Iranian Journal of Insurance Research JR
Scientific Research Quarterly Journal
Vol. 36/ No.3/ Autumn 2021/pp.85-112
Research Paper

# Discovering Profitable Customers by Data Mining Approach

**Maryam Nezhad-Afrasiabi[1], Akbar Esfahanipour[2], Ali Mohammad Kimiagari[3]**

**Received:** 2019/19/01 **Accepted:** 2021/05/07

**Objective:** Today, customers have become a critical factor in directing investors, producers, and even researchers and innovators. For this reason, organizations need to know about their customers and plan for them. Insurance companies and in general the insurance industry in each country, is one of the most important financial institutions active in financial markets, especially the capital market, which in addition to providing security for economic activities, can play a very fundamental role in providing insurance services. In other words, insurance companies play a vital role in the mobility, dynamic of financial markets and the provision of investable funds in economic activities. In this research, it has been attempted to answer one of the most important questions of insurance organizations, namely, predicting the level of customers' losses and investing on profitable customers.

**Methodology:** Data mining methods were used to discover knowledge to meet business needs and customer relationship management strategies. In addition, an overview of the various applications of data mining in customer relationship management in various insurance companies has been done. In the model implementation stage, a real data-set is used to evaluate the proposed model. To perform the data mining techniques in the insurance industry as data of customers, the vehicle body insurance from 2015 to 2017 has been under investigation. The total number of data used in this study from the beginning was more than 19,356, which during data preparation using Rapidminer 7.1 software became 19,356. After the initial processing, an attempt is made to extract good features from the 15 variables in the data-set that is tangible and help this research in its goal. As a result, by using clustering, drivers are divided into separate clusters based on the amount of loss, and the characteristics of each cluster are expressed. In the clustering section, three algorithms of data mining are examined. First, k-means, k-medoids, and DBSCAN implemented on data-set. Then, the conclusion of three algorithms compared with each other based on the time of calculation and accuracy.

**Finding:** Data mining was a good tool in this research, owing to the large volume of data, to discover the needs and identify customers. The data mining technique which

1. Ph.D Student, Department of Industrial Engineering and Management Systems, Amirkabir University of Technology, Tehran, Iran. mafrasiabi@aut.ac.ir.
2. Associate Professor of Department of Industrial Engineering and Management Systems, Amirkabir University of Technology, Tehran, Iran. esfahaa@aut.ac.ir.
3. Associate Professor of Department of Industrial Engineering and Management Systems, Amirkabir University of Technology, Tehran, Iran. **(Corresponding Author)** kimiagar@aut.ac.ir.



**DOI:** 10.22056/JIR.2021.168537.2461

was the main approach of this study fully covered the information needs by methods such as classification, prediction and clustering. The k-means algorithm was selected as the most optimal one in time and accuracy. In the following, the implementation of the algorithms in the modeling step, the decision tree algorithm was selected and by the decision tree related to the forecasting model, it can be predicted future customers by what characteristics would be in what category. It will be valuable for the insurance companies. Using a decision tree, a forecasting model is proposed to help insurance companies to identify profitable customers which can be used for future plans of organizations.

**Conclusion:** The customer plays an important role in today's industry. Through studying the data obtained from customer behavior, appropriate action can be taken for marketing-related planning and customer acquisition. The use of predictive models and preventive roadmaps has always been one of the goals of the tools that various organizations have been looking for. In this research, the insurance industry as one of the most important pillars of economic in developing countries has been chosen. By reviewing the share of the insurance industry in the economy of a developing country, it can be seen that insurance has a significant role compared to other services. In this study, the role of insurance companies in optimizing the investment process and ways to expand the interaction between insurance examined. Customers can lead to the growth and development of the insurance industry and the capital market and thus the growth and development of the national economy. Therefore, in the implementation of this research, the data of insurance customers have been used and a forecasting model has been presented. As a good prediction model, the decision tree with 86.21% accuracy was the best model that reached in this study. The insurers' income criterion is considered as the node root, which shows the used method can help insurance companies make more profit by focusing on profitable customers.

**Keywords:** Behavior of the Insurers, Clustering, Decision tree, K-means, Discover of Profitable Customers
**JEL Classification**: B31, C38, C22, D12

فصلنامه علمی پژوهشی پژوهشنامه بیمه
مقاله پژوهشی
سال سی‌ششم/شماره ۳/پاییز ۱۴۰۰/ شماره مسلسل۱۴۳/صفحات ۱۱۲-۸۵



# کشف مشتریان سودآور با رویکرد داده‌محور

مریم نژاد افراسیابی[۱]، اکبر اصفهانی‌پور[۲]، علی محمد کیمیاگری[۳]

تاریخ دریافت: ۱۳۹۷/۱۰/۲۹ تاریخ پذیرش: ۱۴۰۰/۰۴/۱۴

## چکیده

**هدف**: امروزه مشتریان به عامل بسیار مهم و حیاتی در هدایت سرمایه‌گذاران، تولیدکنندگان و حتی محققان و نوآوران مبدل گشته‌اند. به همین دلیل، سازمان‌ها نیاز دارند مشتریان خود را بشناسند و برای آنان برنامه‌ریزی کنند. در این پژوهش، تلاش شده تا به یکی از اساسی‌ترین سؤالات سازمان‌های بیمه‌ای، یعنی پیش‌بینی سطح خسارت مشتریان، پاسخ داده شود.

**روش تحقیق**: در پژوهش حاضر از ابزار داده‌کاوی برای داده‌های مشتریان صنعت بیمه، بخش بیمه بدنه خودرو از سال ۱۳۹۴ تا ۱۳۹۶ استفاده شده‌است. تعداد کل داده‌ها که از ابتدا در این پژوهش مورد استفاده قرار می‌گیرد بیش از ۱۹۳۵۶ بوده که در ادامه و در طی آماده‌سازی آن‌ها با استفاده از نرم‌افزار Rapidminer 7.1 تعداد داده‌هایی که در نرم‌افزار لحاظ می‌شود ۱۹۳۵۶ است. پس از پردازش اولیه تلاش می‌شود، از بین ۱۵ متغیر موجود در پایگاه داده ویژگی استخراج شود که ملموس باشد و این پژوهش را در هدف خود یاری دهد. بدین منظور با به کارگیری خوشه‌بندی، رانندگان بر اساس میزان مبلغ خسارت به خوشه‌های مجزا تقسیم می‌شوند و ویژگی‌های هر خوشه بیان می‌شود. در قسمت خوشه‌بندی، ابتدا الگوریتم‌های k-means، k-medoids و DBSCAN استفاده شده‌است. سپس الگوریتم‌های بکار رفته به جهت زمان انجام محاسبات و میزان صحت با یکدیگر مقایسه شدند.

**یافته‌ها**: در نهایت الگوریتم k-means به عنوان الگوریتم بهینه برای این مجموعه داده انتخاب شد. در انتها به کمک درخت تصمیم مدلی پیش‌بینی ارایه می‌شود که شرکت‌های بیمه را در جهت سودآوری بیشتر و کشف مشتریان سودآور کمک می‌کند و برای برنامه‌ریزی و تصمیم‌گیری‌های آتی سازمان قابل استفاده‌است.

**نتیجه‌گیری**: برای پیش‌بینی، درخت تصمیم، با میزان صحت ۸۶/۲۱٪ بهترین مدلی بود که در این پژوهش به آن رسیدیم و در مدل درخت تصمیم ارایه شده معیار درآمد بیمه‌گذار به عنوان گره ریشه درنظرگرفته می‌شود که همین نکته نشان‌دهنده آن است روش بکار رفته می‌تواند به شرکت‌های بیمه کمک کند تا با تمرکز بر مشتریان سودآور به درآمد بیشتری برسند.

**واژگان کلیدی**: رفتار بیمه‌گزاران، خوشه‌بندی، درخت تصمیم، k-means، کشف مشتریان سودآور

**طبقه‌بندی موضوعی**: B31, C38, C22, D12

---

۱. دانشجوی دکتری مهندسی صنایع، دانشکده مهندسی صنایع و سیستم‌های مدیریت، دانشگاه صنعتی امیرکبیر، تهران، ایران.
mafrasiabi@aut.ac.ir

۲. دانشیار گروه آموزشی مهندسی مالی، دانشکده مهندسی صنایع و سیستم‌های مدیریت، دانشگاه صنعتی امیرکبیر، تهران، ایران.
esfahaa@aut.ac.ir

۳. دانشیار گروه آموزشی مهندسی مالی، دانشکده مهندسی صنایع و سیستم‌های مدیریت، دانشگاه صنعتی امیرکبیر ، تهران، ایران.
**(نویسنده مسئول)**. kimiagar@aut.ac.ir

## مقدمه

فشار فزاینده در زندگی روزمره، موجب رشد تقاضای محصولات بیمه‌ای می‌شود. در این بین داده‌کاوی می‌تواند بـه شـرکت‌های بیمـه‌ای در کشـف الگوهـای مفیـد نهفتـه در دل بانک‌های اطلاعاتی مشتریان کمک کند (اوماهسـواری و جانـاکیریمن[1]، ۲۰۱۴). هـدف از این پژوهش نیز تبیین این موضوع است که داده‌کاوی چگونـه در صـنعت بیمـه می‌توانـد مفید باشد، فنون آن چه نتایج در بخش بیمه به دنبـال داشـته و چگونـه تصـمیم‌گیری بـا استفاده از داده‌های بیمه‌ای ممکن می‌شود. از فرآیند داده‌کاوی در صنعت بیمه در مسائلی مانند بهینه‌سازی قیمت‌ها، بهینه‌سازی خدمات، جـذب مشـتریان جدیـد، حفـظ مشـتریان کنونی و کشف کلاهبرداری‌ها در زمینه ادعای خسارات می‌توان استفاده کرد. هـر یـک از این موارد جای بحث فراوان دارد. برای این‌که بتوان مشتریان وفادار را در سـازمان حفـظ کرد و به مشتریان امیدوار و در معرض خطر، خدمات بهتری ارائه داد باید در اولین قـدم آنها را شناخت. با شناخت هر دسته از مشتریان می‌تـوان متناسـب بـا نیـاز آنهـا خـدمات شرکت را بهبود بخشید و بـه ایـن صـورت مشـتریان را حفـظ کـرد. پـس از تشـخیص مشتریان، مشخصات آنها از قبیل خواسته‌ها، نیازهـا و انتظـارات هـر طبقـه را شـناخت و سپس شناخت خدمات شرکت، ایرادات و اتلاف‌های حین خدمات، یـافتن فرصـت‌هـا و تهدیدها می‌توان سوآوری شرکت را بالا برد (قره‌نژاد، ۱۳۸۹).

در ادامه ساختار مقاله به این صورت است که ابتدا پیشینه پژوهش مـرور و مبـانی نظـری تحقیق بیان شده است. پس از آن به روش‌شناسی تحقیق پرداخته شده و توضـیحات لازم درباره انواع روش‌های داده‌کاوی ارائه گردیده است. در مرحله بعـدی بـر روی مجموعـه داده‌ها مطالعه انجام شده و روش‌های معرفی شده بر روی مجموعه داده پیاده می‌شود.

1. Umamaheswari & Janakiraman

## ۱. مروری بر پیشینه پژوهش

بایک و بوکا[1] (۲۰۱۷) به بررسی چگونگی تأثیر پیدایش اینترنت به عنوان یک سنسور داده بر بهبود انتخاب شرکت‌های بیمه و ریسک آن پرداخته‌اند. رادل و همکاران[2] یک مجموعه داده گسترده از بیمه سلامت ملی آلمان را با هدف بررسی مداخله مجدد در درمان احیا بیماران دندانپزشکی مورد مطالعه قرار داده‌اند. رحمان و همکاران[3] (۲۰۱۷) با تحلیل داده‌های شرکت‌های بیمه عمر و بررسی واکنش‌های مشتریان در مقابل سیاست‌های مختلف بیمه به پیش‌بینی رفتار آینده بیمه گذاران پرداخته‌اند. وانک و باروس[4] (۲۰۱۵) یک مجموعه داده پانل متوازن[5] مربوط به شرکت‌های بیمه برزیلی را مورد مطالعه قرارداده‌اند و ناهمسانی[6] در قسمت‌های مختلف را بررسی کرده‌اند. نتایج نشان داد این ناهمسانی در عملکرد تأثیرگذار است. ساندارکومار و راوی[7] (۲۰۱۵) یک روش ترکیبی برای مسائل با داده‌های غیرمتوازن ارائه کردند، که قادر به کشف خودکار کلاهبرداری‌ها در شرکت‌های بیمه می‌باشد. اوشینی و کالدرا[8] (۲۰۱۳) با تمرکز بر اجرای تکنیک‌های حفظ مشتری به موضوع داده‌کاوی در بیمه‌های زندگی و همچنین تحلیل بیمه‌گزاران پرداخته‌اند. آنها به این نتیجه رسیده‌اند که اجرای تکنیک‌های داده‌کاوی در حوزه بیمه‌های عمر، به‌راحتی می‌تواند از ریزش بیمه‌گزاران جلوگیری کند. تاکور و سینگ[9] (۲۰۱۳) مطالعه‌ای بر روی مشتریان بیمه خودرو انجام داده و از الگوریتم بهبود یافته‌تری نسبت به روش k-means استفاده کرده‌اند که با توجه به ویژگی‌های خاص مشتریان قدرت پیش‌بینی عملکرد مشتریان را افزایش می‌دهد. بالاجی و اسریواستا[10] (۲۰۱۲) تکنیک‌های مورد استفاده برای پیش‌بینی داده‌ها برای بیمه‌گزاران بیمه عمر را

1. Baecke & Bocca
2. Raedel et al
3. Rahman et al
4. Wanke & Barros
5. Balanced Panel Data Set
6. Heterogeneity
7. Sundarkumar & Ravi
8. Oshini & Caldera
9. Thakur & Sing
10. Balaji & Srivatsa

طبقه‌بندی کرده‌اند. آنها الگوریتم‌های مختلفی چون روش دسته‌بندی نایو بیز و شبکه‌های بیزین را جهت دسته‌بندی داده‌ها مورد ارزیابی قرار داده‌اند. رانجان[1] (۲۰۱۱) با استفاده از از روش موردکاوی، به مطالعه کاربردهای مدیریت ارتباط با مشتری در شرکت‌های بیمه پرداخته است. بی‌هومیک[2] (۲۰۱۱) به بررسی و کشف تخلفات بیمه‌های خودرویی با استفاده از چندین تکنیک کشف تخلف پرداخته‌اند. همچنین با تمرکز بر رفتار بیمه‌گزاران، به دنبال تحلیل مشتریان سودآور برای شرکت‌های بیمه هستند. موریک و کوپک[3] (۲۰۰۴) با استفاده از یک مطالعه موردی موضوع مدیریت حفظ مشتری در صنعت بیمه را مورد بررسی قرار داده و روش‌هایی برای کاهش مشتری از دست رفته ارائه داده‌اند.

## ۲. مبانی نظری

امروزه سیستم‌های بیمه به سرعت در حال پیشرفت هستند و به دلیل افزایش دغدغه در زندگی روزانه، رشد تقاضای شرکت‌های بیمه‌ای افزایش یافته است. صنعت بیمه یکی از حساس‌ترین صنایع به تغییر رفتار و ارتباطات مشتریان است. زیرا در این صنعت، نقشی پررنگی دارد و مشخص‌کننده و جهت‌دهنده سیاست‌های سازمانی و رفتاری شرکت‌های بیمه‌ای است. به همین دلیل، نظارت و کنترل ارتباطات مشتریان در یک شرکت بیمه‌ای امری بسیار مهم است. به نحوی که شرکت‌های بیمه با کنترل کامل چرخه بازاریابی، فروش و خدمات در تمامی رشته‌های بیمه‌ای و رسیدگی دقیق به تمام درخواست‌های مشتریان می‌توانند گامی موثر در جهت حفظ مشتریان بردارند. از این‌رو، شناسایی مولفه‌های اصلی موثر بر رضایتمندی مشتریان باید در اولویت برنامه‌های شرکت‌های بیمه‌ای قرار گیرد (مترجم و نیاکان، ۱۴۰۰). در این میان داده کاوی به شرکت‌های بیمه کمک می‌کند تا الگوهای مفیدی را از بانک اطلاعات مشتری کشف کنند.

1. Ranjan
2. Bhowmik
3. Morik & Köpcke

بنابراین، پژوهش حاضر با هدف بررسی چگونگی استفاده از داده‌کاوی جهت کشف مشتریان سودآور بیمه بدنه خودرو انجام شده است. زیرا داده‌کاوی می‌تواند در تصمیم‌گیری‌های حیاتی کسب‌وکار به شرکت‌های بیمه کمک کند و دانش تازه به دست آمده را به نتایج قابل اقدام در کسب‌وکار شامل محصول، بازاریابی، تحلیل توزیع خسارت، مدیریت دارایی- بدهی و تحلیل توانایی بازپرداخت دیون تبدیل کند (قره‌خانی و ابوالقاسمی، ۱۳۹۰).

## ۳. روش‌شناسی

پژوهش حاضر از نوع داده‌محور است که با استفاده از فرایند استاندارد داده‌کاوی در صنعت[1] انجام شده است. این مراحل شامل درک مسئله کسب و کار، درک داده‌ها، آماده‌سازی داده‌ها، مدل‌سازی، ارزیابی نتایج و به کارگیری مدل است (حاجی‌حیدری و همکاران، ۱۳۹۰).

قدم اول، آماده‌سازی داده‌ها است که در آن انواع داده‌های پرت، ناقص و اشتباه از مجموعه داده حذف و یا اصلاح شدند. در ادامه به معرفی الگوریتم‌های استفاده شده در این تحقیق پرداخته می‌شود. در قسمت مدل‌سازی، از مدل خوشه‌بندی و از الگوریتم k-means استفاده شده است. نتایج این مرحله از آن جهت اهمیت دارند که علاوه‌بر پی بردن به ویژگی‌های هر خوشه، مقادیر خسارت خوشه‌ها، به‌صورت چند سطحی تعریف شده تا بتوان از آنها در ساخت مدل‌های دسته‌بندی و پیش‌بینی استفاده نمود. لازم به توضیح است که داده‌های مورد استفاده برای مدل‌های دسته‌بندی، به دو دسته آموزشی[2] و آزمایشی[3] تقسیم شده‌اند.

1. Cross-Industry Standard Process for Data Mining (CRISP-DM)
2. Training
3. Testing

### ۳-۱. خوشه‌بندی

برای شروع عملیات روی داده‌ها، ابتدا باید آنها خوشه‌بندی شوند. منظور از خوشه‌بندی نیز عملیاتی است که در آن نمونه‌ها و مشاهدات بر اساس ویژگی‌های مشابه بین یکدیگر به دسته‌های گوناگون تقسیم می‌گردند. یک خوشه، شامل یک مجموعه از نمونه‌هاست که بیشترین شباهت را با یکدیگر و بیشترین تفاوت را با خوشه‌های دیگر دارد.

نکته قابل توجه این که خوشه‌بندی با دسته‌بندی متفاوت است. زیرا، خوشه‌بندی جزء مدل‌های «غیر هدایت شونده» است. اما دسته‌بندی با توجه به حجم داده‌های موجود، یک هدف را یاد گرفته و به ازای داده‌های جدید آن هدف را پیش‌بینی می‌کند. تفاوت دیگر خوشه‌بندی با دسته‌بندی این است که در خوشه‌بندی هیچ مشخصه هدفی تعریف نمی‌شود. در حالی که در دسته‌بندی همواره بایستی مشخصه هدف موجود باشد و بر اساس مشخصه‌های ورودی، پیش‌بینی شود.

در خوشه‌بندی با کمک الگوریتم‌های آن، کل داده‌ها به زیرگروه‌ها و یا خوشه‌های همگن بخش‌بندی می‌شود. نکته حائز اهمیت این که در خوشه‌بندی تعداد خوشه‌ها به شکل دلخواه انتخاب می‌شود. ولی در مسائل بزرگتر که ابعاد بیشتری دارد، در انتخاب تعداد خوشه‌ها باید به شاخص‌های ارزیابی خوشه‌ها که در ادامه توضیح داده شده، توجه گردد.

برای اجرای خوشه‌بندی و به منظور یکسان کردن اثر داده‌های مختلف، در قدم اول باید مقادیر داده‌ها نرمالیزه[1] شود که عموماً از دو روش زیر برای نرمال کردن داده‌ها استفاده می‌شود (شهرابی، ۱۳۹۵):

۱. نرمال سازی بیشینه- کمینه:

$$X^* = \frac{X - Min(X)}{Max(X) - Min(X)} \qquad (۱)$$

۲. نرمال سازی با تابع نرمال استاندارد:

$$X^* = \frac{X - Mean(X)}{SD(X)} \qquad (۲)$$

1. Normalization

پس از نرمال‌سازی داده‌ها با استفاده از الگوریتم‌های نظیر k-means یا Kohonen، کم کردن فاصله‌ی درون خوشه‌ای و زیاد کردن فاصله بین خوشه‌ای مد نظر خواهد بود.

### ۳-۲. الگوریتم K-means

مهم‌ترین الگوریتم خوشه‌بندی k میانگین یا k-means است که در این پژوهش نیز از آن استفاده شده است. این الگوریتم به‌طور خلاصه از مراحل زیر تشکیل شده است:

- ✓ تعداد خوشه‌های مدنظر کاربر را از او می‌پرسد (فرض کنیم k خوشه).
- ✓ به صورت تصادفی k نمونه را به عنوان مراکز خوشه‌های kگانه انتخاب می‌کند.
- ✓ برای هریک از نمونه‌های دیگر فاصله تا این k مرکز را حساب کرده و آن نمونه را به نزدیک‌ترین مرکز خوشه اختصاص می‌دهد.
- ✓ برای هریک از خوشه‌ها، مرکز خوشه جدیدی حساب کرده و به قدم ۳ می‌رود.
- ✓ این کار را آنقدر تکرار می‌کند تا شرط توقف حاصل گردد (شهرابی، ۱۳۹۵).

### ۳-۳. الگوریتم K-medoids

در الگوریتم k-means عنصری که به عنوان نماینده خوشه انتخاب می‌شد با توجه به معیار میانگین است. به این صورت که هر بار فاصله تمام اعضا با نماینده محاسبه و میانگین فاصله‌ها به‌عنوان نماینده جدید انتخاب می‌شود.

اما k-medoids الگوریتمی مبتنی بر شی بوده و نماینده خوشه را از میان خود داده‌ها انتخاب می‌کند. معیار انتخاب نماینده عنصر میانه بوده که در این صورت نماینده انتخابی همیشه عضوی از داده‌ها می‌باشد. هدف از انتخاب میانه کم کردن حساسیت خوشه نسبت به مقادیر بسیار بزرگ و یا داده‌های پرت می‌باشد و فاصله اعضای خوشه همیشه با نماینده‌ای که خودش هم عضوی از اعضا می‌باشد سنجیده می‌شود. مراحل اجرای این الگوریتم به صورت زیر است:

- ✓ مانند k-means ابتدا k نمونه را به عنوان نماینده خوشه‌ها انتخاب می‌شود.

- ✓ برای هر نمونه نزدیکترین نماینده مشخص و نمونه مربوطه را در آن خوشه قرار داده می‌شود.
- ✓ آن‌گاه در k خوشه ایجاد شده مجدد میانه‌ها به‌دست آمده و الگوریتم تکرار می‌گردد.
- ✓ این تکرار را تا زمانی ادامه داده که دیگر میانه‌ها در مرحله جدید با میانه‌های انتخاب شده قبل تفاوت نکنند.

## ۳-٤. الگوریتم DBSCAN

الگوریتم خوشه‌بندی مبتنی بر چگالی یا به اختصار DBSCAN[1] یکی از الگوریتم‌های خوشه‌بندی مهم و مطرح در داده‌کاوی می‌باشد. این الگوریتم از آن جهت مورد توجه است که برخلاف سایر الگوریتم‌های خوشه‌بندی مانند k-Mean، k-medoids یا الگوریتم FCM[2] که وابسته به تعداد خوشه می‌باشند و باید از قبل تعداد خوشه‌ها مشخص باشد، خود می‌تواند تعداد خوشه‌ها را مشخص کند و نیازی نیست که تعداد خوشه‌ها به آن اعلام شود. مزیت این روش به نسبت روش‌های دیگری خوشه‌بندی این است که نسبت به شکل داده‌ها حساس نیست و می‌تواند اشکال غیر منظم را نیز در داده‌ها تشخیص دهد.

این الگوریتم نیز مانند دیگر روش‌های خوشه‌بندی نیازمند روشی برای یافتن نزدیکی داده‌ها است. در این الگوریتم می‌توان از فاصله اقلیدسی جهت اندازه‌گیری فاصله (شباهت) استفاده نمود. برای تشریح الگوریتم، نیازمند آشنایی با پارامترهای $\varepsilon$ و $\mu$ می‌باشد که توضیح داده می‌شود:

- ✓ هر نقطه از داده با نقاط دیگر فاصله‌ای دارد. هر نقطه‌ای که فاصله‌اش با یک نقطه مفروض کمتر از $\varepsilon$ شد به عنوان همسایه آن نقطه حساب می‌شود.
- ✓ هر نقطه مفروض که $\mu$ همسایه داشته باشد، یک نقطه مرکزی است.

1. Density-based spatial clustering of applications with noise (DBSCAN)
2. Firebase cloud messaging (FCM)

در ادامه از مدل درخت تصمیم استفاده نموده و در نهایت با مقایسه میزان صحت[1] و اعتبار هر یک از این مدل‌ها، مدل مدنظر برای تحلیل و تصمیم‌گیری‌های آتی سازمان برگزیده می‌شود. بنابرین قدم‌های صورت گرفته تا حصول نتیجه به صورت زیر است:

- ✓ آماده‌سازی و پالایش داده‌ها
- ✓ خوشه‌بندی به کمک الگوریتم‌های k-means، k-medoids و DBSCAN
- ✓ تقسیم داده‌ها به دو دسته‌ی آموزشی و آزمایشی
- ✓ دسته‌بندی به کمک مدل درخت تصمیم مقایسه‌ی میزان صحت هریک از مدل‌ها جهت تعیین مدل مناسب.

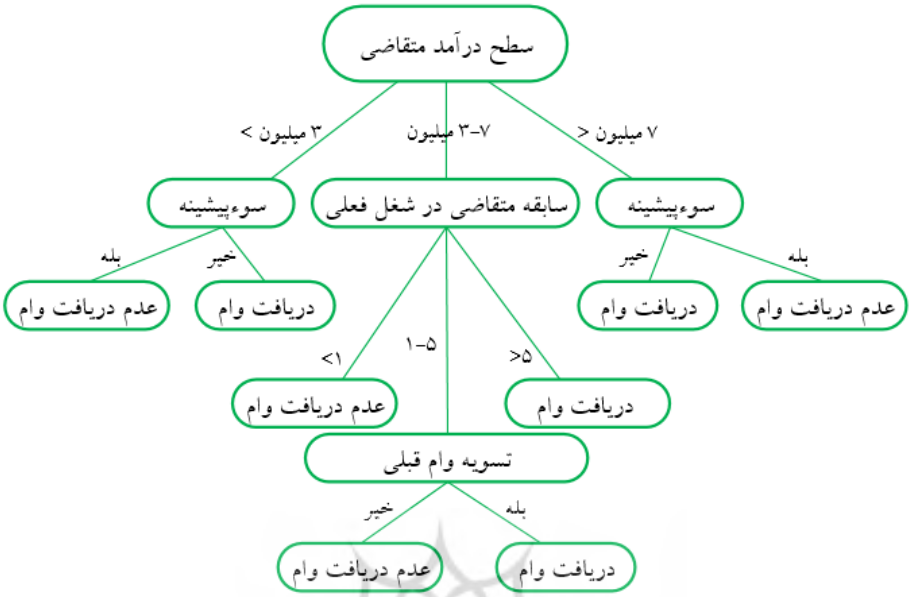
### ۳-۵. درخت تصمیم

درخت تصمیم، روشی معروف برای دسته‌بندی است که نتایج آن در یک نمودار، شبیه ساختار درخت ارائه می‌شود که هر گره نشانگر یک تست بر روی ارزش مشخصه و هر شاخه، خروجی هر تست را نمایش می‌دهد؛ برگ‌های درخت نیز نمایانگر کلاس‌ها هستند.

مفاهیم اصلی در هر درخت تصمیم به شرح زیر می‌باشد:

- ✓ گره[2]: متغیر مستقلی که روی آن آزمون انجام می‌شود.
- ✓ گره ریشه[3]: گره‌ای که در بالاترین نقطه‌ی درخت وجود دارد.
- ✓ برگ[4]: به برچسب دسته، برگ گفته می‌شود.
- ✓ شاخه[5]: مقیاسی که خروجی از آن تعیین می‌شود (غضنفری و همکاران، ۱۳۸۷).

1. Accuracy
2. Node
3. Root Node
4. Leaf
5. Branch




**شکل ۱. نمایش درخت تصمیم برای دریافت وام یک مشتری از بانک**

**منبع:** یافته‌های تحقیق

شکل ۱ ساختار یک درخت تصمیم را نشان می‌دهد که امکان دریافت وام از بانک را برای یک فرد متقاضی نشان می‌دهد. با توجه به این درخت تصمیم بانک برای دادن وام به فرد ابتدا به سطح درآمد او توجه می‌کند و این متغیر به عنوان گره ریشه درنظرگرفته می‌شود. به دنبال آن متغیرهای دیگری نظیر داشتن سوءپیشینه و میزان سابقه فرد در شغل فعلی سایر گره‌های درخت هستند که در تصمیم‌گیری بانک برای دادن وام به فرد متقاضی کمک می‌کند. به‌طور عادی پیچیدگی یک درخت تصمیم با افزایش تعداد مشخصه‌ها افزایش می‌یابد. اگرچه در بعضی شرایط دیده شده است که تنها تعداد کمی از مشخصه‌ها می‌توانند کلاسی را که هر شئ به آن تعلق دارد، تعیین کنند و بقیه مشخصه‌ها کم یا بی‌تأثیرند. اندازه‌گیری کیفیت یک درخت تصمیم، مسئله مهمی است. دقت درخت تعیین شده‌ی دسته‌بندی با استفاده از داده‌های آزمایشی، بطور آشکار یک شاخص مطلوب است.

## ٤. تحلیل داده‌ها

### ٤-١. مجموعه داده‌ها

مجموعه داده‌ها شامل داده‌های مشتریان صنعت بیمه در بخش بیمه بدنه خودرو از سال ۱۳۹۴ تا ۱۳۹۶ است که تعداد ۱۹۳۵۶ داده با استفاده از نرم‌افزار Rapidminer 7.1 آماده‌سازی شد و در نرم‌افزار لحاظ گردیدند. تعداد متغیرهایی که برای این پژوهش درنظر گرفته شد، شامل ۱۵ متغیرِ سن، جنسیت، شهر، گروه خودرو، نوع خودرو، ارزش خودرو، جمع کل حق بیمه، مقصر، محل حادثه، علت حادثه، نوع حادثه، درصد حادثه، تعداد مصدوم، علت پرداخت و مبلغ خسارت بود. البته، ابتدا بر روی داده‌های مشتریان پردازش صورت گرفت. سپس سه روش مختلف خوشه‌بندی k-means، k-medoids و DBSCAN روی داده‌های آموزشی پیاده شد تا بهترین گزینه از بین آن‌ها انتخاب شود. معیار انتخاب روش بهتر نیز سرعت الگوریتم و خوشه‌بندی منطقی و میزان صحت در داده‌ها بود. برای اعتبارسنجی متقابل[1] نیز از 10-fold استفاده گردید.

### ٤-٢. آماده‌سازی داده‌ها

در این مرحله سعی شد با توجه به داده‌های جمع‌آوری شده تحلیلی آماری صورت گیرد. در این مرحله تحلیل‌های مختلفی می‌توان انجام داد، که در ادامه جزییات آن تشریح می‌شود. رکوردهای موجود در این پایگاه داده، شامل اطلاعات نامعتبر زیادی به شکل داده‌های پرت، داده‌های تکراری، رکوردهای ناقص و رکوردهای اشتباه بود. عملیات آماده‌سازی داده‌ها به کمک نرم‌افزارهای Minitab و Excel صورت گرفت. تعداد بسیار کمی از این رکوردها قابل اصلاح بود و تصحیح شدند. به‌طور مثال سال تولد فرد به‌جای ۱۳۳۳، بصورت ۱۳۳۳۳، نوشته شده بود. البته، تعداد چنین مواردی بسیار کم بود. در بسیاری از رکوردها، در قسمت سال تولد فرد به اشتباه، سالی ثبت شده بود که اطلاعات بیمه‌ای جمع‌آوری شده بودند (یکی از سال‌های ۱۳۹۴ تا ۱۳۹۶). بنابراین، در چنین مواردی تنها به خاطر یک ثبت نادرست ناگزیر باید رکورد حذف می‌شد. تعدا قابل

1. Cross Validation

توجهی رکورد تکراری نیز در داده‌های این سه سال وجود داشت که اطلاعات یک فرد، اغلب ۲ بار پشت سر هم و در معدود دفعاتی تا ۳ بار پشت سر هم ثبت شده بود که همه این موارد تکراری از پایگاه داده حذف گردیدند. داده‌های پرت از طریق نمودار جعبه‌ای شناسایی شد و به کمک ابزارSubset Worksheet در Minitab و ابزار Filter در Excel، مقادیر نامعتبر از مجموعه داده، شناسایی و حذف گردیدند که در نهایت تعداد ۱۹۳۵۶ رکورد برای انجام مدل‌سازی و تحلیل داده‌ها مورد استفاده قرار گرفتند. اصلاح دیگری که در داده‌ها صورت گرفت، اصلاحاتی بر روی اسامی شهرها بود. از آنجایی که تعداد شهرها در این مجموعه داده بسیار زیاد بود (۲۹۶ نام مجزا)، در صورتی که بدون تغییر باقی می‌ماند، نرم‌افزار، نوع فیلد مورد نظر را تشخیص نمی‌داد و از فرایند مدل‌سازی حذف می‌شدند. بنابراین برای حفظ تأثیر این عامل، نام ۱۶ شهر که بیش از ۷۷ درصد داده‌ها را تشکیل می‌داد، حفظ و مابقی به "Others" تغییر نام یافتند. نام این ۱۶ شهر به شرح جدول ۱ است:

**جدول ۱. نام شهرها**

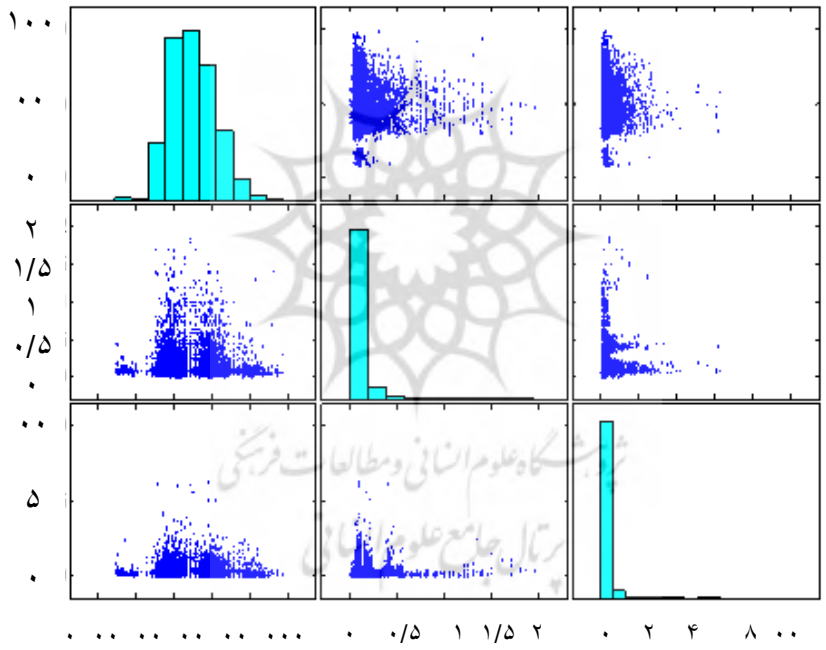
| | | | |
|---|---|---|---|
| اهواز | اراک | بندر عباس | بندر انزلی |
| اصفهان | همدان | کرج | مشهد |
| قزوین | قم | رشت | شیراز |
| تبریز | تهران | تنکابن | یزد |

**منبع:** یافته‌های تحقیق

در مورد فیلد «گروه وسیله نقلیه» نیز، به علت تفاوت در فراوانی‌ها، رکوردهای مربوط به آن به ۳ دسته »شخصی»، «بارکش» و «مابقی» تقسیم شدند.

پس از آن تلاش شد، از بین متغیرهای موجود در پایگاه داده ویژگی‌ای استخراج شود که ملموس باشد و پژوهش را در دستیابی به هدف یاری نماید. جمع کل حق بیمه در واقع داده استخراجی از مجموع حق بیمه شکست، حق بیمه سرقت، حق بیمه سیل، حق بیمه پاشیدن رنگ، حق بیمه ماده ۱۰، حق بیمه حوادث، حق بیمه حوادث شخصی، حق بیمه

برخورد قطعات، حق بیمه بند ۷ ماده ۹، حق بیمه ایاب و ذهاب، حق بیمه خطر اصلی وسایل اضافی و حق بیمه خطر اصلی می‌باشد. سن نیز داده‌ای است که با توجه به سال تولد مشتریان که در سیستم اطلاعات بیمه ثبت شده و با در نظر گرفتن سال جاری محاسبه شده و به‌عنوان یکی از ویژگی‌های اصلی مورد استفاده قرار گرفت. در ادامه ارتباط بین ویژگی‌های استخراج شده سن، جمع کل حق بیمه و ارزش خودرو مورد توجه قرار گرفت. از آنجا که ارزش خودرو از ملاک‌های مطلوب برای شرکت‌های بیمه است، در ادامه نمودار اسکتر پلات[1] ارزش خودرو به همراه دو ویژگی استخراج شده سن و جمع کل حق بیمه مطابق شکل ۲ ترسیم شد.




**شکل ۲. نمودار اسکتر پلات سن، ارزش خودرو و جمع کل حق بیمه**

**منبع:** یافته‌های تحقیق

1. Scatter plot

در نهایت متغیرهای سن، جنسیت، شهر، گروه خودرو، نوع خودرو، ارزش خودرو، جمع کل حق بیمه، مقصر، محل حادثه، علت حادثه، نوع حادثه، درصد حادثه، تعداد مصدوم و علت پرداخت به عنوان متغیر مستقل و عوامل تأثیرگذار در سانحه و مبلغ خسارت به عنوان متغیر وابسته مدنظر قرار گرفت (به شرح جدول ۲).

**جدول ۲. داده‌های آماده‌شده و ویژگی آن‌ها**

| ردیف | ویژگی | نوع ویژگی | سطوح اندازه‌گیری | شرح |
|---|---|---|---|---|
| ۱ | سن | عددی | فاصله | سن راننده در سال |
| ۲ | جنسیت | رده‌ای | باینری | زن، مرد |
| ۳ | شهر | رده‌ای | ترتیبی | محل اقامت راننده |
| ۴ | گروه خودرو | رده‌ای | ترتیبی | سواری، بارکش، موتورسیکلت، ... |
| ۵ | نوع خودرو | رده‌ای | ترتیبی | انواع خودرو |
| ۶ | ارزش خودرو | عددی | فاصله | قیمت خودرو |
| ۷ | جمع کل حق بیمه | عددی | فاصله | مبلغ حق بیمه |
| ۸ | مقصر | رده‌ای | ترتیبی | بیمه‌گذار، ناشناخته و ... |
| ۹ | محل حادثه | رده‌ای | ترتیبی | شهرهای مختلف وقوع حادثه |
| ۱۰ | علت حادثه | رده‌ای | ترتیبی | انواع علت حادثه |
| ۱۱ | نوع حادثه | رده‌ای | ترتیبی | شکست شیشه، آتش سوزی و ... |
| ۱۲ | درصد حادثه | عددی | نسبی | درصدهای مختلف از میزان حادثه |
| ۱۳ | تعداد مصدوم | عددی | فاصله | تعداد مصدومین در حادثه |
| ۱۴ | علت پرداخت | رده‌ای | ترتیبی | خسارت، کارشناسی و بازیافت |
| ۱۵ | مبلغ خسارت | پیوسته | فاصله | انواع هزینه‌ها |

**منبع:** یافته‌های تحقیق

### ٤-٣. پیاده‌سازی مدل

قدم اول در خوشه‌بندی، نرمال‌کردن داده‌ها است. اطلاعات مربوط به میزان خسارت در ابتدا از طریق فرمول $x^* = \frac{X-Mean(X)}{SD(X)}$ در نرم افزار نرمال شده و سپس از مدل k-means برای خوشه‌بندی استفاده شده است. در قدم دوم ویژگی مبلغ خسارت به‌عنوان کلاس هدف درنظر گرفته شد. سپس در مرحله اول روش k-means پیاده‌سازی شد. مطابق با آن داده‌ها به سه سطح مطابق جدول ۳ تقسیم شدند.

**جدول ۳. خوشه‌بندی با استفاده از مبلغ خسارت**

| عنوان خوشه | مبلغ خسارت |
|---|---|
| سطح ۱ | [-∞-1900350] |
| سطح ۲ | [1900350-4427651] |
| سطح ۳ | [4427651-∞] |

**منبع:** یافته‌های تحقیق

سرعت اجرای این الگوریتم نزدیک به ۱ دقیقه است.

در مرحله دوم روش k-medoids پیاده می‌شود که در این روش زمان الگوریتم ۳۶ دقیقه است و همه داده‌ها در یک خوشه قرارمی‌گیرند.

در مرحله سوم الگوریتم DBSCAN با مقدار $\varepsilon$ برابر ۱ و $\mu$ برابر ۵ اجرا می‌شود. مدت اجرای الگوریتم ۴ دقیقه و ۳۷ ثانیه است و به دنبال آن تعداد خوشه‌های حاصل شده ۵۵۸ است و بسیار بالاست و هیچ‌گونه الگوی منطقی برای تحلیل خوشه‌ها نمی‌توان یافت.

با مقایسه نتایج سه الگوریتم بکار رفته در این پژوهش، برای خوشه‌بندی از داده‌های مربوط به «مبلغ خسارت» استفاده شده‌است و مطابق با روش k-means به سه سطح تقسیم می‌شوند که در جدول ۱ قابل مشاهده است. لازم به ذکر است که انجام خوشه‌بندی، بسته به هدف کار، هدف سازمان و نظر خبره می‌تواند تغییر کند. مثلا این امر نیز امکان‌پذیر بود که از تمام فیلدها جهت خوشه‌بندی استفاده شود که این شکل از

خوشه‌بندی نتایجی متفاوت برای هدفی متفاوت را به دنبال خواهد داشت. مثلا در این حالت، شهر بیمه‌گر، سن، جنسیت و از این حیث ویژگی‌های فردی که ممکن است در بسیاری از افراد نیز مشترک باشد، ملاکی برای خوشه‌بندی می‌شوند و نتایج را تغییر می‌دهند.

### ٤-٤. بهبود مدل

تعداد کل داده‌ها بیمه بدنه که از ابتدا مورد استفاده قرار گرفته بیش از ۱۹۳۵۶ بود. اما پس از آماده‌سازی داده‌ها، تعداد داده‌هایی که برای خوشه‌بندی در نرم‌افزار Rapidminer 7.1 لحاظ شد ۱۹۳۵۶ است.

### ٤-٤-١. گام یک

ابتدا با استفاده از کلیه ویژگی‌هایی که از شرکت بیمه استخراج شده بود، درخت تصمیم پیاده‌سازی شد. نتیجه‌ای که از اجرای مدل در نرم‌افزار حاصل شد به این صورت بود که ویژگی سال ساخت خودرو به عنوان ریشه درنظر گرفته شد. از آنجا که به لحاظ منطقی چنین ویژگی نه برای شرکت‌های بیمه و نه برای مشتریان معیار اصلی برای تصمیم‌گیری نیست، نباید آن‌قدر مهم باشد که ریشه قرار بگیرد. در گام بعدی سعی شد ویژگی‌های مفید نگهداشته شود و بقیه ویژگی‌ها حذف شوند.



### ٤-٤-٢. گام دوم

ویژگی‌های درنظر گرفته شده مطابق جدول ۴ می‌باشد. نوع ویژگی مطابق گزینه‌ای است که در نرم‌افزار انتخاب شده است.

**جدول ٤. ویژگی‌های در نظر گرفته شده**

| ردیف | ویژگی | نوع ویژگی | مقادیر گم‌شده |
|---|---|---|---|
| ۱ | سن | عدد صحیح | ۰ |
| ۲ | جنسیت | چندجمله‌ای | ۰ |
| ۳ | شهر | چندجمله‌ای | ۰ |
| ۴ | گروه خودرو | چندجمله‌ای | ۰ |
| ۵ | نوع خودرو | چندجمله‌ای | ۰ |
| ۶ | ارزش خودرو | عدد صحیح | ۰ |
| ۷ | جمع کل حق بیمه | عدد صحیح | ۰ |
| ۸ | مقصر | چندجمله‌ای | ۰ |
| ۹ | محل حادثه | چندجمله‌ای | ۰ |
| ۱۰ | علت حادثه | چندجمله‌ای | ۰ |
| ۱۱ | نوع حادثه | چندجمله‌ای | ۰ |
| ۱۲ | درصد حادثه | چندجمله‌ای | ۲ |
| ۱۳ | تعداد مصدوم | عدد صحیح | ۹۹۹۹ |
| ۱۴ | علت پرداخت | چندجمله‌ای | ۰ |
| ۱۵ | مبلغ خسارت | عدد صحیح | ۰ |

**منبع**: یافته‌های تحقیق

نتیجه‌ای که از اجرای مدل در نرم‌افزار حاصل شد به این صورت بود که ویژگی ارزش خودرو به عنوان گره ریشه درخت درنظر گرفته شد.

روش ارائه شده در هر پژوهشی باید از لحاظ اعتبار، مورد ارزیابی قرار گیرد. بنابرین از آنجا که این پژوهش از نوع «داده‌محور» است، میزان صحت هر مدل از طریق ماتریس اختلال[1] سنجیده می‌شود. در حالت کلی اگر دو دسته «درست[2]» و «نادرست[3]» داشته باشیم، چهار حالت ممکن برای نتایج وجود خواهد داشت و ماتریس اختلال مطابق جدول ۵ می‌باشد. چهار حالت بیان شده در ماتریس اختلال جدول ۵ شامل به درستی

1. Confusion matrix
2. True
3. False

نتیجه آزمون مثبت باشد(TP[1])، به اشتباه نتیجه آزمون مثبت باشد(FP[2])، نتیجه آزمون به اشتباه منفی باشد(FN[3]) و نتیجه آزمون به درستی منفی باشد(TN[4]) است

**جدول ۵. ماتریس اختلال**

| | | مشاهده شده | |
|---|---|---|---|
| | | درست | نادرست |
| پیش‌بینی | درست | TP | FP |
| | نادرست | FN | TN |

منبع: یافته‌های تحقیق

در این پژوهش با سه سطح ایجاد شده ناشی از مرحله خوشه‌بندی، شش حالت ممکن در ماتریس اختلال ایجاد شد که نتایج در جدول ۶ خلاصه شده است. میزان صحت[5] مدل به‌کار رفته مطابق جدول ۶ نزدیک به ۵۵/۲۸ درصد بدست آمده‌است.

**جدول ٦. ماتریس اختلال حاصل از داده‌های پژوهش**

| Accuracy | ٪۵۵,۲۸ | | | | |
|---|---|---|---|---|---|
| Confusion matrix | | | | | |
| Class precision | سطح ۳ | سطح ۲ | سطح ۱ | | |
| ٪۸۰,۳۶ | ۶۰۷ | ۱۹۴۲ | ۲۸۱۰ | [-∞-1900350] | سطح ۱ |
| ٪۵۰,۴۷ | ۳۷ | ۲۸ | ۱۶ | [1900350-4427651] | سطح ۲ |
| ٪۸۴,۸۵ | ۲۶۸۹ | ۱۳۳۷ | ۵۳۳ | [4427651-∞] | سطح ۳ |
| | ٪۴۵,۷۸ | ٪۹۶,۳۱ | ٪۶۴,۳۲ | Class recall | |

منبع: یافته‌های تحقیق

1. True Positives
2. False Positives
3. True Negatives
4. False Negatives
5. Accuracy

مقدار مربوط به میزان صحت، دقت کلاس[1] و فراخوانی کلاس[2] به‌ترتیب از روابط ۳، ۴ و ۵ پیروی می‌کند[۱۱]:

$$\text{Accuracy} = (TP + TN)/(TP + TN + FP + FN) \quad (۳)$$

$$\text{Precision} = TP/(TP + FP) \quad (۴)$$

$$\text{Recall} = TP/(TP + TN) \quad (۵)$$

با نتیجه حاصله از نرم‌افزار می‌توان چنین تحلیل کرد که مدل بهبود یافته است. زیرا ریشه درخت معیار ارزش خودرو درنظر گرفته شده که نسبت به حالت مشابه در گام یک معیار بهتری است. اما دقت مدل پایین می‌باشد. به دنبال بهبود مدل در گام بعدی سعی شد در داده‌های مورد استفاده تحلیلی عمیق‌تری صورت گیرد تا ویژگی اضافه مدنظر حذف و ویژگی‌های مفیدتری استخراج گردد و در ادامه مقادیر گم شده در داده‌ها نیز بهبود یابد.

### ٤-٤-٣. گام سوم

با تحلیل و مطالعه بیشتر در داده‌ها این نتیجه حاصل شد که با استفاده از دو ویژگی مبلغ خسارت و مجموع حق بیمه می‌توان در نهایت خالص درآمد شرکت را به‌دست آورد. با استخراج ویژگی به نام درآمد یا عایدی شرکت بیمه، تعداد ویژگی‌های درنظر گرفته شده ۱۶ است که در جدول ۷ خلاصه شده‌اند.



1. Class Precision
2. Class Recall

**جدول ۷. ویژگی‌های نهایی در نظرگرفته شده**

| ردیف | ویژگی | نوع ویژگی | مقادیر گم‌شده |
|---|---|---|---|
| ۱ | سن | عدد صحیح | ۰ |
| ۲ | جنسیت | چندجمله‌ای | ۰ |
| ۳ | شهر | چندجمله‌ای | ۰ |
| ۴ | گروه خودرو | چندجمله‌ای | ۰ |
| ۵ | نوع خودرو | چندجمله‌ای | ۰ |
| ۶ | ارزش خودرو | عدد صحیح | ۰ |
| ۷ | جمع کل حق بیمه | عدد صحیح | ۰ |
| ۸ | مقصر | چندجمله‌ای | ۰ |
| ۹ | محل حادثه | چندجمله‌ای | ۰ |
| ۱۰ | علت حادثه | چندجمله‌ای | ۰ |
| ۱۱ | نوع حادثه | چندجمله‌ای | ۰ |
| ۱۲ | درصد حادثه | چندجمله‌ای | ۰ |
| ۱۳ | تعداد مصدوم | عدد صحیح | ۰ |
| ۱۴ | علت پرداخت | چندجمله‌ای | ۰ |
| ۱۵ | مبلغ خسارت | عدد صحیح | ۰ |
| ۱۶ | درآمد | عدد صحیح | ۰ |

**منبع**: یافته‌های تحقیق



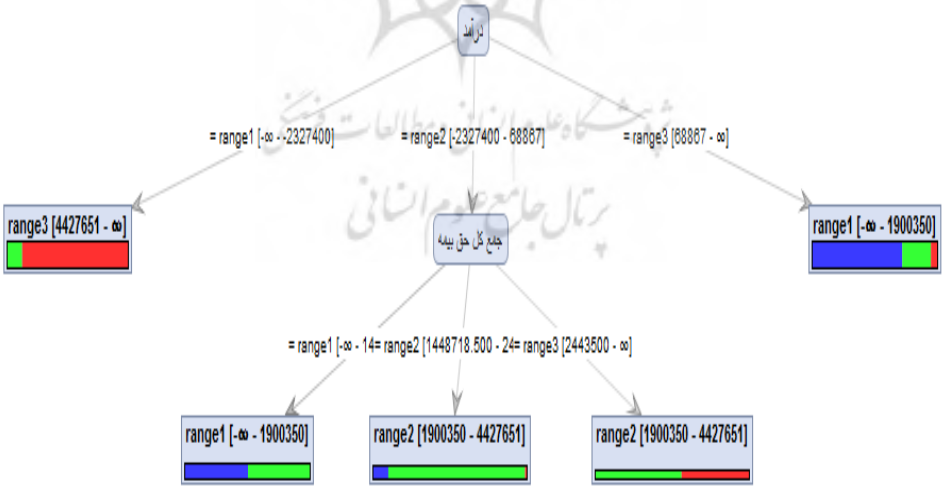
با استخراج ویژگی به نام درآمد نرم‌افزار دوباره اجرا شد و نتیجه ماتریس اختلال مطابق جدول ۸ است.

**جدول ۸. ماتریس اختلال مدل بهبود یافته**

| Accuracy | ٪۸۶,۲۱ | | | | |
|---|---|---|---|---|---|
| Confusion matrix | | | | | |
| | | سطح ۱ | سطح ۲ | سطح ۳ | Class precision |
| سطح ۱ | [-∞-1900350] | ۳۲۴۱ | ۱۰۴۰ | ۲۳ | ٪۸۱,۶۷ |
| سطح ۲ | [1900350-4427651] | ۱۰۵ | ۲۱۵۱ | ۸۲ | ٪۹۴,۵۲ |
| سطح ۳ | [4427651-∞] | ۱۳ | ۱۱۶ | ۳۲۲۸ | ٪۹۷,۶۳ |
| Class recall | | ٪۹۷,۲۶ | ٪۷۶,۱۴ | ٪۹۷,۶۹ | |

**منبع:** یافته‌های تحقیق

در جدول ۸ مشاهده می‌شود که دقت مدل در هر سه سطح در مقایسه با جدول ۶ بهبود یافته و به حد قابل قبولی رسیده است. علاوه‌بر آن درخت تصمیم حاصله مطابق شکل ۵ می‌باشد که در آن ویژگی درآمد به‌عنوان گره ریشه درنظرگرفته شده که یک معیار اصلی برای شرکت‌های بیمه محسوب می‌شود. با توجه به این درخت تصمیم مشتریان با جمع کل حق بیمه کمتر و درآمد بالاتر برای بیمه‌گزاران سود بیشتری به ارمغان می‌آورند.




**شکل ۵. درخت تصمیم (به ترتیب رنگ‌های آبی، سبز و قرمز نشاندهنده روند سودآوری بیشتر به کمتر است)**

**منبع:** یافته‌های تحقیق

## ۵. جمع‌بندی و پیشنهادها

استفاده از روش‌های داده کاوی برای تحلیل داده‌های مشتریان در میان شرکت‌های بزرگ و کوچک در سال‌های اخیر رشد فزاینده‌ای داشته است. لذا صاحبان بسیاری از کسب‌وکارها می‌خواهند بدانند چگونه با استفاده از ابزارهایی مانند داده‌کاوی می‌توانند به فروش بیشتری دست یابند و به دنبال آن رویکرد بازاریابی داده‌محور، داده‌کاوی و تحلیل رفتار مشتریان را در کسب‌وکار خود به اجرا در آورند. در این راستا، پژوهش حاضر سعی نمود با تکیه بر فنون مختلف داده‌کاوی، به یکی از نیازهای اساسی صنعت بیمه، یعنی میزان سوآوری حاصل از مشتریان پاسخ دهد. برای این منظور، در ابتدا تعداد بیش از ۱۹۳۵۶ داده از یک شرکت بیمه بدنه خودرو از سال ۱۳۹۴ تا سال ۱۳۹۶ مورد مطالعه قرار گرفت. پس از آن پردازش اولیه روی داده‌ها صورت گرفت و با استفاده از نرم‌افزار Rapidminer 7.1 داده‌ها پاکسازی شدند. در ادامه برای کشف مشتریان سودآور، تنها به یک یا دو الگوریتم اکتفا نشد و الگوریتم‌های k-means، k-medoids و DBSCAN که وظیفه خوشه‌بندی را به عهده دارند، به کار گرفته شد تا ضمن آشنایی با عملکرد آن‌ها و این‌که هر یک چگونه سازمان را در نیل به هدف یاری می‌کنند، با مقایسه میزان صحت هر یک، الگوریتمی که بیش از سایرین قابل اعتماد است، برای تصمیم‌گیری‌های آتی، انتخاب و مورد بررسی قرار گیرد که نتایج نشان داد از میان سه روش بالا، روش k-means برای خوشه‌بندی بهتر است. برای پیش‌بینی نیز درخت تصمیم با میزان صحت ۸۶/۲۱ درصد بهترین مدلی بود که این پژوهش به آن دست یافت و در مدل درخت تصمیم ارایه شده معیار درآمد بیمه‌گذار به‌عنوان گره ریشه درنظرگرفته شد که این نکته نشان‌دهنده آن است روش به‌کار رفته می‌تواند به شرکت‌های بیمه کمک کند تا با تمرکز بر مشتریان سودآور به درآمد بیشتری برسند. لازم به توضیح است که ممکن است با کاهش تعداد داده‌ها میزان صحت مدل تغییر کند و حتی افزایش یابد. ولی از آنجایی که از ابتدای انجام این پژوهش سعی بر این بود که یک مساله تجاری حل شود، حجم زیاده داده‌ها بدون کاستن تعداد خاصی از آن‌ها بکار گرفته شد. شایان ذکر است که انجام خوشه‌بندی، بسته به هدف کار، هدف سازمان و نظر خبره می‌تواند تغییر کند. همان‌طور که عنوان شد در این پژوهش معیار مبلغ خسارت به‌عنوان کلاس هدف برای خوشه‌بندی انتخاب شد که می‌توان از بین داده‌های موجود معیار دیگری برای خوشه‌بندی انتخاب نمود.

## ملاحظات اخلاقی

### حامی مالی

این مقاله حامی مالی ندارد.

### مشارکت نویسندگان

تمام نویسندگان در آماده سازی این مقاله مشارکت کرده‌اند.

### تعارض منافع

بنا به اظهار نویسندگان، در این مقاله هیچ‌گونه تعارض منافعی وجود ندارد.

### تعهد کپی‌رایت

## منابع

مترجم، کیومرث و نیاکان، لیلی. (۱۴۰۰). سنجش و ارزیابی رضایت‌مندی مشتریان بیمه‌های زندگی. پژوهشنامه بیمه، ۳۶(۱): ۱۱۹-۸۷.

حاجی‌حیدری، نسترن.، خالهء، سامرند و فراهی، احمد. (۱۳۹۰). طبقه‌بندی میزان ریسک بیمه‌گذاران بیمه بدنه خودرو با استفاده از الگوریتم‌های داده‌کاوی (مورد مطالعه: یک شرکت بیمه). پژوهشنامه بیمه، ۲۶(۴): ۱۲۹-۱۰۷.

شهرابی، جمال. (۱۳۹۵). داده‌کاوی. تهران: انتشارات دانشگاه صنعتی امیرکبیر(پلی تکنیک تهران)، چاپ اول.

غضنفری، مهدی، علیزاده، سمیه و تیمورپور، بابک. (۱۳۸۷). داده‌کاوی و کشف دانش. تهران: انتشارات دانشگاه علم و صنعت، چاپ اول.

قره‌نژاد، سحر. (۱۳۸۹). لزوم حفظ مشتریان بیمه با استفاده از ابزارهای داده‌کاوی. *تازه‌های جهان بیمه*، ۱۳(۱۵۰-۱۵۱): ۲۳-۱۵.

قره‌خانی، محسن و ابوالقاسمی، مریم. (۱۳۹۰). کاربردهای داده‌کاوی در صنعت بیمه. *تازه‌های جهان بیمه*، ۱۴(۱۵۸)، ۲۱-۵.

Baecke, P. & Bocca, L. (2017). The value of vehicle telematics data in insurance risk selection processes. *Decision Support Systems*, 98: 69–79.

Balaji, S. & Srivatsa, S. K. (2012). Decision tree induction based classification for mining life insurance databases. *International Journal of Computer Science and Information Technology & Security*, 2(3): 699-703.

Bhowmik, R. (2011). Detecting auto insurance fraud by data mining techniques. *Journal of Emerging Trends in Computing and Information Sciences*, 2(4): 156-162.

Morik, K. & Köpcke, H. (2004). Analyzing customer churn in insurance data –A case study. *European Conference on Principles of Data Mining and Knowledge Discovery*, 325-336.

Oshini, G. T. L. & Caldera H. A. (2013). Mining life insurance data for customer attrition analysis. *Journal of Industrial and Intelligent Information*, 1(1): 52-58.

Raedel, M., Hartmann, A., Priess, H. W., Bohm, S., Samietz, S., Konstantinidis, I. & Walter, M. H. (2017). Re-interventions after

restoring teeth—mining an insurance database. *Journal of Dentistry*, 57: 14–19.

Rahman, M. S., Arefin, K. Z., Masud, S., Sultana, S. & Rahman, R. M. (2017). Analyzing life insurance data with different classification techniques for customers' behavior analysis. *Asian Conference on Intelligent Information and Database Systems*, 15-25.

Ranjan, R. (2011). Self insurance and insurance demand under self-deception. *Asia-Pacific Journal of Risk and Insurance*, 5(2): 1-27.

Sundarkumar, G. G. & Ravi, V. (2015). A novel hybrid under sampling method for mining unbalanced datasets in banking and insurance. *Engineering Applications of Artificial Intelligence*, 37: 368–377.

Thakur, S. SÙ& Sing, J. K. (2013). Mining customer's data for vehicle insurance prediction system using k-means clustering - An Application. *International Journal of Computer Applications in Engineering Sciences*, 3(4): 148-153.

Umamaheswari, K. & Janakiraman, S. (2014). Role of data mining in insurance industry. *An International Journal of Advanced Computer Technology*, 3(6): 961-966.

Wanke, P. & Barros, C. P. (2015). Efficiency drivers in Brazilian insurance: A two-stage DEA meta frontier-data mining approach. *Economic Modelling*, 53(C): 8–22.