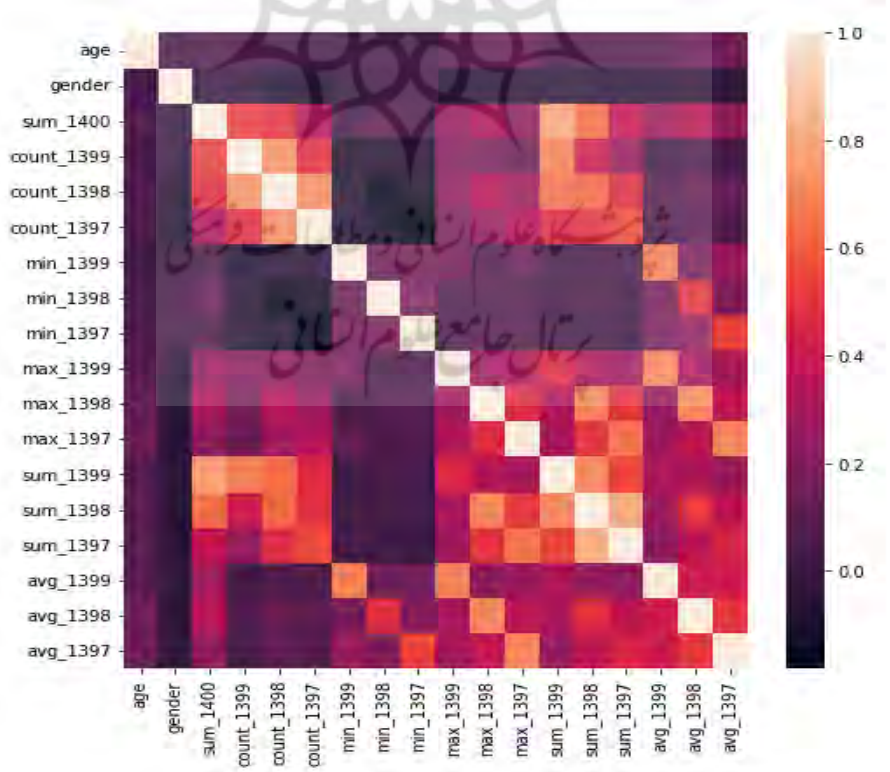
کدگذاری هدف: کدگذاری هدف فرایند جایگزینی مقادیر یک ویژگی دسته‌ای با میانگین متغیر هدف به‌ازای آن مقدار است. در روش پیشنهادی برای متغیر شهر از کدگذاری هدف استفاده شد. به‌طور مثال به‌جای «تهران» در این ستون، میانگین متغیر هدف در کل رکوردهایی که شهر آن‌ها تهران است جایگزین شد. این روش برای متغیرهای اسمی که مقادیر گوناگونی دارند (مثل متغیر شهر) مناسب است. یکی از مزایای این روش همبستگی بالای این کدگذاری با متغیر هدف است.

انتخاب زیرمجموعه‌ای از ویژگی‌ها: ویژگی‌هایی همچون کد ملی، نام و نام خانوادگی که تأثیری در پیش‌بینی ندارند حذف شدند. در ادامه برای انتخاب ویژگی‌ها از تحلیل همبستگی و اهمیت داده‌ها استفاده شد.

برای بررسی تأثیر ویژگی‌ها طبق شکل ۵ نقشهٔ حرارتی رسم شد. نقشهٔ حرارتی رابطهٔ میان ویژگی‌ها را نشان می‌دهد. با تحلیل

count_98_99	sum_98_99	avg_98_99	count_97_99	sum_97_99	avg_97_99

شکل ۴: ویژگی‌های جدید تولیدشده
Fig. 4: New features produced



شکل ۵: نقشهٔ حرارتی
Fig. 5: Heat map

همبستگی، از هر گروه ویژگی‌هایی که همبستگی بالایی با هم دارند می‌توانیم فقط یکی را انتخاب کنیم. در این مرحله دو ویژگی «جمع مبالغ خسارت در سال ۱۳۹۷» و «نسبت میانگین خسارت در سال ۱۳۹۹ به سال ۱۳۹۸» از مجموعهٔ ویژگی‌ها حذف شد.

اهمیت ویژگی: «اهمیت ویژگی» به راهبردهایی اشاره دارد که به متغیرهای ورودی براساس قدرت پیش‌بینی آن‌ها برای متغیر هدف، مقداری اختصاص می‌دهند. امتیازهای اهمیت ویژگی اجزای حیاتی یک پروژهٔ مدل‌سازی پیش‌بینی‌کننده هستند، زیرا بینش‌هایی را در مورد داده‌ها و مدل، و پایه‌ای برای کاهش ابعاد و انتخاب ویژگی ارائه می‌دهند، که می‌تواند کارایی و اثربخشی یک مدل پیش‌بینی را بر روی مسئله بهبود بخشد (Boutahir *et al.*, 2022). در این مرحله از خاصیت اهمیت ویژگی الگوریتم LightGBM استفاده شد و مطابق شکل ۵ متغیرهای «سن»، «نسبت مجموع خسارت در سال ۱۳۹۹ به ۱۳۹۷» و «تعداد خسارت در سال ۱۳۹۷» به‌دلیل اهمیت بسیار پایینی که داشتند از ویژگی‌ها حذف شدند.

تعیین متغیر هدف باینری برای مسئله: برای حل مسئله و پیش‌بینی آینده در روش پیشنهادی علاوه‌بر جنبهٔ کمّی مسئله بر جنبهٔ کیفی مسئله مانند شهر و جنسیت نیز توجه شد و مسئله به‌صورت کلاس‌بندی باینری تعریف شد. برای این منظور پرونده‌های خسارت به سه دسته تقسیم شد. به این صورت که میانهٔ خسارت‌های مجموعه داده محاسبه شد و ۲۰٪ کمتر از میانه، به‌عنوان حد پایین و ۲۰٪ بیشتر از میانه، به‌عنوان حد بالا تعریف شد. پرونده‌هایی که بین حد پایین و حد بالا قرار گرفتند به‌عنوان کلاس «ریسک متوسط» از مجموعه داده حذف شد تا الگوریتم‌ها قدرت بیشتری برای تفکیک دو کلاس «ریسک بالا» و «ریسک پایین» داشته باشند. نمودار شکل ۶ تفکیک این سه کلاس را نشان می‌دهد، به‌طوری که کلاس ۰ تعداد رکوردهای «ریسک پایین» و کلاس ۱ تعداد رکوردهای «ریسک بالا» و کلاس ۲ تعداد رکوردهای «ریسک متوسط» را نشان می‌دهد.

مدل‌سازی

در ادامه از روش‌های شبکهٔ عصبی پرسپترون چندلایه، ماشین بردار پشتیبان XGBoost، LightGBM (Ke *et al.*, 2017)، (Vapnik, 1999).

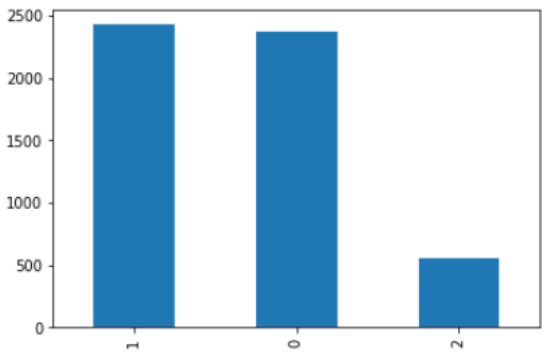
(Chen and He, 2016)، رگرسیون لجستیک (Lee *et al.*, 2006) و جنگل تصادفی (Ho, 1995) برای پیش‌بینی استفاده شد.

شبکهٔ عصبی پرسپترون چندلایه، ساده‌ترین مدل شبکهٔ عصبی است که از نورون‌هایی به نام پرسپترون تشکیل شده است. از چندین ورودی، پرسپترون، یک خروجی را با توجه به وزن‌ها و توابع فعال‌سازی غیرخطی آن محاسبه می‌کند. اساساً شبکهٔ عصبی پرسپترون چندلایه از لایهٔ ورودی، یک یا چند لایهٔ پنهان و لایهٔ خروجی پرسپترون محاسباتی تشکیل شده است. شبکهٔ عصبی پرسپترون چندلایه مدلی برای یادگیری نظارت‌شده است که از الگوریتم پس‌انتشار استفاده می‌کند.

XGBoost یک روش یادگیری ماشین پیشرفته است که براساس ایدهٔ اصلی گرادیان تقویتی توسعه یافته است. این الگوریتم با استفاده از الگوریتم یادگیری ترکیب چندین درخت تصمیم قدرتمند، یک مدل از ابتدا تا انتها ایجاد می‌کند که قادر است مشکلات پیچیده را تشخیص دهد و دسته‌بندی دقیقی انجام دهد. مزیت استفاده از رویکرد XGBoost این است که استخراج امتیازهای مربوط به هر ویژگی پس از تولید درختان تقویت‌شده به‌طور منطقی آسان است. هرچه صفت متداول‌تر در درخت‌های تصمیم‌گیری برای قضاوت‌های مهم به کار گرفته شود، اهمیت نسبی آن بیشتر می‌شود. سپس اهمیت ویژگی‌ها در تمام درخت‌های تصمیم در مدل به‌طور میانگین محاسبه می‌شود (Brownlee, 2016).

الگوریتم LightGBM نسخه‌ای قوی از روش تقویتی است که شبیه به XGBoost است، اما در چند جنبهٔ کلیدی متفاوت است، مانند نحوهٔ ایجاد درخت یا پایهٔ یادگیرندگان. در مقایسه با دیگر راهبردهای این گروه، LightGBM درختان را برگ‌به‌برگ توسعه می‌دهند، که خطا را در طول مرحلهٔ تقویت متوالی به حداقل می‌رساند (Effrosynidis and Arampatzis, 2021).

ماشین بردار پشتیبان توسط Vapnik (1999)، در حوزهٔ نظریهٔ یادگیری آماری و به‌منظور به حداقل رساندن ریسک ساختاری پیشنهاد شد. این الگوریتم در بسیاری از مسائل تشخیص الگو و تخمین رگرسیون استفاده شده و برای مسائل تخمین وابستگی، پیش‌بینی و ساخت ماشین‌های هوشمند به کار رفته است.



شکل ۶: نتایج کلاس‌بندی داده‌ها
Fig. 6: Data classification results

جنگل تصادفی، یک روش ترکیبی است که در آن یک طبقه‌بندی‌کننده با ترکیب چندین درخت تصمیم پایهٔ مستقل مختلف ساخته می‌شود. این روش به‌عنوان bagging یا bootstrap aggregation شناخته می‌شود. در این الگوریتم، هر درخت تصمیم با بخشی از ویژگی‌ها که به شکل تصادفی انتخاب می‌شوند، آموزش داده می‌شود.

رگرسیون لجستیک نیز یک روش مدل‌سازی آماری پرکاربرد است، می‌تواند مدلی با متغیر هدف دوکلاسه بسازد و به‌عنوان یک الگوریتم قدرتمند شناخته می‌شود (Lee *et al.*, 2006).

اعتبارسنجی

در این مطالعه، از روش اعتبارسنجی متقاطع K-fold برای ارزیابی و اعتبارسنجی استفاده شده است. در این روش، مجموعه داده به K زیرمجموعه تقسیم می‌شود، که در این پژوهش ۵ در نظر گرفته شده است و در هر مرحله، یک زیرمجموعه برای اعتبارسنجی و K-1 زیرمجموعه دیگر برای آموزش استفاده می‌شود. این فرایند K بار تکرار می‌شود و تمام داده‌ها به‌صورت کامل برای اعتبارسنجی استفاده می‌شوند. نتیجهٔ ارزیابی در کل تکرارها، به‌عنوان ارزیابی نهایی در نظر گرفته می‌شود.

ترکیب مدل‌ها به روش میانگین‌گیری وزنی

در نهایت، با استفاده از رش میانگین وزنی و تنظیم وزن هریک از روش‌ها، پیش‌بینی نهایی حاصل شد. در این روش، وزن‌ها براساس آزمایش‌ها و بررسی‌های انجام‌شده تنظیم می‌شود. ایدهٔ اصلی در این روش کاهش کل خطاها با تجمیع پیش‌بینی‌های چند طبقه‌بندی مختلف است. منطق این روش به این صورت است که در ابتدا فرض می‌کنیم هر الگوریتم در آموزش و پیش‌بینی خود، اشتباهات مختلف را دارد. سپس، مجموعه‌ای از الگوریتم‌ها با تنوع زیاد آموزش می‌دهیم و خروجی آن‌ها را با هم ترکیب می‌کنیم. بدین‌ترتیب، مجموع خطاهای پیش‌بینی نهایی پس از انجام تجمیع به‌نحوی کاهش می‌یابد.

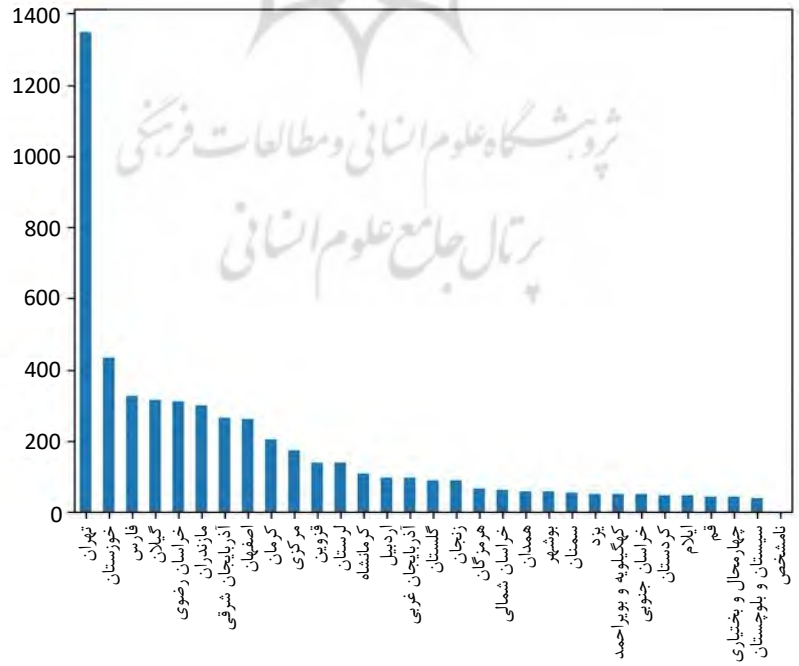
نتایج و بحث

برای ارزیابی، روش پیشنهادی در نسخهٔ ۳ پایتون و محیط ژوپیتر پیاده‌سازی شد. شکل ۷ فراوانی شهرهای مختلف را در مجموعه داده و شکل ۸ توزیع متغیر هدف برحسب جنسیت نشان می‌دهد.

شکل ۹ هیستوگرام متغیر «مجموع خسارت سال ۱۴۰۰» را که پایهٔ محاسبهٔ متغیر هدف است، نشان می‌دهد. با توجه به اینکه این متغیر دارای داده‌های پرت با مقادیر بسیار بزرگ است، در شکل ۱۰ هیستوگرام لوگاریتم همان متغیر با توزیع نرمال‌شده، مشاهده می‌شود.

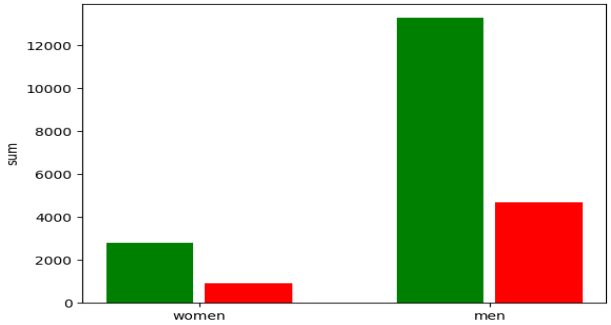
شکل ۱۱، هیستوگرام متغیر سن در مجموعه داده را نشان می‌دهد. همان‌طور که در شکل دیده می‌شود، متغیر سن دارای توزیع نرمال است و میانه‌ای حدود ۴۰ سال دارد.

برای مقایسهٔ روش پیشنهادی و مقایسهٔ آن با روش‌های دیگر از مفهوم منحنی راک استفاده شد. مشخصهٔ عملکرد سیستم یا منحنی راک، نموداری گرافیکی است که توانایی تشخیص یک سیستم اندازه‌گیری طبقه‌بندی باینری را نشان می‌دهد. ناحیهٔ زیرمنحنی راک که AUC نامیده می‌شود، عددی بین صفر تا یک است و نشان

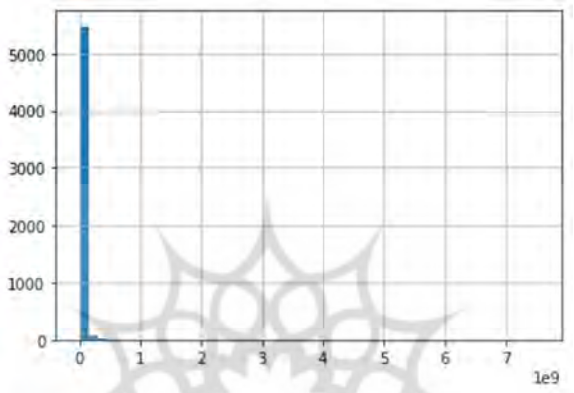


شکل ۷: تعداد شهرهای مختلف در مجموعه داده

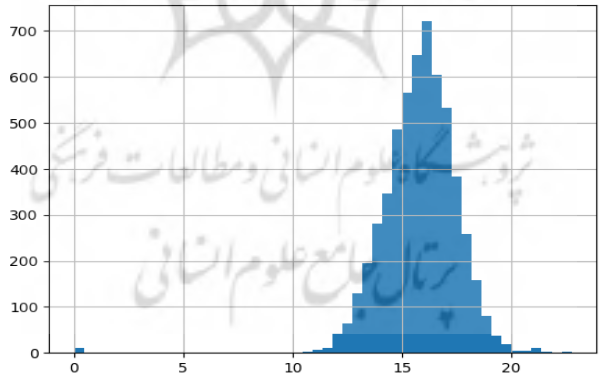
Fig. 7: Number of different cities in the dataset



شکل ۸: توزیع متغیر هدف برحسب جنسیت
Fig. 8: Distribution of target variable by gender



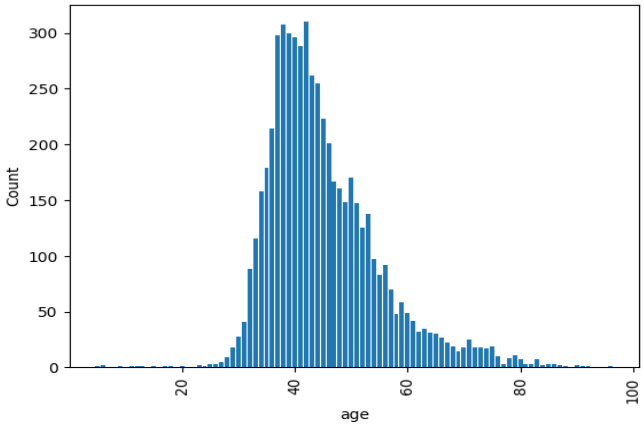
شکل ۹: هیستوگرام متغیر «مجموع خسارت سال ۱۴۰۰»
Fig. 9: Histogram of the variable “total damage in 1400”



شکل ۱۰: هیستوگرام لگاریتم متغیر «مجموع خسارت سال ۱۴۰۰»
Fig. 10: Histogram of the variable logarithm “Total damage in 1400”

می‌دهد قدرت تشخیص یا درستی نتایج یک آزمون یا طبقه‌بند چقدر است. این معیار مخصوصاً زمانی کاربرد دارد که هزینهٔ مثبت غلط و منفی غلط (اشتباهات در تشخیص هریک از کلاس‌ها) متفاوت باشد، زیرا این معیار، تعادل بین نرخ مثبت صحیح (حساسیت) و نرخ مثبت غلط (خاصیت) در آستانه‌های مختلف را در نظر می‌گیرد. با تنظیم آستانه، می‌توان به طبقه‌بندی رسید که با توجه به هزینهٔ مثبت غلط و منفی غلط یک مسئلهٔ خاص، بین حساسیت و خاصیت اولویت‌بندی کند. نتیجهٔ AUC روش‌های مختلف در جدول ۱ ذکر شده است.

در نهایت با روش میانگین‌گیری وزنی، نتیجهٔ نهایی AUC=0.7334 حاصل شد که بهتر از تمام روش‌های پایه است. وزن بهینه که برای هر الگوریتم مشخص شد در جدول ۲ آمده است. شکل ۱۲ اهمیت هریک از ویژگی‌ها در الگوریتم LightGBM



شکل ۱۱: هیستوگرام متغیر سن در مجموعه داده
Fig. 11: Histogram of the age variable in the dataset

جدول ۱: AUC در روش‌های مختلف
Table 1: AUC in different methods

روش	AUC
XGBoost	0.725
جنگل تصادفی	0.726
رگرسیون لجستیک	0.726
ماشین بردار پشتیبان	0.727
شبکهٔ عصبی پرسپترون چندلایه	0.728
LightGBM	0.727

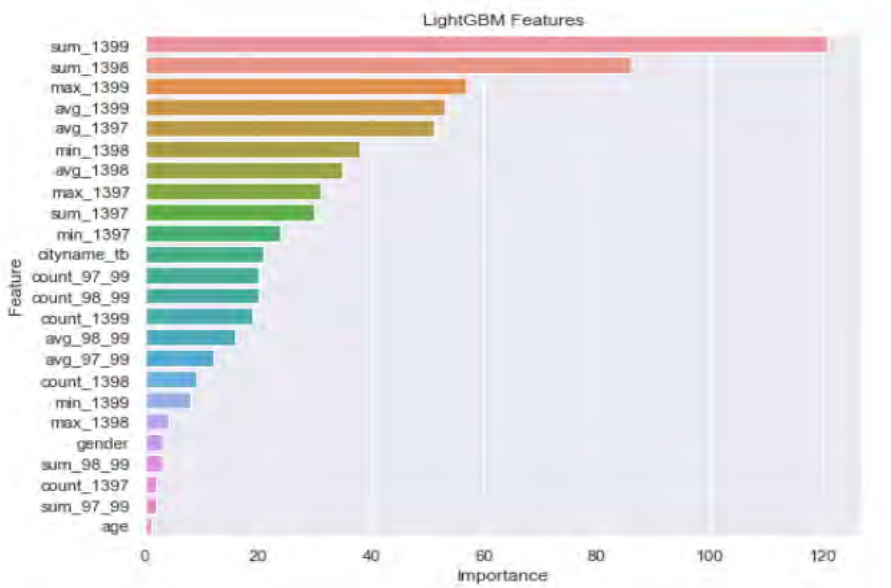
جدول ۲: وزن الگوریتم‌ها در روش ترکیبی
Table 2: The weight of algorithms in the hybrid method

روش	وزن
XGBoost	0.05
جنگل تصادفی	0.1
رگرسیون لجستیک	0.1
ماشین بردار پشتیبان	0.25
شبکهٔ عصبی پرسپترون چندلایه	0.25
LightGBM	0.25

را نشان می‌دهد. همان‌طور که در این شکل مشخص است مجموع خسارت دو سال اخیر به‌عنوان مهم‌ترین ویژگی‌ها در پیش‌بینی ریسک سال آینده شناخته شده است.

جمع‌بندی و پیشنهادها

رشتهٔ درمان پس از رشتهٔ ثالث یکی از پرتقاضاترین رشته‌ها در صنعت بیمه است. از مهم‌ترین چالش‌های صنعت بیمه، پیش‌بینی هزینه‌های بیمهٔ درمانی افراد است. برای برآورد هزینه‌های بیمه و نرخ‌گذاری بیمه، پیش‌بینی هزینهٔ بیمهٔ درمانی مورد نیاز است. در حال حاضر هزینه‌های درمانی معمولاً براساس ترکیبی از پارامترها، مانند ویژگی‌های فیزیکی فرد بیمه‌شونده، سنجیده می‌شود که کاری پیچیده است. در این پژوهش با استفاده از روش‌های داده‌کاوی سعی در پیش‌بینی هزینه‌های درمانی افراد بیمه‌شونده می‌شود. روش پیشنهادی از اعتبارسنجی متقابل برای ارزیابی نتایج مدل‌سازی استفاده می‌کند، همچنین روش پیشنهادی با یادگیری جمعی و با ترکیب روش‌های رگرسیون لجستیک، شبکه‌های عصبی پرسپترون چندلایه، ماشین بردار پشتیبان، و جنگل تصادفی، XGBoost، LightGBM، به AUC برابر با ۰/۷۳ دست می‌یابد که اثربخشی آن را در پیش‌بینی افراد پرخطر و کم‌خطر نشان می‌دهد. با استفاده از علم داده و روش‌های یادگیری ماشین، شرکت‌های



شکل ۱۲: اهمیت ویژگی‌ها نسبت به ویژگی هدف از نظر الگوریتم LightGBM

Fig. 12: The importance of features compared to the target feature according to the LightGBM algorithm

بیمه می‌توانند دقت برآورد هزینهٔ خود را بهبود بخشند و ریسک را بهتر مدیریت کنند. این رویکرد می‌تواند به شرکت‌های بیمه کمک کند تا پوشش بیمه‌ای و قیمت‌گذاری دقیق‌تری را برای افراد ارائه دهند که به رضایت بیشتر مشتریان و کاهش زیان‌های مالی منجر می‌شود.

با توجه به ویژگی‌های محدودی که در این پژوهش استفاده شده است، برای کارهای آتی پیشنهاد می‌شود که از متغیرهای محاسباتی بیشتر برای افزایش دقت استفاده شود. در ضمن به‌جز روش‌های یادگیری ماشین، برای قیمت‌گذاری می‌توان از روش‌های تصمیم‌گیری با معیارهای چندگانه نیز استفاده کرد. روش‌هایی مانند تاپسیس یا فرایند تحلیل سلسله‌مراتبی مناسب برای این امر هستند. همچنین می‌توان در کارهای آتی به‌جای طبقه‌بندی افراد در دو طبقهٔ پرخطر و کم‌خطر، تعداد سه طبقه را در نظر گرفت و افراد با هزینهٔ درمانی متوسط را نیز در آن لحاظ کرد.

مشارکت نویسندگان

مهسا تجددی نودهی: جمع‌آوری داده‌ها و گردآوری مطالب و نگارش مقاله، سمانه حسینی خطیبانی: مدل‌سازی و گردآوری مطالب و نگارش مقاله، محسن یزدینژاد: استاد راهنما و پیاده‌سازی الگوریتم، سمیه زلفی: گردآوری مطالب و نگارش مقاله.

تشکر و قدردانی

از پیشنهادهای داوران محترم که به غنای علمی مقاله کمک کردند، بسیار سپاسگزاریم.

تعارض منافع

نویسندگان اعلام می‌کنند که هیچ تضاد منافعی در مورد انتشار پژوهش ثبت‌شده وجود ندارد. علاوه‌براین، موارد اخلاقی از جمله سرقت ادبی، رضایت آگاهانه، رفتار نادرست، جعل و/یا جعل داده‌ها، انتشار مضاعف و یا سوءرفتار به‌طور کامل ازسوی نویسندگان رعایت شده است.

یادداشت ناشر

ناشر نشریهٔ پژوهشنامهٔ بیمه با توجه به مرزهای حقوقی در نقشه‌های منتشرشده بی‌طرف باقی می‌ماند.

منابع

Ahmadlou, Y.; Pourebrahimi, A.;Tanha, J.; Rajabzadeh, A., (2023). Presenting a hybrid model for identifying claims of suspicious damages in agricultural insurance. J. Insur. Res., 12(1): 63-78 **(16 Pages)**. [In Persian]

Albalawi, S.; Alshahrani, L.; Albalawi, N.; Alharbi, R., (2023). Prediction of healthcare insurance costs. Comput. Inf., 3(1): 9-18 **(10 Pages)**.

Anwar ul Hassan, Ch.; Iqbal, J.; Hussain, S.; AlSalman, H.; Mosleh, M.A.; Sajid Ullah, S., (2021). A computational intelligence approach for predicting medical insurance cost. Math. Probl. Eng., 2021: 1-13 **(13 Pages)**.

Babichev, S.; Korobchynskyi, M.; Lahodynskyi, O.; Korchomnyi, O.; Basanets, V.; Borynskyi, V., (2018). Development of a technique for the reconstruction and validation of gene network models based on gene expression profiles. J. Enterp. Technol., 1(4): 19-32 **(14 Pages)**.

Benedek, B.; Ciumas, C.; Nagy, B.Z., (2022). Automobile insurance fraud detection in the age of big data – A systematic and comprehensive literature review. J. Financ. Regul. Compliance., 30(4): 503-523 **(21 Pages)**.

Bhardwaj, N.; Anand, R., (2020). Health insurance amount prediction. J. Eng. Res. Technol., 9: 1008-1011 **(4 Pages)**.

Bodyanskiy, Y.; Vynokurova, O.; Pliss, I.; Peleshko, D., (2017). Hybrid adaptive systems of computational intelligence and their online learning for green it in energy management tasks., 229-244 **(16 Pages)**.

Bogaert, M.; Ballings, M.; Bergmans, R.; Van den Poel, D., (2021). Predicting self-declared movie watching behavior using facebook data and information-fusion sensitivity analysis. J. Decis. Sci. Inst., 52(3): 776-810 **(35 Pages)**.

Boutahir, M.K.; Farhaoui, Y.; Azrour, M.; Zeroual, I.; El Allaoui, A., (2022). Effect of feature selection on the prediction of direct normal irradiance. Big. Data. Min. Anal., 5(4): 309-317 **(9 Pages)**.

Brownlee, J., (2016). Feature importance and feature selection with XGBoost in python.

Chen, T.; He, T., (2016). XGBoost: A scalable tree boosting system. In proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining., 785-794 **(10 Pages)**.

Chou, Y.C.; Chuang, H.H.C.; Chou, P.; Oliva, R., (2023). Supervised machine learning for theory building and testing: Opportunities in operations management. J. Oper. Manage., 69(4): 643-675 **(33 Pages)**.

Christobel, Y.A.; Subramanian, S., (2022). An empirical study of machine learning regression models to predict health insurance cost. Webology., 19(2).

Chyrun, L.; Vysotska, V.; Kis, I.; Chyrun, L., (2018). Content analysis method for cut formation of human psychological state. ., 139-144 **(6 Pages)**.

Doroshenko, A., (2018). Piecewise-linear approach to classification based on geometrical transformation model for imbalanced dataset., 231-235 **(5 Pages)**.

Drewe-Boss, P.; Enders, D.; Walker, J.; Ohler, U., (2022). Deep learning for prediction of population health costs. BMC. Med. Inf. Decis. Making., 22(1): 1-10 **(10 Pages)**.

Du, Y.; Yang, C.; Zhao, B.; Hu, C.; Zhang, H.; Yu, Z.; Wang, H., (2023). Optimal design of a supercritical carbon dioxide recompression cycle using deep neural network and data mining techniques. Energy., 271.

Effrosynidis, D.; Arampatzis, A., (2021). An evaluation of feature selection methods for environmental data., 61.

Eriksson, K.; Estep, D.; Johnson, C., (2004). Applied mathematics: Body and soul. ., 1: 741-753 **(13 Pages)**.

Fauzan, M.A.; Murfi, H., (2018). The accuracy of XGBoost for insurance claim prediction. Int. J. Adv. Soft. Comput. Appl., 10(2): 159-171 **(13 Pages)**.

Goundar, S.; Prakash, S.; Sadal, P.; Bhardwaj, A., (2020). Health insurance claim prediction using artificial neural networks. J. Syst. Dyn. Appl., 9(3): 40-57 **(18 Pages)**.

Hanafy, M.; Omar, M.A.M., (2021). Predict health insurance cost by using machine learning and DNN regression models. Int. J. Innov. Technol. Explor. Eng., 10(3): 137-143 **(7 Pages)**.

Ho, T.K., (1995). Random decision forests. In proceedings of 3rd international conference on document analysis and recognition., 1: 278-282 **(5 Pages)**.

Kafuria., A.D., (2022). Predictive model for computing health insurance premium rates using machine learning algorithms. J. Comput., 44(1): 21-38 **(18 Pages)**.

Kaushik, K.; Bhardwaj, A.; Dwivedi, A.D.; Singh, R., (2022). Machine learning-based regression framework to predict health insurance premiums. J. Environ. Res. Public. Health., 19(13).

Ke, G.; Meng, Q.; Finley, T.; Wang, T.; Chen, W.; Ma, W.; Ye, Q.; Liu, T.Y., (2017). LightGBM: A highly efficient gradient boosting decision tree. Adv. Neural. Inf. Process. Syst., 30: 3149-3157 **(9 Pages)**.

Kumar Sharma, D.; Sharma, A., (2020). Prediction of health insurance emergency using multiple linear regression technique. Eur. J. Mol. Clin. Med., 7: 95-105 **(11 Pages)**.

Kumar, M.; Ghani, R.; Mei, Z.S., (2010). Data mining to predict and prevent errors in health insurance claims processing. In proceedings of the 16th ACM SIGKDD international conference on knowledge discovery and data mining., 65-74 **(10 Pages)**.

Lakshmanarao, A.; Koppireddy, C.S.; Kumar, G.V., (2020). Prediction of medical costs using regression algorithms. J. Inf. Comput. Sci., 10(5): 751-757 **(7 Pages)**.

Lee, T.S.; Chiu, C.C.; Chou, Y.C.; Lu, C.J., (2006). Mining the customer credit using classification and regression tree and multivariate adaptive regression splines. Comput. Stat. Data. Anal., 50(4): 1113-1130 **(18 Pages)**.

Marmolejo-Ramos, F.; Tejo, M.; Brabec, M.; Kuzilek, J.; Joksimovic, S.; Kovanovic, V.; Ospina, R., (2023). Distributional regression modeling via generalized additive models for location, scale, and shape: An overview through a data set from learning analytics. Wiley. Interdiscip. Rev. Data. Min. Knowl. Discovery., 13(1).

Milovic, B.; Milovic, M., (2012). Prediction and decision making in health care using data mining. Kuwait. Chapter. Arabian. J. Bus. Manage. Rev., 1(12): 1-11 **(11 Pages)**.

Morid, M.A.; Kawamoto, K.; Ault, T.; Dorius, J.; Abdelrahman, S., (2017). Supervised learning methods for predicting

healthcare costs: Systematic literature review and empirical evaluation., 2017: 1312-1321 **(10 Pages)**.

Park, S.B.; Oh, S.K.; Kim, E.H.; Pedrycz, W., (2023). Rule-based fuzzy neural networks realized with the aid of linear function prototype-driven fuzzy clustering and layer reconstruction-based network design strategy. Expert. Syst. Appl., 219.

Perova, I.; Pliss, I., (2017). Deep hybrid system of computational intelligence with architecture adaptation for medical fuzzy diagnostics. Int. J. Intell. Syst. Appl., 9(7): 12-21 **(10 Pages)**.

Pfutzenreuter, T.C.; Lima, E.P., (2021). Machine learning in healthcare management for medical insurance cost prediction.

Sepahvand, S.; Ramandi, S.; Mahmoudvand, R., (2022). Identifying customers' risk in auto insurance and calculating distorted insurance premiums. Iran. J. Insur. Res., 11(4): 321-338 **(18 Pages)**.

Shakhovska, N.; Melnykova, N.; Chopiyak, V., (2022). An ensemble methods for medical insurance costs prediction task. Comput. Mater. Continua., 70(2).

Shakhovska, N.; Veres, O.; Bolubash, Y.; Bychkovska-Lipinska, L., (2015). Data space architecture for big data managering. In 2015 Xth international scientific and technical conference computer sciences and information technologies (CSIT)., 184-187 **(4 Pages)**.

Shyamala Devi, M.; Swathi, P.; Purushotham Reddy, M.; Deepak Varma, V.; Praveen Kumar Reddy, A.; Vivekanandan, S.; Moorthy, P., (2021). Linear and ensembling regression based health cost insurance prediction using machine learning. In smart computing techniques and applications: Proceedings of the fourth international conference on smart computing and informatics., 2.

Sommers, B.D., (2020). Health insurance coverage: What comes after the ACA?. Health. Aff., 39(3): 502-508 **(7 Pages)**.

Tkachenko, R.; Izonin, I.; Kryvinska, N.; Chopyak, V.; Lotoshynska, N.; Danylyuk, D., (2018). Piecewise-linear approach for medical insurance costs prediction using SGTM neural-like structure. IDDM., 21: 170-179 **(10 Pages)**.

Vapnik, V., (1999). An overview of statistical learning theory. IEEE. Trans. Neural. Netw., 10(5): 988-999 **(12 Pages)**.

Vijayalakshmi, V.; Selvakumar, A.; Panimalar, K., (2023). Implementation of medical insurance price prediction system using regression algorithms. In 2023 5th international conference on smart systems and inventive technology (ICSSIT)., 1529-1534 **(6 Pages)**.

Yang, C.; Delcher, C.; Shenkman, E.; Ranka, S., (2018). Machine learning approaches for predicting high cost high need patient expenditures in health care. Biomed. Eng. Online., 17(1): 1-20 **(20 Pages)**.

Zhang, J.; Li, C.; Yin, Y., (2023). Applications of artificial neural networks in microorganism image analysis: A comprehensive review from conventional multilayer perceptron to popular convolutional neural network and potential visual transformer. Artif. Intell. Rev., 56(2): 1013-1070 **(58 Pages)**.

معرفی نویسندگان — AUTHOR(S) BIOSKETCHES

مهسا تجددی نودهی، دانشجوی کارشناسی ارشد مهندسی کامپیوتر، گروه کامپیوتر، دانشکده کامپیوتر، موسسه آموزش عالی آل طه، تهران، ایران

- Email: *mahsa.tajaddodi@gmail.com*
- ORCID: 0009-0003-7310-4345
- Homepage: https://aletaha.ac.ir/fa

سمانه حسینی خطیبانی، دانشجوی کارشناسی ارشد مهندسی کامپیوتر، گروه کامپیوتر، دانشکده کامپیوتر، موسسه آموزش عالی آل طه، تهران، ایران

- Email: *Samaneh.hosseini62@gmail.com*
- ORCID: 0009-0003-9668-6093
- Homepage: https://aletaha.ac.ir/fa

محسن یزدی نژاد، دانشجوی دکترای هوش مصنوعی، گروه کامپیوتر، دانشکده کامپیوتر دانشگاه اصفهان، اصفهان، ایران

- Email: *Mohsen.yazdinejad@eng.ui.ac.ir*
- ORCID: 0000-0001-7805-6344
- Homepage: http://www.fad.ir/Teacher/Details/1039

سمیه زلفی، دانشجوی کارشناسی ارشد مهندسی کامپیوتر، گروه کامپیوتر، دانشکده کامپیوتر، موسسه آموزش عالی آل طه، تهران، ایران

- Email: *S.zolfi1365@gmail.com*
- ORCID: 0009-0001-2621-626x
- Homepage: https://aletaha.ac.ir/fa

HOW TO CITE THIS ARTICLE

Tajaddodi Nodehi, M.; Hosseini Khatibani, S.; Yazdinejad, M.; Zolfi, S., (2024). Predicting people's health insurance costs using machine learning and ensemble learning methods. Iran. J. Insur. Res., 13(1): 1-14.

DOI: 10.22056/ijir.2024.01.01

URL: https://ijir.irc.ac.ir/article_160311.html?lang=en