



ORIGINAL RESEARCH PAPER

Providing the concept of risk package instead of risk factor in order to classify the risk of policyholders more accurately

M. Esna-Ashari¹, F. Khamesian^{2*}, F. Khanizadeh¹

¹ Department of property and Casualty Insurance, Iranian Insurance Research Center, Tehran, Iran

² Department of General Insurance, Iranian Insurance Research Center, Tehran, Iran

ARTICLE INFO

Article History:

Received 10 June 2023

Revised 30 July 2023

Accepted 01 October 2023

Keywords:

Clustering

Risk package

Third party

Unsupervised algorithm

*Corresponding Author:

Email: khamesian@irc.ac.ir

Phone: +9821 22084084

ORCID: [0000-0001-6113-4246](https://orcid.org/0000-0001-6113-4246)

ABSTRACT

BACKGROUND AND OBJECTIVES: The accurate and scientific assessment of the risk to issue an insurance policy is one of the most critical and important stages of risk assessment frameworks. This leads companies to identify high-risk customers and determine the policy rates in accordance with their risks, and as a result, the claims will be covered appropriately through the insurance premiums. In this paper, a new method is presented to define the concept of risk factor in more practical, flexible and accurate way. In this method, which is based on an unsupervised clustering algorithm, initially, every single factor is examined based on different ranges and their corresponding impact on customer loss levels. Then, considering their connection with the ranges of other factors in terms of creating similar levels of customer loss, they are combined to form a package. Thus, different packages are created, each of which is considered a risk factor and comprise the ranges of factors affecting different levels of loss.

METHODS: The k-means clustering method was used to divide insurers into clusters with similar risks, which correspond to the risk packages associated with the customers' risk level. The number of desired clusters should be determined in advance, which is the main challenge of using this algorithm. Two main approaches for validation, namely the silhouette score and the elbow method, were presented.

FINDINGS: Based on the elbow plot and silhouette coefficient, as well as considering the practical and realistic evaluation needed by insurance companies, four clusters were obtained. Cluster 2 and 3 are similar and can be merged to form a cluster of medium risk level. Therefore, three clusters were considered the best outcome for categorizing insurance policyholders.

CONCLUSION: The risk packages can be introduced from the examination of the 3 clusters including People with high, medium and low age (confidence interval) with low price car whose gender is male can be introduced as the highest level of risk; People with medium and high ages (confidence interval) with medium and high car prices can be considered as medium risks, and Middle-aged and older people (confidence interval) with expensive cars were considered the lowest level of risk. From the results of these risk packages, it can be concluded that although a significant population of older policyholders falls into the first package (first cluster), they have the highest level of risk. On the other hand, the older people in the third package (even though their average age is the highest among the clusters) have the lowest level of risk. Another important point is that the risk level decreases as income increases simultaneously with age.

DOI: [10.22056/ijir.2024.01.02](https://doi.org/10.22056/ijir.2024.01.02)

This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).





مقاله علمی

ارائه مفهوم ریسک پکیج به جای ریسک فاکتور به منظور طبقه‌بندی دقیق تر ریسک بیمه‌گذاران

مریم اثنی‌عشری^۱، فرزانه خامسیان^{۲*}، فرید خانی‌زاده^۱

^۱ گروه پژوهشی بیمه‌های اموال و مسئولیت، پژوهشکده بیمه، تهران، ایران

^۲ گروه پژوهشی مطالعات عمومی بیمه، پژوهشکده بیمه، تهران، ایران

چکیده:

پیشینه و اهداف: ارزیابی صحیح و علمی ریسک صدور بیمه‌نامه یکی از حساس‌ترین و مهم‌ترین مراحل ارزیابی ریسک است و انجام آن باعث شناسایی مشتریان پرریسک و تعیین نرخ بیمه‌نامه، متناسب با ریسک مشتریان و در نتیجه پوشش مناسب خسارت‌های مالی ادعا شده به وسیله حق بیمه‌های دریافتی می‌شود. در این پژوهش روشی جدید برای تبیین دقیق‌تر و کاربردی‌تر از ریسک فاکتور ارائه شده است. در این روش که مبتنی بر الگوریتم بدون نظارت خوشه‌بندی است، ابتدا بازه‌های مختلف هر عامل مؤثر بر خسارت بررسی و با توجه به میزان تأثیرگذاری بر سطوح خسارت مشتریان به چند ریسک فاکتور تقسیم می‌شوند. سپس با توجه به میزان ارتباط آن با بازه دیگر عوامل، از لحاظ ایجاد سطوح خسارت مشابه در مشتریان، با آن‌ها ترکیب می‌شود و پکیجی شامل بازه‌های عوامل تأثیرگذار بر سطوح مختلف خسارت را تشکیل می‌دهد. به این ترتیب به جای یک ریسک فاکتور، پکیج‌های مختلفی ایجاد می‌شود که هر کدام از آن‌ها یک عامل ریسک یا همان ریسک فاکتور در نظر گرفته می‌شوند.

روش شناسی: با استفاده از روش خوشه‌بندی کامیانتگین، بیمه‌گذاران به خوشه‌هایی با ریسک همگن که در واقع ریسک پکیج‌های متناظر با میزان پرخطر بودن مشتریان هستند، تقسیم شده‌اند. بر اساس ساختار الگوریتم کامیانتگین تعداد خوشه‌های مورد نظر باید از پیش تعیین شود. این موضوع چالش اصلی استفاده از الگوریتم مزبور است. در همین راستا دو رویکرد اصلی اعتبارسنجی سایه‌نما (ضریب سیلوئت) و روش آرنج برای حل این مشکل ارائه شده است.

یافته‌ها: با توجه به نمودار آرنج و ضریب سیلوئت و همچنین در نظر گرفتن نیاز شرکت‌های بیمه به ارزیابی کاربردی و منطبق بر واقعیت، ۴ خوشه به دست آمد که با توجه به اینکه خوشه ۲ و ۳ در یک طیف نزدیک به هم و در نتیجه قابل پیوستن به یکدیگر هستند و خوشه با سطح ریسک متوسط را تشکیل می‌دهند، ۳ خوشه به‌عنوان بهترین خروجی دسته‌بندی بیمه‌گذاران لحاظ شد.

نتیجه‌گیری: از بررسی ویژگی‌های به‌دست آمده در ۳ خوشه مطرح شده می‌توان پکیج‌های ریسک ذیل را معرفی کرد.
- افراد با سنین بالا، متوسط و پایین (چگال در بازه ۳۰ تا ۵۸ سال) با ماشین ارزان‌قیمت و دارای جنسیت مرد را می‌توان به‌عنوان بیمه‌گذاران با بالاترین سطح ریسک معرفی کرد.

- افراد با سنین متوسط و بالا (چگال در بازه ۳۲ تا ۵۳ سال) با ارزش ماشین متوسط و بالا را می‌توان بیمه‌گذاران دارای ریسک‌های متوسط در نظر گرفت.

- افراد با سنین متوسط به بالا (چگال در بازه ۵۱ تا ۶۳ سال) با ماشین گران‌قیمت را می‌توان بیمه‌گذاران با پایین‌ترین سطح ریسک در نظر گرفت.

اطلاعات مقاله

تاریخ‌های مقاله:

تاریخ دریافت: ۳۰ خرداد ۱۴۰۲

تاریخ داوری: ۰۸ مرداد ۱۴۰۲

تاریخ پذیرش: ۰۹ مهر ۱۴۰۲

کلمات کلیدی:

الگوریتم بدون نظارت

خوشه‌بندی

ریسک پکیج

شخص ثالث

* نویسنده مسئول:

ایمیل: khamesian@irc.ac.ir

تلفن: ۰۸۴۰۸۴۰۲۲۰۹۸۲۱+

ORCID: 0000-0001-6113-4246

DOI: 10.22056/ijir.2024.01.02

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

در بیمه‌های اتومبیل، این عامل‌ها معمولاً متغیرهای قابل مشاهده‌ای در مورد راننده، نوع وسیله نقلیه و نوع کاربری هستند (Xie, 2021).

مروری بر پیشینه پژوهش

اصلی‌ترین و اثرگذارترین عامل‌های ریسکی که بنا بر مطالعات و پژوهش‌های گوناگون در رشته شخص ثالث احصا می‌شوند و شرکت‌های بیمه برای پیش‌بینی و ارزیابی ریسک از آن‌ها استفاده می‌کنند، متغیرهایی نظیر سن راننده، جنسیت، سوابق تصادفات یا ادعای خسارت، تاریخ گواهینامه، نوع وسیله نقلیه، نوع کاربری و محل سکونت هستند (Doerpinghaus et al., 2008; McCart et al., 2009; Ayuso et al., 2019). این متغیرها با نرخ خسارات همبستگی دارند و در نتیجه می‌توانند برای پیش‌بینی خسارات آینده مفید باشند. البته باید دقت داشت، ریسک فاکتورهایی که شناسایی و در نظر گرفته می‌شوند نباید به راحتی قابل دستکاری کردن باشند. به طور مثال، اگر میزان مسافت پیموده شده از طریق کیلومترشمار نصب شده بر روی اتومبیل سنجیده شود و اگر این کیلومترشمار قابل دستکاری باشد، این عامل ریسک فاکتور مناسبی نیست (Desyllas and Sako, 2013). یک رویکرد معمول برای انتخاب صحیح عوامل خطر یا همان ریسک فاکتورها، مبتنی بر روش‌های آماری چندمتغیره مانند رگرسیون خطی یا GLM است (David, 2015; McCullag, 2019). اما یکی از مشکلات این روش‌ها در این است که هنوز مقدار زیادی ناهمگونی در کلاس‌های مختلف به جا می‌گذارند. منابع مختلفی در این رابطه وجود دارد که از آن جمله می‌توان (Arvidsson, 2010) را نام برد. در حالی که این مدل‌ها ساده و به راحتی قابل توضیح‌اند، اما غالباً برای یادگیری و انعکاس اثرات پیچیده بسیار محدودند. رشته‌های بیمه اموال و مسئولیت از جمله بیمه اتومبیل، خطراتی را پوشش می‌دهد که از ترکیب چند منبع (علت) از جمله علل رفتاری ناشی می‌شود و در نتیجه به ندرت یک رابطه خطی برای مدل‌سازی رفتارهای پیچیده کافی خواهد بود. به عبارتی، در این رشته‌ها تبدیل‌های غیرخطی و تعاملات بین متغیرها می‌توانند واقعیت را با دقت بیشتری منعکس کنند که برای گنجاندن این اثرات در GLM ها، بیم‌سنج باید این ویژگی‌ها را به صورت دستی ایجاد کند و در مدل بگنجانند. اما مدل‌های یادگیری ماشین این مشکل را رفع کرده‌اند و تبدیل‌های غیرخطی و روابط بین متغیرها را بدون تعیین دستی آن‌ها یاد می‌گیرند (Spedicato et al., 2018; Burka et al., 2021). این کار معمولاً با مدل‌هایی نظیر شبکه عصبی و مدل‌های مبتنی بر درخت به راحتی انجام می‌شود (Hanafy and Ming, 2021). استفاده از این مدل‌ها رویکردی نو و منعطف است و آن‌ها اغلب می‌توانند سطح بالایی از دقت پیش‌بینی را ارائه دهند. محققان زیادی با استفاده از این مدل‌ها به تعیین ریسک فاکتورهای انواع رشته‌های بیمه‌ای و طبقه‌بندی مشتریان براساس میزان ریسکشان و همچنین قیمت‌گذاری و تعیین نرخ برای بیمه‌نامه‌ها پرداخته‌اند (Dugas et al., 2003; Yeo, 2009). در مسائل قیمت‌گذاری، به دلیل پیچیدگی مشخصات مدل و اجرای

شرکت‌های بیمه مانند دیگر بنگاه‌های اقتصادی با انواع مختلفی از ریسک‌ها روبه‌رو هستند (Hoy, 1982). در بین ریسک‌هایی که شرکت‌های بیمه و بیمه‌گران با آن مواجه‌اند، ریسک صدور اهمیت ویژه‌ای دارد. در اغلب نظام‌های توانگری مالی، بیش از ۲۲ درصد ریسک‌های شرکت‌های بیمه را ریسک‌های صدور بیمه‌نامه تشکیل می‌دهند (Eling et al., 2007). آنچه در اغلب موارد ممکن است ریسک صدور را افزایش دهد و شرکت بیمه را دچار مشکل کند، ارائه بیمه‌نامه به مشتریان پریسک و همچنین متناسب نبودن نرخ بیمه‌نامه با ریسک مشتریان و در نتیجه عدم پوشش مناسب خسارت‌های ادعا شده به وسیله حق بیمه‌های دریافتی است. با توجه به اهمیت و تأثیر ریسک صدور یا بیمه‌گری بر عملکرد شرکت‌های بیمه، این موارد ممکن است حتی به ورشکستگی یک شرکت بیمه منجر شود. کاهش و کنترل ریسک صدور به عنوان یکی از عوامل مهم و مؤثر بر بهبود فرایند بیمه‌گری و در نتیجه عملکرد شرکت‌های بیمه مطرح است و نقش اساسی در تداوم انجام این فرایند و بقای شرکت‌های بیمه دارد. به عبارت دیگر، می‌توان گفت اگر یک شرکت بیمه در شناسایی و ارزیابی میزان ریسک مشتریان و ارائه پوشش بیمه‌ای متناسب با ریسکشان به آن‌ها دچار مشکل شود، زیان هنگفتی را متوجه خود خواهد کرد. حال آنکه در صورت وجود یک نظام کارا و هوشمند برای شناسایی و ارزیابی سریع میزان ریسک مشتریان، شرکت بیمه می‌تواند تا حد زیادی این مشکل را کاهش دهد و محصولات و پوشش‌های بیمه‌ای خود را به صورت کارآمد به مشتریان تخصیص دهد. شرکت‌های بیمه در همه رشته‌ها از جمله بیمه اتومبیل، سعی می‌کنند بیمه‌نامه‌های خود را در طبقات تعرفه‌ای همگن طبقه‌بندی کنند و به بیمه‌گذاران و بیمه‌نامه‌هایی که متعلق به یک طبقه ریسک هستند حق بیمه یکسانی را اختصاص دهند تا به این ترتیب از بیمه‌گذاران حق بیمه عادلانه دریافت کنند. بنابراین انتخاب مجموعه مناسبی از ریسک فاکتورها برای پیش‌بینی صحیح نرخ و همچنین میزان خسارت‌های بیمه‌گذاران، می‌تواند برای یک شرکت بیمه بسیار مهم باشد. هنگامی که از یک مدل ریاضی پیچیده به منظور نرخ‌گذاری و تعیین قیمت و یا کشف الگوی خسارات استفاده می‌شود، مطالعه تأثیر یا انجام تحلیل حساسیت ریسک فاکتورها در مدل بسیار مهم است (Asmussen and Rubinstein, 1999). در شرایطی که نرخ‌های بیمه‌نامه‌ها را قانون‌گذار تعیین می‌کند (مانند بیمه شخص ثالث در کشور ایران) و شرکت‌های بیمه ناگزیر به پیروی از این نرخ‌ها هستند، باز هم تعیین دقیق ریسک فاکتورها و پیش‌بینی سطح خسارت بیمه‌گذاران می‌تواند مفید باشد، زیرا تعیین دقیق عامل‌های ریسک به جلوگیری از انتخاب نامطلوب بیمه‌نامه‌ها کمک می‌کند (Dionne et al., 1999). این عامل‌ها ویژگی‌ها و متغیرهایی هستند که همچنین به شرکت‌های بیمه کمک می‌کنند تا مبلغ خسارات خود را در دوره زمانی معینی (که معمولاً یک ساله است)، پیش‌بینی کنند. به عبارت دیگر، شرکت‌های بیمه میزان خسارات احتمالی خود را براساس این عامل‌ها مدل‌سازی می‌کنند.

اندازه دقیق تأثیرگذاری ریسک فاکتورها که چه بسا در شرایط مختلف متغیر نیز هست، محاسبه نمی‌شود. از طرف دیگر، این اقدام برای حالتی که ریسک فاکتوری در یک شرایط، عامل اصلی پریسک و در شرایطی دیگر، عامل اصلی کم‌ریسک بودن است، توضیحی ندارد. به‌طور نمونه در مشاهدات پرونده بیمه‌ها مشخص شد در رشته بیمه‌های شخص ثالث و بدنه (اتومبیل) بسیاری از افراد با سن بالا پریسک‌اند و در خوشه دیگر افراد با همین میانگین سن، کمترین ریسک را دارند. در نتیجه نمی‌توان در خصوص افزایش سن به‌عنوان یک ریسک فاکتور با قاطعیت نظر داد.

با توجه به آنچه گفته شد و سوابق مطالعاتی، در این پژوهش روشی جدید برای بررسی ریسک فاکتورها ارائه می‌دهیم تا از طریق آن مبنای دقیق‌تری از ریسک فاکتور به دست آید و ایراد مطرح‌شده در روش‌های قبلی را کمی مرتفع سازد. در روش جدید که معرفی می‌شود، ریسک فاکتورها با توجه به سطوح مقادیرشان به متغیرهای کوچک‌تر و جدا از هم تقسیم می‌شوند. سپس هر کدام از سطوح مختلف ریسک فاکتورهای اصلی که هم‌خوشه هستند، ریسک فاکتور جدیدی را تشکیل می‌دهند و به‌این ترتیب پکیج‌های مختلفی ایجاد می‌شود که هر کدام از آن‌ها یک عامل ریسک یا همان ریسک فاکتور هستند. به‌عبارتی دیگر، بازه‌های مقادیر ریسک فاکتورهای سابق در ترکیب با یکدیگر ریسک فاکتور جدیدی را می‌سازند که باید دوباره ارزیابی شود. ملاک و عملیات جداسازی بازه‌های مقادیر ریسک فاکتورهای اولیه و چگونگی ترکیب آن با بازه مقادیر همدیگر از طریق مدل خوشه‌بندی انجام‌پذیر است. در رابطه با مدل مورد استفاده برای خوشه‌بندی بیمه‌گذاران بیمه شخص ثالث، هدف این پژوهش نوآوری در ارائه مدل یادگیری ماشین نیست، بلکه ارائه دقیق‌ترین برآورد مورد نظر بوده است. در حقیقت با توجه به توانایی خوشه‌بندی، ابزار یادگیری ماشین کاربردی بسیار مهم در طبقه‌بندی ریسک دارند و از همین جهت در مقاله استفاده شده‌اند، زیرا چنانچه به‌جای اینکه همانند ادبیات پژوهش به این نتیجه برسیم که سن یک ریسک فاکتور است، با کاربرد یادگیری ماشین به این نتیجه برسیم که بازه‌های از تغییرات سن به‌عنوان عامل افزایش ریسک است و بازه‌های دیگر با توجه به بازه‌های از عامل ریسک دیگر مانند قیمت خودرو عامل کاهش ریسک است، می‌توانیم مفهوم جدیدی را معرفی کنیم که در ادبیات ریسک فاکتورها وجود نداشته است و کاربرد دقیق‌تری از آن است. شایان ذکر است نرم‌افزار مورد استفاده پایتون است. در اکثر مدل‌های یادگیری ماشین در زمان تخمین در پایتون، داده‌ها به دو مجموعه آموزشی (Training Set) و آزمایشی (Testing Set) تقسیم می‌شود و آزمون صحت مدل و برآوردها از طریق راستی‌آزمایی پیش‌بینی‌های مجموعه آزمایشی انجام می‌شود.

سؤالات پژوهش

با توجه به توضیحات ارائه‌شده در بخش قبل، سؤالات پژوهش به شرح ذیل قابل بررسی‌اند:

الف) چه مقادیر و چه ترکیبی از متغیرهای مورد بررسی،

آن، برای برآورده کردن شفافیت قیمت ارائه‌شده، توضیح مدل از طریق ارزیابی اهمیت متغیرهای مورد استفاده ضروری است. به‌عبارت‌دیگر، همان‌گونه که گفته شد در مسائل قیمت‌گذاری نیز قدم اول شناسایی ریسک فاکتورهای اصلی و بررسی تأثیر آن‌ها بر متغیرهای پاسخی نظیر فراوانی یا شدت خسارت است. (Xie (2021) در پژوهش خود با استفاده از روش شبکه عصبی مصنوعی به تجزیه و تحلیل ریسک فاکتورهای اصلی در بیمه‌های اتومبیل پرداخت و اهمیت این متغیرها را با یکدیگر مقایسه کرد.

(Cheong et al., 2008) با استفاده از مدل‌های GLM به قیمت‌گذاری بیمه‌های اتومبیل در مالزی پرداختند. آن‌ها ابتدا اصلی‌ترین متغیرهای تأثیرگذار بر فراوانی و شدت خسارات را شناسایی و پس از مدل‌سازی فراوانی و شدت خسارات به ارائه حق بیمه خالص و ناخالص برای این رشته بیمه‌ای پرداختند. (Segovia-Vargas et al., 2015) نیز با استفاده از روش‌های هوش مصنوعی و نظریه راف به تعیین ریسک فاکتورها در بیمه اتومبیل پرداختند. به گفته آن‌ها شواهد تجربی نشان می‌دهد که ریسک فاکتورهایی از قبیل سن راننده، جنسیت و ... که به‌صورت معمول از سوی شرکت‌های بیمه در این رشته در نظر گرفته می‌شود، متغیرهای توضیحی خوبی برای طبقه‌بندی مشتریان این رشته هستند و علاوه بر این متغیرهای معمول، پاداش جریمه کمی قدرت توزیع را افزایش می‌دهد. (Dugas et al., 2003) ریسک فاکتورهای اصلی مورد استفاده در نرخ‌گذاری بیمه‌های اتومبیل در منطقه آمریکای شمالی را بررسی کردند. آن‌ها قدرت تمایز و اثرگذاری هر یک از این متغیرها را بر روی نرخ بیمه‌نامه تحلیل کردند و عملکرد چندین مدل را در پنج دسته کلی رگرسیون خطی، مدل‌های خطی تعمیم‌یافته، درخت‌های تصمیم، شبکه‌های مصنوعی و ماشین‌های بردار پشتیبان بررسی کردند. مقایسه این مدل‌ها به‌صورت کیفی انجام شد و آن‌ها در نهایت نشان دادند که چگونه شبکه‌های عصبی می‌توانند وابستگی‌های غیرخطی مرتبه بالا را با تعداد کمی از پارامترها که هر یک بر روی نسبت بزرگی از داده‌ها تخمین زده می‌شوند، نشان دهند.

حال مسئله اینجاست که با بررسی مشاهدات ریسک فاکتورها برای هر نوع بیمه‌نامه‌ای متوجه نوعی تناقض رفتاری ریسک فاکتورها می‌شویم. ممکن است یک عامل ریسک در بسیاری از شرایط عامل اصلی پریسک بودن بیمه‌گذار یا مورد بیمه باشد، ولی در شرایط دیگر همین عامل ریسک اهمیت کمتری در پریسک بودن بیمه‌گذار دارد و حال چه بسا مشاهدات ما از انبوه داده‌ها نشان داده است بسیاری از متغیرها یا ریسک فاکتورها در یک وضعیت عامل اصلی پریسک بودن مشتری و همین ریسک فاکتور در وضعیتی دیگر عامل کم‌ریسک بودن مشتری را نشان داده است. در سوابق مطالعاتی و پژوهشی عمدتاً دلیل این امر را در همبستگی ریسک‌ها و ریسک فاکتورها می‌بینند و با حذف اثرات همبستگی درصد تحلیل ریسک فاکتورها برمی‌آیند (Barsotti et al., 2016; Meyers et al., 2003). مشکل این دیدگاه در این است که با لحاظ همبستگی بین ریسک فاکتورها، اثرات متغیرها به‌طور میانگین محاسبه می‌شوند و

پکیج‌های مختلف ریسک (از کم‌ریسک به پرریسک) را تشکیل می‌دهند؟

ب) آیا ریسک پکیج به‌عنوان روش اندازه‌گیری جدید در ارزیابی ریسک، نتایج دقیق‌تری نسبت به ریسک فاکتور ارائه می‌دهد؟ در ادامه مقاله، روش پژوهش استفاده‌شده برای بررسی و شناسایی پکیج‌های مختلف ریسک معرفی می‌شود. پس از تحلیل هر ریسک پکیج که به‌صورت خوشه‌های گوناگون معرفی شده‌اند، پاسخ پرسش‌ها و برتری این رویکرد نسبت به روش کلاسیک که تأثیر ریسک فاکتورها را مدنظر قرار می‌دهد نیز مشخص می‌شود.

مبانی نظری و روش‌شناسی پژوهش

خوشه‌بندی فرایند گروه‌بندی مجموعه‌ای از داده‌ها با تقسیم داده‌ها به گروه‌ها یا خوشه‌ها با هدف به حداکثر رساندن شباهت در داخل هر گروه و به حداقل رساندن شباهت بین گروهی بوده، به‌طوری که اعضای یک خوشه شباهت زیادی داشته، ولی اعضای خوشه‌های مختلف متفاوت باشند. فرایند خوشه‌بندی اساساً شامل سه مرحله اصلی است: (۱) تعریف معیار تشابه (Similarity)؛ (۲) تعیین معیاری برای فرایند ساخت خوشه؛ (۳) الگوریتم مناسب برای ساخت خوشه‌ها براساس معیار انتخاب‌شده. بنابراین اولین اقدام، در نظر گرفتن معیاری مناسب برای ارزیابی «فاصله» (عدم تشابه (Dissimilarity)) بین اشیاء است (Han et al., 2012; Likas et al., 2003).

در اصل، معیار تشابه تابع $d: D \times D \rightarrow \mathbb{R}_+$ است که بر روی مجموعه‌ای از اشیاء D اعمال می‌شود و ویژگی‌های خاصی دارد. از نظر مفهومی، می‌توان گفت که معیار تشابه معکوس فاصله است. بنابراین، اصطلاح معیار عدم تشابه، به‌عنوان فاصله بین دو شیء نیز در نظر گرفته می‌شود. بنابراین به‌عنوان یک تابع، فاصله معیاری است که ویژگی‌های اساسی یک متریک را دارد (Chen et al., 2009).

۱- غیرمنفی بودن و اصل این‌همانی تمایزناپذیرها (Identity of Indiscernibles)

$$d(A, B) \geq 0, \quad d(A, B) = 0 \leftrightarrow A = B$$

۲- تقارن

$$d(A, B) = d(B, A)$$

۳- نامساوی مثلثی

$$d(A, B) + d(B, C) \geq d(A, C)$$

در ادامه تعدادی از رایج‌ترین توابع فاصله ارائه می‌شود. برای اندازه‌گیری تشابه بین دو شیء l نمونه، آن‌ها را به‌عنوان بردار در نظر می‌گیریم که به جهت سادگی فرض می‌کنیم که هم‌بعد باشند: $X = (x_1, x_2, \dots, x_n)$ و $Y = (y_1, y_2, \dots, y_n)$. شایان ذکر است که در برخی موارد، ممکن است (با برخی اصلاحات)

بردارهایی با ابعاد مختلف در نظر گرفته شوند.

با در نظر گرفتن بردارهای بالا، فاصله اقلیدسی را می‌توان به‌سادگی به‌عنوان کوتاه‌ترین فاصله بین ۲ نقطه بدون توجه به ابعاد تعریف کرد. این شیوه رایج‌ترین راه برای یافتن فاصله است. براساس فرمول فاصله اقلیدسی، فاصله بین دو نقطه عبارت است از:

$$d_{Euc}(X, Y) = \left(\sum_{i=1}^n |x_i - y_i|^2 \right)^{1/2}$$

فاصله اقلیدسی را می‌توان با استفاده از نرم مینکوفسکی (Minkowski) که به نرم p نیز معروف است تعمیم داد، که در این صورت برای دو نقطه X, Y خواهیم داشت:

$$d_p(X, Y) = \left(\sum_{i=1}^n |x_i - y_i|^p \right)^{1/p}, \quad p \in \mathbb{N}.$$

همان‌طور که مشاهده می‌شود، برای حالت $p = 2$ فاصله مینکوفسکی به تابع فاصله اقلیدسی تبدیل می‌شود. معمولاً مقادیر ۱ و ۲ برای متغیر p استفاده می‌شود. اگر مقدار p را برابر یک قرار دهیم، تابع فاصله منهتن (Manhattan) به‌دست خواهد آمد:

$$d_{cb}(X, Y) = \left(\sum_{i=1}^n |x_i - y_i| \right)$$

در حالت‌های حدی نیز فاصله چیبیشف را خواهیم داشت:

$$\lim_{p \rightarrow \infty} \left(\sum_{i=1}^n |x_i - y_i|^p \right)^{1/p} = \max \{ |x_i - y_i| \}$$

$$\lim_{p \rightarrow -\infty} \left(\sum_{i=1}^n |x_i - y_i|^p \right)^{1/p} = \min \{ |x_i - y_i| \}$$

الگوریتم کا-میانگین (K-means) یکی از شناخته‌شده‌ترین الگوریتم‌های پرکاربرد خوشه‌بندی‌های غیرسلسله‌مراتبی (Non-hierarchical) در حوزه یادگیری بدون نظارت است. از نظر تاریخی، همچنان کا-میانگین، در میان سایر الگوریتم‌های گروه‌بندی، مناسب‌ترین رویکرد گروه‌بندی داده است. الگوریتم کا-میانگین دارای قابلیت گروه‌بندی تعداد انبوهی از داده‌ها با زمان محاسبات نسبتاً سریع و کارآمد است. از نظر فنی، مراحل الگوریتم به شرح زیر است (MacKay, 2003).

- ۱- فرض کنید N نقطه داده به شکل $X^l = (x_1^l, x_2^l, \dots, x_n^l)$ وجود دارد که $l = 1, 2, \dots, N$.
- ۲- مجموعه‌ای از k بردار نماینده (مراکز خوشه) C_j را پیدا کنید ($j = 1, 2, \dots, k$).
- ۳- نقاط داده را به k زیرمجموعه مجزا S_j حاوی N_j نقطه تقسیم کنید، به‌گونه‌ای که تابع زیر به حداقل برسد:

خوشه‌ها تا مرکز متناظر با آن را ارائه می‌دهد.

$$WCSS = \sum_{C_k}^{C_n} \left(\sum_{p_i \in C_k}^{p_m} dist(p_i, C_k)^2 \right)$$

در رابطه بالا C مراکز خوشه و i نقاط موجود در خوشه است و هدف به حداقل رساندن مجموع بالاست. فرض کنید n مشاهده در یک مجموعه داده وجود دارد و ما n تعداد خوشه را مشخص کنیم ($K = n$)، سپس WCSS صفر می‌شود، زیرا نقاط داده خودشان به‌عنوان مرکز عمل می‌کنند و فاصله نقاط تا مرکز صفر خواهد بود که در این حالت به n خوشه خواهیم رسید که طبیعتاً تعدادی بدیهی و غیرمنطقی است. بنابراین یک مقدار آستانه برای K وجود دارد که می‌توانیم آن را با استفاده از نمودار آرنج پیدا کنیم. برای این منظور به‌طور تصادفی الگوریتم ک- میانگین را برای محدوده‌ای از مقادیر K مقداردهی اولیه می‌کنیم و مقدار WCSS را برای هر K رسم می‌کنیم.

برای شکل ۲، مقدار بهینه برای K عدد ۵ خواهد بود. همان‌طور که می‌بینیم با افزایش تعداد خوشه‌ها، مقدار WCSS کاهش می‌یابد. مقدار K براساس میزان کاهش WCSS انتخاب می‌شود. به‌طور مثال، از خوشه ۱ به ۲ و ۳ در نمودار بالا شاهد افت ناگهانی و شدید WCSS هستیم. اما از مقدار ۵ به بعد شاهد افت حداقلی هستیم و از این‌رو ۵ به‌عنوان مقدار بهینه برای K انتخاب می‌شود. برای مطالعه فاصله جدایی بین خوشه‌های حاصل می‌توان از تحلیل سایه‌نما یا سیلوئت نیز استفاده کرد. نمودار سیلوئت میزان نزدیکی هر نقطه در یک خوشه را نسبت به نقاط موجود در خوشه‌های همسایه نشان می‌دهد و بنابراین راهی برای ارزیابی بصری تعداد خوشه‌ها ارائه می‌دهد. ضرایب سیلوئت از رابطه زیر محاسبه می‌شود:

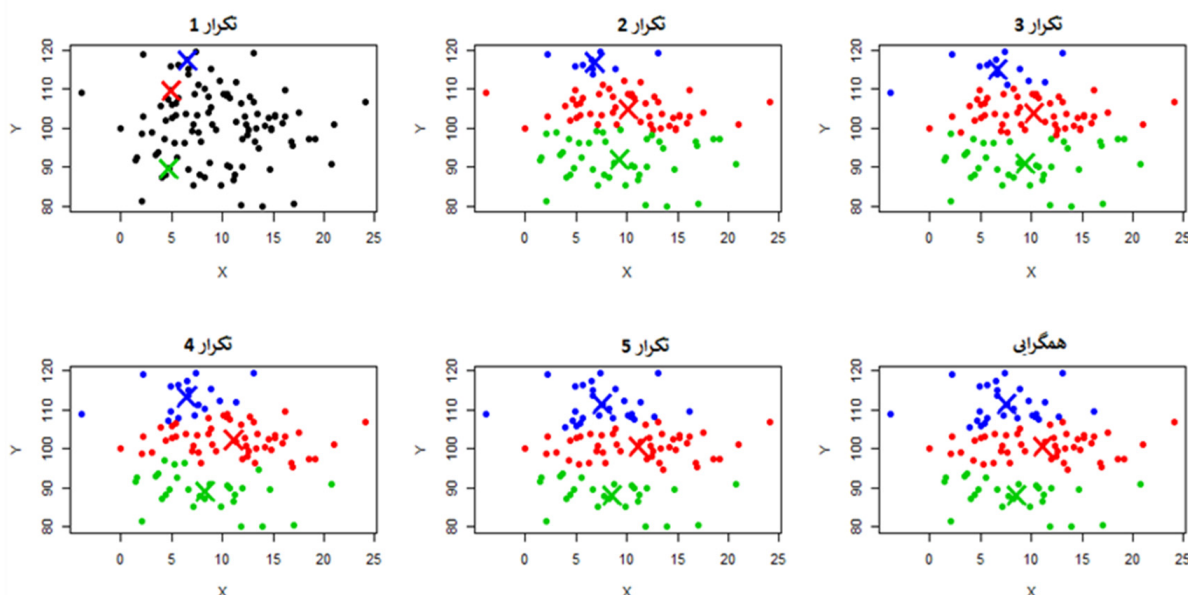
$$J = \sum_{j=1}^k \sum_{l \in S_j} X^l - c_j^2$$

که در آن c_j میانگین نقاط مجموعه S_j است که از رابطه زیر به‌دست می‌آید:

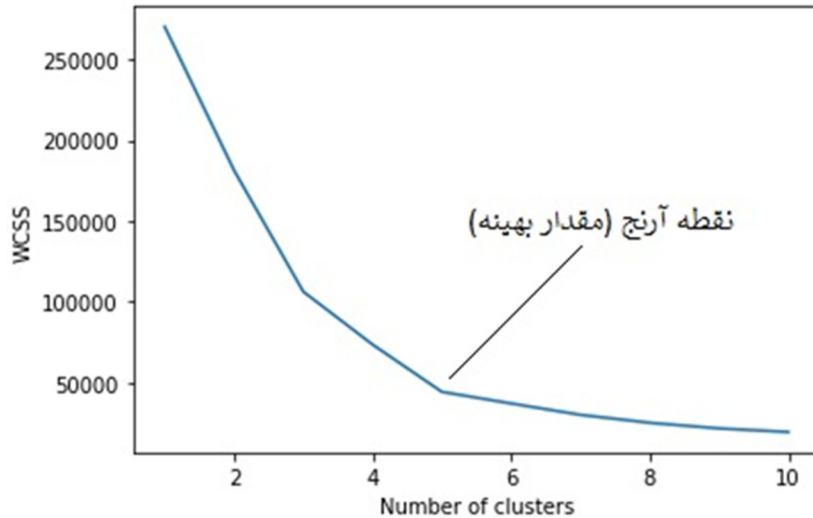
$$c_j = \frac{\sum_{l \in S_j} X^l}{N_j}$$

به‌عبارتی دیگر، به‌طور مثال در حالتی که با فاصله اقلیدسی کار می‌کنیم مقداردهی اولیه مراکز خوشه و سایر محاسبات را آن‌قدر تکرار می‌کنیم که خوشه تازه تشکیل شده در نقاط مرکزی تغییر نکند. شکل ۱ مراحل رسیدن از داده خام به خوشه‌های نهایی را در پنج تکرار نمایش می‌دهد.

همان‌طور که مشاهده کردید، براساس ساختار الگوریتم ک- میانگین تعداد خوشه‌های مورد نظر می‌بایست از پیش تعیین شود. این موضوع چالش اصلی استفاده از الگوریتم مزبور است. برای حل مشکل تعیین بهینه تعداد خوشه‌ها دو رویکرد اصلی سایه‌نما (Silhouette score) و روش آرنج (Elbow method) ارائه شده است (Singh et al., 2013; Pham et al., 2005; Llet et al., 2004; Kodinariya and Makwana, 2013) برای مطالعه فاصله جدایی بین خوشه‌های حاصل می‌توان از روش آرنج استفاده کرد. در واقع هر خوشه با محاسبه و مقایسه فواصل نقاط داده در یک خوشه با مرکز آن تشکیل می‌شود. در همین راستا روشی ایده‌آل برای تعیین تعداد مناسب خوشه‌ها، محاسبه مجموع مربعات درون خوشه‌ها ((Within-Cluster-Sum-of-Squares (WCSS)) است. مقدار این متغیر مجموع مربعات فواصل هر نقطه داده در همه



شکل ۱: خوشه‌بندی ک- میانگین ($K=3$)
Fig. 1. K-means clustering ($K=3$)



شکل ۲: نمودار WCSS در مقابل تعداد خوشه
Fig. 2. Plot of WCSS versus number of clusters

گرفت و شکل ۳ به عنوان خروجی مدل به دست آمد: براساس رویکرد آرنج و با توجه به نمودار بالا، تعداد خوشه بهینه که به بیشترین کاهش در مقدار WCSS منجر می شود مقدار آستانه ۴ است. شایان ذکر است مقدار ۳ نیز می تواند به عنوان یک کاندید برای تعداد بهینه خوشه در نظر گرفته شود. به همین دلیل از نمودارهای سیلوئت برای تصمیم گیری مناسب تر نیز استفاده می شود. در شکل زیر نمودار سیلوئت برای ۳ تا ۶ خوشه بندی ترسیم شده است

همان طور که در شکل ۴ نیز مشاهده می شود بین ضرایب سیلوئت برای تعداد ۳ و ۴ خوشه تفاوت چشمگیری وجود ندارد، اما این ضرایب برای حالتی که مجموعه داده به ۵ یا ۶ خوشه تقسیم می شود دچار کاهش زیادی می گردد. براساس ضرایب سیلوئت (که بیانگر تعداد بهینه خوشه های متمایز است) مقادیر ۳ و ۴ کاندیدهای اصلی برای تعداد بهینه خوشه است. نتایج نشان دهنده ۴ خوشه است، اما همان طور که در نمودارها مشاهده می شود، خوشه دوم و سوم طیفی از طبقه ریسکی متوسط اند و برای ارائه مدلی منطبق بر واقعیت و کاربردی تر برای شرکت های بیمه، بهتر است خوشه دوم و سوم را در یک طبقه ریسک متوسط قرار دهیم. تا در نهایت سه طبقه ریسکی کم ریسک، ریسک متوسط و پرریسک را از هم متمایز سازیم. به همین دلیل با وجود اینکه در خروجی نرم افزار ۴ خوشه داریم، در نهایت سه طبقه ریسک به عنوان نتیجه برآورد و تعداد بهینه خوشه بندی انتخاب می شود.

پس از پیدا کردن تعداد بهینه خوشه های مربوط به بیمه گذاران، برای ترسیم آن از تحلیل مؤلفه های اصلی استفاده شده و شکل ۵ به عنوان خروجی به دست آمده است:

پس از تعیین تعداد بهینه خوشه ها، توزیع متغیرهای مختلف در هر خوشه بررسی شد. این امر به تحلیل رفتاری بیمه گذاران و توصیف گروه بندی های مختلف کمک می کند. در همین راستا نمودارهای زیر

$$s(o) = \frac{b(o) - a(o)}{\max\{a(o), b(o)\}}$$

که در آن:

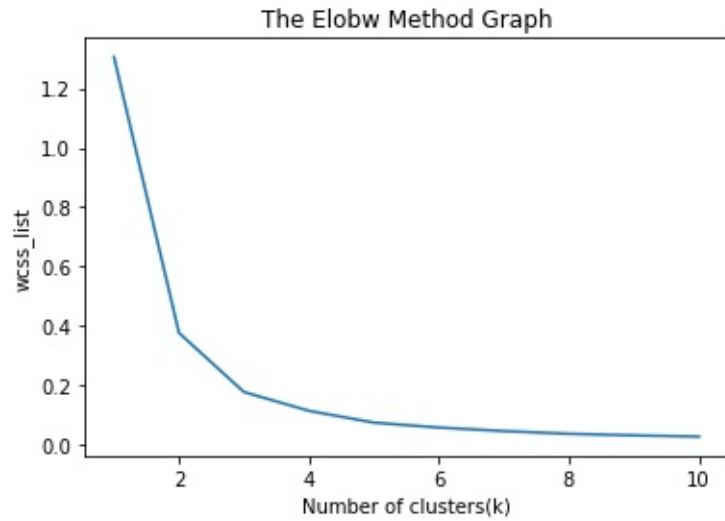
- $s(o)$ ضریب سیلوئت متناظر با نقطه O ،
- $a(o)$ میانگین فاصله بین O و سایر نقاط داده در خوشه ای که O به آن تعلق دارد و
- $b(o)$ حداقل میانگین فاصله از O تا خوشه هایی هستند که O به آنها تعلق ندارد.

ضرایب در بازه $[-1, 1]$ تعریف می شوند. مقادیر نزدیک به ۱ نشان می دهد که نمونه داده مورد نظر از خوشه های همسایه دور است. مقدار صفر بیانگر آن است که نمونه بسیار نزدیک و یا روی مرز تصمیم بین دو خوشه همسایه قرار دارد و مقادیر منفی نشان می دهد که آن نمونه ها ممکن است به خوشه اشتباهی اختصاص داده شده باشند.

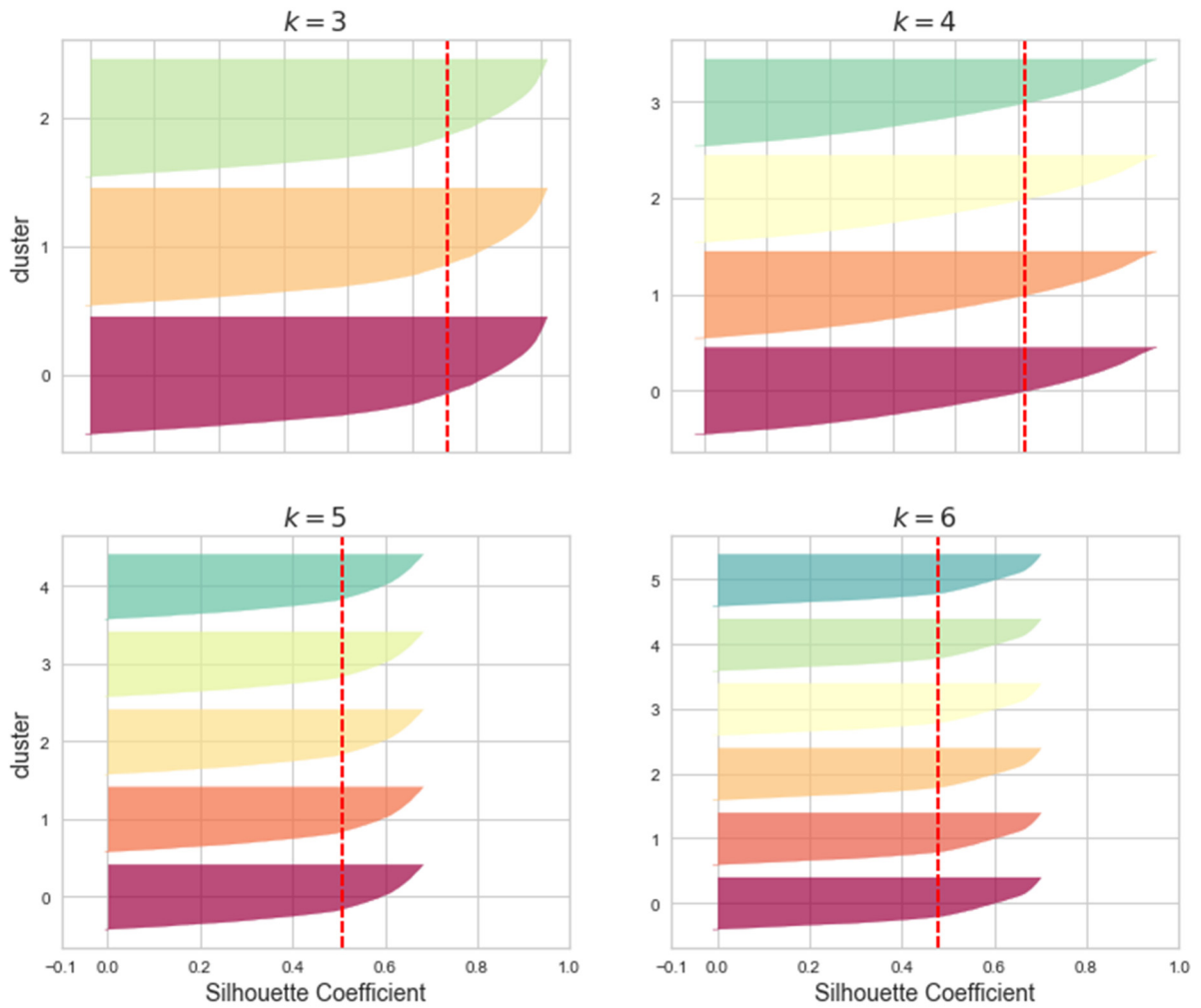
نتایج و بحث

در این مقاله، مطالعه بر روی داده های خسارت مالی مربوط به پرونده های خسارت بیمه شخص ثالث صورت گرفته است. مجموعه داده مزبور شامل ۲۰۶۰۶ نمونه و چهار متغیر (ویژگی) با عناوین سن بیمه گذار، سطح درآمد بیمه گذار (براساس میزان حق بیمه پرداختی)، جنسیت بیمه گذار و خسارت مالی پرداختی به بیمه گذار است. قبل از اعمال مدل بر روی داده ها، مراحل پیش پردازش داده شامل کشف و جایگزینی مقادیر گم شده، کدگذاری متغیرهای اسمی و نرمال سازی (تبدیل دامنه تغییرات مشاهدات به بازه صفر و یک) مقادیر متغیرها صورت گرفته است. تمامی تحلیل ها با استفاده از نرم افزار پایتون انجام شده است.

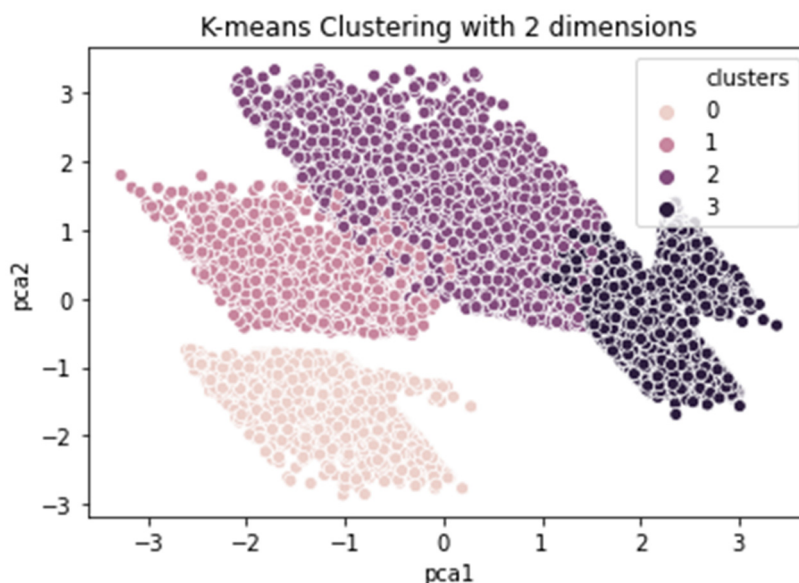
پس از آماده سازی داده ها، تعداد خوشه های متفاوت برای دسته بندی بیمه گذاران به گروه های مناسب بر روی داده ها انجام



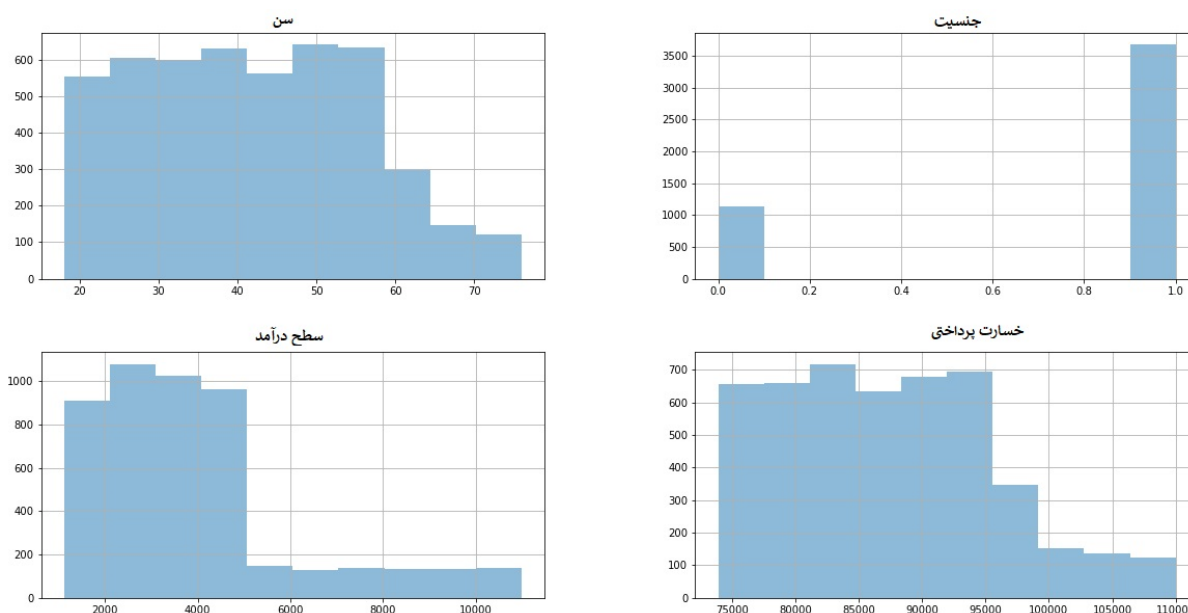
شکل ۳: نمودار آرنج برای ۱۰ خوشه‌بندی روی داده
Fig. 3. Elbow diagram for 10 event clustering



شکل ۴: نمودار سیلوئت برای تعداد ۳، ۴، ۵ و ۶ خوشه
Fig. 4. Silhouette diagram for the number of 3, 4, 5 and 6 clusters



شکل ۵: خوشه‌بندی بهینه از طریق تحلیل مؤلفه‌های اصلی
 Fig. 5. Optimal clustering through principal component analysis



شکل ۶: توزیع متغیرها در خوشه اول
 Fig. 6. Distribution of variables in the first cluster

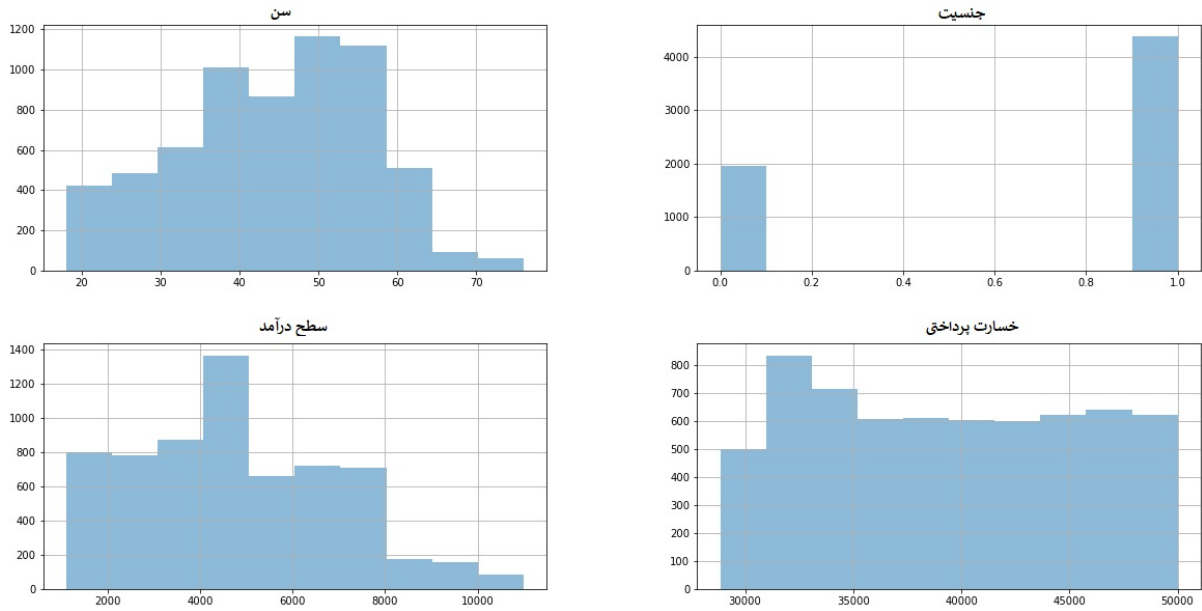
شکل ۸ بیانگر شیوه توزیع متغیرهای مجموعه داده‌های در اختیار در خوشه سوم است.

در نهایت شکل ۹ نمایانگر نمودار توزیع چهار متغیر سن، جنسیت، سطح درآمد و خسارت پرداختی در خوشه و گروه چهارم بیمه‌گذاران است. در همین راستا در بخش بعد با توجه به خروجی‌های به‌دست‌آمده از اعمال مدل و خوشه‌بندی بیمه‌گذاران به تحلیل هر خوشه جداگانه پرداخته می‌شود.

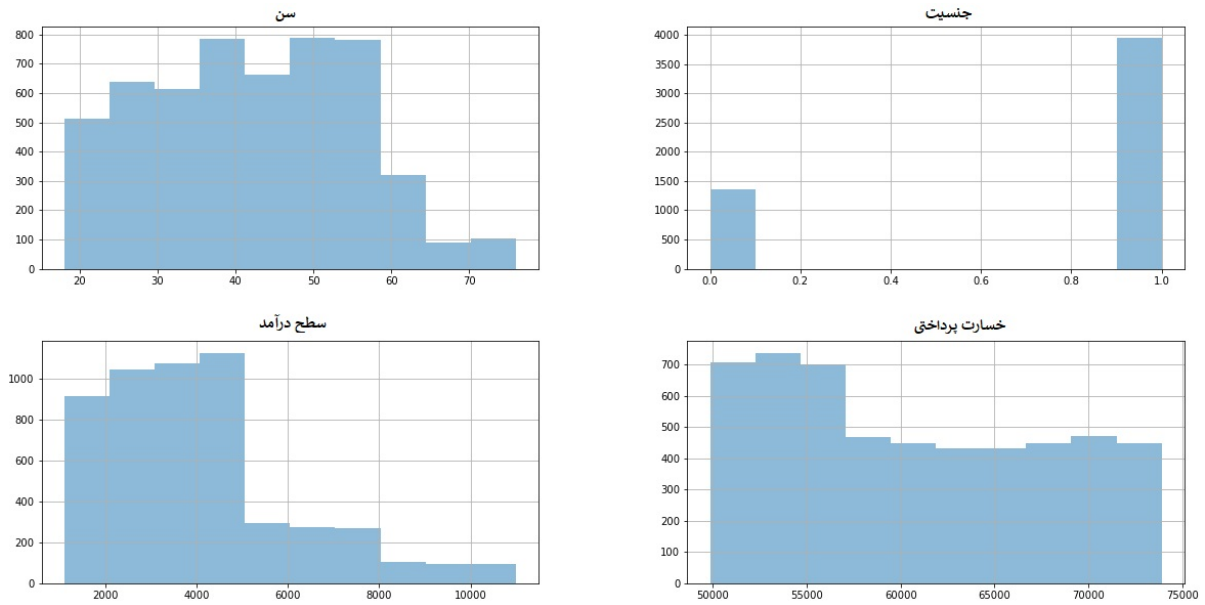
به ترتیب شرایط موجود در خوشه‌های ۱ تا ۴ را برای متغیرهای مختلف نشان می‌دهد.

در شکل ۶ تعداد و توزیع چهار متغیر اصلی مورد بررسی در مطالعه در خوشه اول مشاهده می‌شود.

چهار نمودار ترسیم‌شده در شکل ۷، میزان توزیع متغیرهای سن، جنسیت، سطح درآمد و خسارت پرداختی را در خوشه دوم نمایش می‌دهد.



شکل ۷: توزیع متغیرها در خوشه دوم
Fig. 7. Distribution of variables in the second cluster



شکل ۸: توزیع متغیرها در خوشه سوم
Fig. 8. Distribution of variables in the third cluster

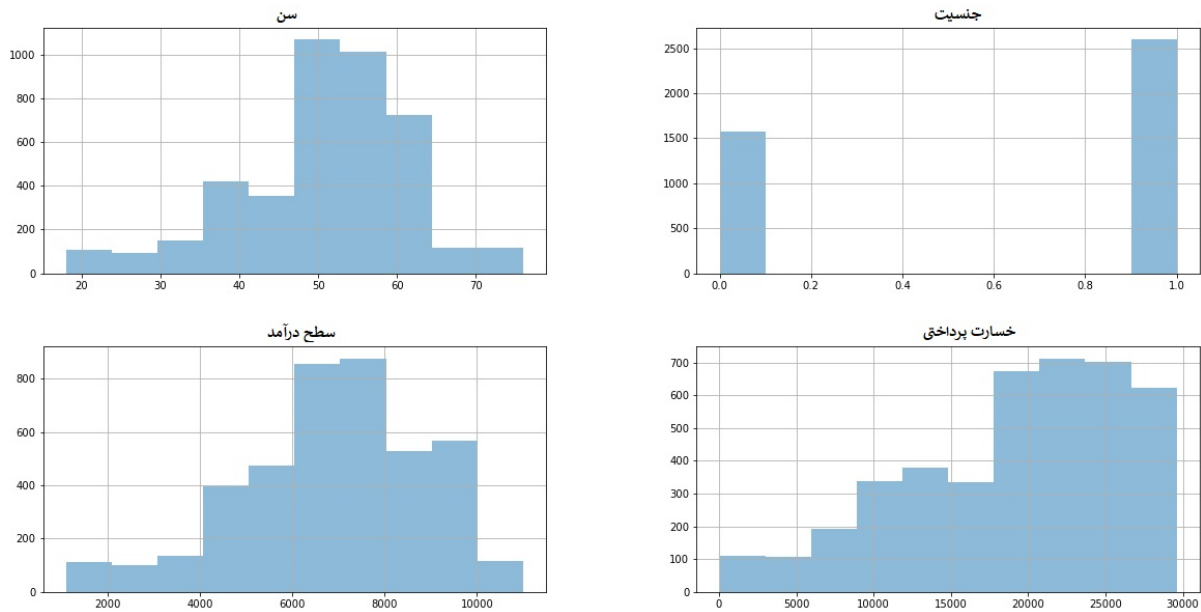
خروجی دسته‌بندی بیمه‌گذاران است. در خوشه اول همان‌طور که در نمودار مشخص است، پهنه سنی بیمه‌گذاران از کم‌سن تا سنین بالا مشاهده می‌شود، اما میانگین سنین ۴۱ سال با انحراف معیار ۱۴ است. مد در این خوشه ۵۰ سال است. البته توزیع داده‌ها مشابه توزیع یکنواخت و شامل همه گروه‌های سنی است. ویژگی دیگر این خوشه، نسبت ۰,۷ مردان به زنان است. از ویژگی‌های مهم دیگر این خوشه این است که معیار سطح درآمد آن در پایین‌ترین

جمع‌بندی و پیشنهادها

با توجه به تحلیل‌های انجام‌گرفته در بخش ۵، در پاسخ به پرسش اول مبنی بر مقادیر و چگونگی ترکیب متغیرها برای تشکیل ریسک‌پکیج‌های مختلف، خوشه‌های به‌دست‌آمده به شرح ذیل تفسیر می‌شوند.

خوشه اول

آنچه از نتایج برآوردها مشاهده می‌شود، ۴ خوشه بهترین



شکل ۹: توزیع متغیرها در خوشه چهارم
Fig. 9. Distribution of variables in the fourth cluster

بالاتر مردان در گروه‌های سطح درآمدی پایین است و ویژگی دیگری ارائه نمی‌دهد.

خوشه چهارم

این خوشه میانگین سنی متفاوت با سایر خوشه‌ها دارد. میانگین سنی ۵۰ و انحراف معیار ۱۰ نشان‌دهنده بیمه‌گذاران مسن در این خوشه است. همان‌طور که در نمودار نیز مشاهده می‌شود چولگی نیز به سمت راست است. از طرف دیگر معیار سطح درآمد این خوشه نیز بسیار بالاتر از دیگر خوشه‌ها و چوله به راست است. نسبت مردان به زنان در سطح ۰,۶ است که جزء نسبت پایین در خوشه‌هاست. این خوشه کمترین میزان ریسک (سطح ۱) را در مقابل سایر خوشه‌ها دارد. ویژگی این خوشه افراد با سنین بالا و ماشین‌های گران‌قیمت و سطح درآمدی بالاست.

از بررسی ویژگی‌های به‌دست‌آمده در ۴ خوشه مطرح‌شده می‌توان پکیج‌های ریسک ذیل را معرفی کرد. همان‌گونه که در بخش ۵ ذکر شد، چنانچه بخواهیم تمایز خوشه‌ها را که نشان‌دهنده طبقه ریسکی متفاوت است، شفاف‌تر کنیم و همپوشانی‌های خوشه‌ها را کمتر کنیم، بهتر است خوشه دوم و سوم را در یک طبقه ریسک متوسط قرار دهیم. - افراد با سنین بالا، متوسط و پایین (چگال در بازه ۳۰ تا ۵۸ سال) با ارزش ماشین ارزان‌قیمت با جنسیت مرد را می‌توان به‌عنوان بالاترین سطح ریسک معرفی کرد.

- افراد با سنین متوسط و بالا (چگال در بازه ۳۲ تا ۵۳ سال) با ارزش ماشین متوسط و بالا را می‌توان به‌عنوان ریسک‌های متوسط در نظر گرفت.

سطح میان ۴ خوشه ارائه شده است. درعین‌حال بازه ریسک این گروه بیشترین سطح ریسک را نشان می‌دهد. در تفسیر این خوشه می‌توان به گروه‌هایی از بیمه‌گذاران با درآمد کم و بیمه‌گذاران مرد که دارای خودروی ارزان‌قیمت‌تر از متوسط است، اشاره کرد. با توجه به پارامترهای بیان‌شده، شاخصه اصلی پکیج‌سازی از پریسک‌ترین بیمه‌گذاران مستخرج از خوشه اول، درآمد پایین، ماشین‌های ارزان‌قیمت و نسبت بالای مردان است.

خوشه دوم

در این خوشه، میانگین سنی کمی بیشتر از خوشه اول برابر ۴۴ سال و انحراف معیار ۱۲ است. البته با توجه به چولگی توزیع که در نمودار هم قابل مشاهده است، بیمه‌گذاران این خوشه با سن عموماً متوسط‌اند. معیارهای سطح درآمد نیز در میان ۴ خوشه در حد متوسط است. نسبت مردان به زنان نیز کمترین سطح درمیان خوشه‌هاست. سطح ریسک این گروه در حد متوسط رو به پایین یا طبقه ریسک ۲ قرار دارد. ویژگی‌های مهم این خوشه برای تشکیل پکیج درآمد متوسط با سنین متوسط است.

خوشه سوم

این خوشه با ویژگی‌های میانه‌ای از خوشه اول و دوم است. میانگین سنی ۴۲ سال با انحراف معیار ۱۴ و سطح درآمدی متوسط گروه‌های ۱ و ۲ را داراست. سطح ریسک این خوشه نیز در سطح متوسط خوشه اول و دوم است و در سطح ۳ قرار دارد. این خوشه تأکیدی بر تقابل سه ویژگی سن و سطح درآمد و وجود ریسک

مشارکت نویسندگان

مریم اثنی‌عشری: جمع‌آوری مطالعات مرتبط، جمع‌آوری و اخذ داده‌ها، تفسیر داده‌ها، نگارش و ویرایش. فرزانه خامسیان: ارائه مدل، تجزیه و تحلیل و تفسیر داده‌ها، تجزیه و تحلیل و تفسیر داده‌ها و نظارت و سرپرستی. فرید خانی‌زاده: اجرای مدل‌ها، تفسیر نتایج و خروجی‌ها.

تشکر و قدردانی

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

تعارض منافع

نویسندگان اعلام می‌دارند که در خصوص انتشار این مقاله تضاد منافع وجود ندارد. علاوه بر این، موضوعات اخلاقی شامل سرقت ادبی، رضایت آگاهانه، سوءرفتار، جعل داده‌ها، انتشار و ارسال مجدد و مکرر از سوی نویسندگان رعایت شده است.

دسترسی آزاد

کپی‌رایت نویسنده (ها) ©2024: این مقاله تحت مجوز بین‌المللی Creative Commons Attribution 4.0 اجازه استفاده، اشتراک‌گذاری، اقتباس، توزیع و تکثیر را در هر رسانه یا قالبی مشروط بر درج نحوه دقیق دسترسی به مجوز CC و منوط به ذکر تغییرات احتمالی در مقاله می‌داند. لذا به استناد مجوز مذکور، درج هرگونه تغییرات در تصاویر، منابع و ارجاعات یا سایر مطالب از اشخاص ثالث در این مقاله باید در این مجوز گنجانده شود، مگر اینکه در راستای اعتبار مقاله به اشکال دیگری مشخص شده باشد. در صورت عدم درج مطالب یادشده و یا استفاده فراتر از مجوز بالا، نویسنده ملزم به دریافت مجوز حق نسخه‌برداری از شخص ثالث است.

به‌منظور مشاهده مجوز بین‌المللی Creative Commons Attribution 4.0 به نشانی زیر مراجعه شود:
<http://creativecommons.org/licenses/by/4.0>

یادداشت ناشر

ناشر نشریه پژوهشنامه بیمه با توجه به مرزهای حقوقی در نقشه‌های منتشرشده بی‌طرف باقی می‌ماند.

- افراد با سنین متوسط به بالا (چگال در بازه ۵۱ تا ۶۳ سال) با ماشین گران‌قیمت را پایین‌ترین سطح ریسک در نظر گرفت. همان‌طور که در نتایج پکیج‌ها مشاهده می‌شود، با اینکه جمعیت قابل توجهی از بیمه‌گذارانی با سن بالا در پکیج اول (خوشه اول) هستند، بالاترین سطح ریسک را دارند، افرادی با سنین بالا در پکیج سوم (به‌طوری که حتی میانگین سنی این پکیج در بالاترین سطح در خوشه‌هاست) کمترین سطح ریسک را دارند. نکته درخور توجه دیگر اینکه با افزایش دهک درآمدی و هم‌زمان با افزایش سن (بررسی خوشه ۱ تا ۳) از سطح ریسک کاسته می‌شود. این موضوع بیانگر برتری رویکرد مزبور نسبت به روش کلاسیک ریسک‌فاکتور است که پاسخ به پرسش دوم این پژوهش نیز هست. در خصوص آزمون اینکه ریسک پکیج بهتر از ریسک‌فاکتور نشان‌دهنده ریسک افراد است، با توجه به اینکه در نمونه مثلاً افرادی با سنین بالا ولی خسارت کم و بالعکس در سنین پایین و خسارت بالا نیز به میزان چشمگیری (که در نمودارهای انتهای مقاله مشاهده می‌شود) وجود دارند، نشان‌دهنده این است که سن نماینده دقیقی برای نشان دادن ریسک بیمه‌گذاران نیست، ولی روش جدید با دسته‌بندی ریسک پکیج و ترکیب ریسک‌فاکتورها پکیج‌هایی ایجاد شده که توانسته در توضیح‌دهندگی ارتباط پکیج با خسارت‌ها دست‌های تفکیک‌پذیری ایجاد کند، به‌طوری که تفاوت میانگین و واریانس خسارت‌ها در کلاسترهای بیمه‌گذاران با ریسک‌های قابل تمیز از هم را توانسته توضیح دهد، بدون آنکه هم‌پوشانی حالت ریسک‌فاکتور را داشته باشد.

در نهایت شایان ذکر است که همچنان ترکیب‌هایی از متغیرهای استفاده‌شده در این تحقیق می‌توان تولید کرد که خوشه‌بندی‌های انجام‌شده در این مقاله نمی‌توانند توجیه و دسته‌بندی دقیقی برای آن ارائه دهند. در واقع در این مرحله مبحث رزولوشن به میان می‌آید. به این معنا که اگر بخواهیم تحلیل‌هایی با رزولوشن و دقت بالاتری ارائه دهیم که دربرگیرنده حالت‌های بیشتری باشد به تعداد متغیرهای مستقل بیشتری نیاز داریم. لیکن دسترسی به متغیرهای بیشتر که مقادیر تمیزی در آن‌ها وارد شده باشد امر ساده‌ای نیست. باوجود این موضوع مزبور توسط تیم نویسندگان این مقاله در دست انجام است و پس از بررسی‌ها و انجام مطالعات، خروجی پژوهش در قالب مقاله‌ای با نتایج مقاله فعلی مقایسه خواهد شد.

منابع

- Arvidsson, S., (2010). Does private information affect the insurance risk?: Evidence from the automobile insurance market. *Work. Pap.*, 1-54 **(54 Pages)**.
- Asmussen, S.; Rubinstein, R.Y., (1999). Sensitivity analysis of insurance risk models via simulation. *Manage. Sci.*, 45(8): 1125-1141 **(17 Pages)**.
- Ayuso, M.; Guillen, M.; Nielsen, J.P., (2019). Improving automobile insurance ratemaking using telematics: Incorporating mileage and driver behaviour data. *Transp.*, 46: 735-752 **(18 Pages)**.
- Barsotti, F.; Milhaud, X.; Salhi, Y., (2016). Lapse risk in life insurance: Correlation and contagion effects among policyholders' behaviors. *Insur. Math. Econ.*, 71: 317-331 **(15 Pages)**.
- Burka, D.; Kovács, L.; Szepesváry, L., (2021). Modelling MTPL insurance claim events: Can machine learning methods overperform the traditional GLM approach?. *Hung. Stat. Rev.*, 4(2): 34-69 **(36 Pages)**.
- Chen, S.; Ma, B.; Zhang, K., (2009). On the similarity metric and the distance metric. *Theor. Comput. Sci.*, 410(24-25): 2365-2376 **(12 Pages)**.
- Cheong, P.; Jemain, A.A.; ISMAIL, N., (2008). Practice and pricing in non-life insurance: The malaysian experience. *J. Qual. Meas. Anal. JQMA.*, 4(1): 11-24 **(14 Pages)**.
- David, M., (2015). Auto insurance premium calculation using generalized linear models. *Procedia. Econ. Finance.*, 20: 147-156 **(10 Pages)**.
- Desyllas, P.; Sako, M., (2013). Profiting from business model innovation: Evidence from Pay-As-You-Drive auto insurance. *Res. Policy.*, 42(1): 101-116 **(16 Pages)**.
- Dionne, G.; Gouriéroux, C.; Vanasse, C., (1999). Evidence of adverse selection in automobile insurance markets. In *Automobile Insurance: Road safety, new drivers, risks, insurance fraud and regulation.*, 20(1).
- Doerpinghaus, H.I.; Schmit, J.T.; Yeh, J.J.H., (2008). Age and gender effects on auto liability insurance payouts. *J. Risk. Insur.*, 75(3): 527-550 **(24 Pages)**.
- Dugas, C.; Bengio, Y.; Chapados, N.; Vincent, P.; Denoncourt, G.; Fournier, C., (2003). Statistical learning algorithms applied to automobile insurance ratemaking. *CAS. Forum.*, 1(1): 179-214 **(36 Pages)**.
- Eling, M.; Schmeiser, H.; Schmit, J.T., (2007). The Solvency II process: Overview and critical analysis. *Risk. Manage. Insur. Rev.*, 10(1): 69-85 **(17 Pages)**.
- Han, J.; Kamber, M.; Pei, J., (2012). *Data Mining: Concepts and techniques*. Elsevier.
- Hanafy, M.; Ming, R., (2021). Machine learning approaches for auto insurance big data. *Risk.*, 9(2): 1-42 **(42 Pages)**.
- Hoy, M., (1982). Categorizing risks in the insurance industry. *Q. J. Econ.*, 97(2): 321-336 **(16 Pages)**.
- Kodinariya, T.M.; Makwana, P.R., (2013). Review on determining number of cluster in K-Means clustering. *Int. J.*, 1(6): 90-95 **(6 Pages)**.
- Likas, A.; Vlassis, N.; Verbeek, J.J., (2003). The global k-means clustering algorithm. *Pattern. Recognit.*, 36(2): 451-461 **(11 Pages)**.
- Lleti, R.; Ortiz, M.C.; Sarabia, L.A.; Sánchez, M.S., (2004). Selecting variables for k-means cluster analysis by using a genetic algorithm that optimises the silhouettes. *Anal. Chim. Acta.*, 515(1): 87-100 **(14 Pages)**.
- MacKay, D., (2003). *Information theory, Inference, and learning algorithms*. Cambridge university press
- McCartt, A.T.; Mayhew, D.R.; Braitman, K.A.; Ferguson, S.A.; Simpson, H.M., (2009). Effects of age and experience on young driver crashes: Review of recent literature. *Traffic. Inj. Prev.*, 10(3): 209-219 **(11 Pages)**.
- McCullagh, P., (2019). *Generalized linear models*. Routledge.
- Meyers, G.G.; Klinker, F.L.; Lalonde, D.A., (2003). The Aggregation and Correlation of Reinsurance Exposure. *Casualty. Actuarial. Soc. Forum.*, 69-151 **(82 Pages)**.
- Pham, D.T.; Dimov, S.S.; Nguyen, C.D., (2005). Selection of K in K-means clustering. *Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science.*, 219(1): 103-119 **(17 Pages)**.
- Segovia-Vargas, M.J.; Camacho-Miñano, M.D.M.; Pascual-Ezama, D., (2015). Risk factor selection in automobile insurance policies: A way to improve the bottom line of insurance companies. *Rev. brasileira de gestão de negócios.*, 1228-1245 **(18 Pages)**.
- Singh, A.; Yadav, A.; Rana, A., (2013). K-means with Three different distance metrics. *Int. J. Comput. Appl.*, 67(10): 13-17 **(5 Pages)**.
- Spedicato, G.A.; Dutang, C.; Petrini, L., (2018). Machine learning methods to perform pricing optimization. A comparison with standard GLMs. *Variance.*, 12(1): 69-89 **(21 Pages)**.
- Xie, S., (2021). Improving explainability of major risk factors in artificial neural networks for auto insurance rate regulation. *Risk.*, 9(7).
- Yeo, A.C., (2009). *Neural networks for automobile insurance pricing*. *Encycl. Inf. Sci. Technol. Second. Ed.*, 2794-2799 **(6 Pages)**.

AUTHOR(S) BIOSKETCHES	معرفی نویسندگان
<p>مریم اثنی عشری، استادیار گروه پژوهشی بیمه‌های اموال و مسئولیت، پژوهشکده بیمه، تهران، ایران</p> <ul style="list-style-type: none">Email: esnaashari@irc.ac.irORCID: 0000-0001-5337-9665Homepage: https://www.irc.ac.ir/fa-IR/Irc/5020/page <p>فرزان خامسیان، استادیار گروه پژوهشی عمومی بیمه، پژوهشکده بیمه، تهران، ایران</p> <ul style="list-style-type: none">Email: khamesian@irc.ac.irORCID: 0000-0001-6113-4246Homepage: https://www.irc.ac.ir/fa-IR/Irc/5015/page <p>فرید خانی زاده، استادیار گروه پژوهشی بیمه‌های اموال و مسئولیت، پژوهشکده بیمه، تهران، ایران</p> <ul style="list-style-type: none">Email: kanizadeh@irc.ac.irORCID: 0000-0002-0565-2046Homepage: https://www.irc.ac.ir/fa-IR/Irc/5019	
<p>HOW TO CITE THIS ARTICLE</p> <p><i>Esna-Ashari, M.; Khamesian, F.; Khanizadeh, F., (2024). Providing the concept of risk package instead of risk factor in order to classify the risk of policyholders more accurately. Iran. J. Insur. Res., 13(1): 15-28.</i></p> <p>DOI: 10.22056/ijir.2024.01.02</p> <p>URL: https://ijir.irc.ac.ir/article_160308.html?lang=en</p>	