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

Predicting the Purchase of Self-Employed Pension Schemes in the Iranian Social Security Organization Using Decision Tree and Random Forest Classification Algorithms

Yasin Ghasemi[1] | **Abbas Khandan**[2*] | **Narges Akbarpour Roshan**[3] |

1. Master Student, Faculty of Economics, Kharazmi University, Tehran, Iran. Email: Yasin8769@gmail.com
2. Assistant Professor, Faculty of Economics, Kharazmi University, Tehran, Iran. (Corresponding Author), Email: khandan.abbas@khu.ac.ir
3. Assistant Professor, Faculty of Economics, Institute for Humanities and Cultural Studies, Tehran, Iran. Email: n.akbarpour@ihcs.ac.ir

Article Info

Article type:
Research Article

Article history:
Received:
2022/11/15

Received in revised form:
2022/11/15

Accepted:
2023/04/15

Keywords:
Pension Insurance of self-employed, Social Security Organization, Data mining, Machine learning, Classification

JEL:
H55 ،C81 ،J26.

ABSTRACT

The pension coverage of the Iranian Social Security Organization for self-employed workers is offered at three contribution rates of 12, 14 and 18 percent, but looking at the statistics shows that the demand for these types of insurances is low. This research investigates the characteristics of these insured groups by using data mining and applying two machine learning algorithms, decision tree and random forest, and predicts their behavior by providing a classification model. This will help the Social Security Organization to improve customer relationship management. For this purpose, the information of 1286174 insured persons of self-employed in 2020 was used, which includes the characteristics of age, gender, average monthly income, the years of service, and the type of self-employed pension scheme. The obtained results show that women mainly apply for the scheme with 12 percent contribution, while men tend to be covered by schemes with contribution rates of 14

and 18 percent due to the burden of supporting the family. Also, for men, the demand for schemes of 14 and 18 percent increases with the increase of age, income and years of service, but there are no such trends for women. According to the obtained results, years of service and then gender are decisive in choosing the type of pension scheme in such a way that according to the prediction of the model, people with less than 4.5 years of service are known as definite applicants for 12 percent self-employed pension scheme.

Cite this article: Ghasemi, Yasin; Khandan, Abbas; Akbarpour Roshan, Narges. (2023). Predicting the Purchase of Self-Employed Pension Schemes in the Iranian Social Security Organization Using Decision Tree and Random Forest Classification Algorithms. *Journal of Economic Modeling Research*, 13 (47), 115-165.
DOI: 0000000000000000000

Publisher: Kharazmi University

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

پیش‌بینی خرید بیمه‌نامه حرف و مشاغل آزاد سازمان تامین اجتماعی با استفاده از الگوریتم طبقه‌بندی درخت تصمیم و جنگل تصادفی

یاسین قاسمی[۱] | عباس خندان*[۲] | نرگس اکبرپور روشن[۳]

۱. دانشجوی کارشناسی ارشد مهندسی صنایع، دانشکده اقتصاد، دانشگاه خوارزمی، تهران، ایران yasein8769@gmail.com
۲. استادیار اقتصاد، دانشکده اقتصاد، دانشگاه خوارزمی، تهران، ایران. (نویسنده مسئول) khandan.abbas@khu.ac.ir
۳. استادیار اقتصاد، پژوهشکده اقتصاد، پژوهشگاه علوم انسانی و مطالعات فرهنگی، تهران، ایران n.akbarpour@ihcs.ac.ir

اطلاعات مقاله

نوع مقاله:
مقاله پژوهشی

تاریخ دریافت:
۱۴۰۱/۰۸/۲۴

تاریخ ویرایش:
۱۴۰۱/۰۸/۲۴

تاریخ پذیرش:
۱۴۰۲/۰۱/۲۶

واژه‌های کلیدی:
بیمه بازنشستگی حرف و مشاغل آزاد، سازمان تأمین اجتماعی، داده‌کاوی، یادگیری ماشین، طبقه‌بندی.

طبقه‌بندی JEL:
H55 ،C81 ،J26.

چکیده

پوشش بیمه سازمان تامین اجتماعی برای حرف و مشاغل آزاد به صورت اختیاری در سه نرخ ۱۲، ۱۴ و ۱۸ درصد ارائه می‌شود اما نگاه به آمار نشان می‌دهد که تقاضای این بیمه‌نامه‌ها بسیار پایین است. این پژوهش با استفاده از داده‌کاوی و با به‌کارگیری دو الگوریتم یادگیری ماشین یعنی درخت تصمیم و جنگل تصادفی به بررسی مشخصه‌های خریداران این نوع بیمه‌نامه‌ها پرداخته و با ارائه یک مدل طبقه‌بندی، رفتار آن‌ها را پیش‌بینی می‌کند تا از این طریق به سازمان تأمین اجتماعی در جهت بهبود مدیریت ارتباط با مشتری کمک کند. برای این منظور، از اطلاعات ۱۲۸۶۱۷۴ نفر از خریداران انواع بیمه‌نامه‌های حرف و مشاغل آزاد سال ۱۳۹۹ استفاده شد که مشخصه‌های سن، جنسیت، متوسط درآمد ماهانه، میزان سابقه کار و نوع بیمه‌نامه خریداری شده را در بر می‌گیرد. نتایج به دست آمده نشان می‌دهند که زنان به‌طور عمده متقاضی بیمه‌نامه با نرخ ۱۲ درصد هستند در حالی‌که مردان به دلیل بر عهده داشتن بار تکفل خانواده عمدتاً تمایل به خرید بیمه‌نامه‌های با نرخ ۱۴ و ۱۸ درصدی دارند. همچنین، در مردان با افزایش سن، درآمد و سابقه، تقاضای بیمه‌های با نرخ ۱۴ و ۱۸ درصد افزایش می‌یابد، اما چنین روندهایی برای زنان وجود ندارد. طبق نتایج به‌دست آمده متغیرهای میزان سابقه‌کار و پس از آن جنسیت در انتخاب نوع بیمه‌نامه تعیین‌کننده هستند، به‌گونه‌ای که طبق پیش‌بینی مدل افراد با سابقه کار کمتر از ۴/۵ سال متقاضیان قطعی بیمه‌نامه ۱۲ درصدی شناخته

شده‌اند. با توجه به نتایج و انگیزه پایین زنان و جوانان برای انتخاب بیمه‌های با خدمات گسترده‌تر، سازمان تأمین اجتماعی می‌تواند ازطریق ارائه مشوق‌ها یا خدمات کوتاه‌مدت، جذابیت این نوع بیمه‌ها با خدمات گسترده‌تر را در بین این گروه خاص افزایش دهد.

استناد: قاسمی، یاسین؛ خندان، عباس؛ اکبرپور روشن، نرگس. (۱۴۰۱). پیش‌بینی خرید بیمه‌نامه حرف و مشاغل آزاد سازمان تامین اجتماعی با استفاده از الگوریتم طبقه‌بندی درخت تصمیم و جنگل تصادفی؛ *تحقیقات مدل‌سازی اقتصادی*، ۱۳ (۴۷)، ۱۶۵-۱۱۵.

DOI: 00000000000000000000

ناشر: دانشگاه خوارزمی.

۱. مقدمه

پوشش ریسک کاهش درآمد در دوران سالمندی از اساسی‌ترین نیازهای انسان است که به‌لحاظ اقتصادی، سیاسی و اجتماعی بسیار اهمیت داشته و همواره مورد توجه دولت‌ها بوده است. سازمان تامین اجتماعی ایران بزرگ‌ترین سازمان فعال در حوزه بیمه‌های اجتماعی و بازنشستگی کشور محسوب می‌شود و طبق آخرین سالنامه منتشر شده در سال ۱۳۹۹ جمعیتی حدود ۴۴ میلیون و پانصد هزار نفر را تحت پوشش دارد. این سازمان یک بیمه‌گر اجتماعی است که ماموریت اصلی آن پوشش کارگران مزد و حقوق‌بگیر (به‌صورت اجباری) و صاحبان حرف و مشاغل آزاد (به‌صورت اختیاری) در برابر ریسک‌های مختلفی از جمله ریسک فوت، ازکارافتادگی و کاهش درآمد در دوران سالمندی است. بازنشستگان و ازکارافتادگان تحت پوشش تا زمان حیات خود از مستمری برخوردار می‌شوند، و در صورت فوت بیمه‌شده اصلی، چه در دوران اشتغال و چه بازنشستگی، مستمری به بازماندگان مشمول آن‌ها پرداخت می‌شود. این وجه در واقع ریسک فوت بیمه شده اصلی را پوشش می‌دهد.

بیمه حرف و مشاغل آزاد تامین اجتماعی که در گروه بیمه‌های خویش‌فرما قرار می‌گیرد، شامل افرادی می‌شود که حرفه و شغل آزاد دارند (به‌تنهایی) یا تحت عنوان کارفرما (با داشتن کارگر) در حال فعالیت هستند که می‌توانند پس از بازنشستگی از مزایای مستمری آن استفاده نمایند. طبق آخرین سالنامه منتشرشده، بیمه‌شدگان اصلی سازمان حدود ۱۴ میلیون و ۵۰۰ هزار نفر هستند که از این تعداد حدود ۱ میلیون و ۱۰۰ هزار نفر (۸ درصد) بیمه‌شده حرف و مشاغل آزاد هستند. این آمار نشان می‌دهد استقبال شاغلین حرف و مشاغل آزاد از این نوع بیمه‌نامه‌ها کم است و سازمان نیازمند ایجاد سازوکارهایی برای جذب حداکثری افراد به این نوع بیمه است. مطالعات بسیار کمی به بررسی بیمه حرف و مشاغل آزاد سازمان تأمین اجتماعی و انگیزه‌های افراد از تقاضای آن پرداخته‌اند، اما همین مطالعات اندک هم به خوبی نشان می‌دهند که هنوز بسیاری از مردم متقاضی برخورداری از این بیمه‌نامه‌ها نیستند و در بسیاری از موارد قراردادهای این بیمه‌نامه‌ها پس از یک مدت بیمه‌پردازی، قطع می‌شود.

از این جهت، مطالعه و بررسی ویژگی‌های افرادی که متقاضی این نوع بیمه‌نامه‌ها هستند بسیار اهمیت دارد، چراکه می‌تواند در رفع مشکل ضریب نفوذ پایین این بیمه‌نامه‌های بازنشستگی اختیاری کمک‌کننده باشد. البته دستیابی به این هدف و بررسی ویژگی‌های افراد تحت پوشش این نوع بیمه‌ها با پیچیدگی‌هایی همراه است، اما ابزارهای جدید مانند یادگیری ماشین[1] به‌راحتی و بدون مداخله مستقیم انسانی در روند کار، می‌توانند با حجم زیادی از داده‌های چندبعدی و چندمتغیره کار کنند و الگوها و روندهای موجود را شناسایی و در نتیجه باعث افزایش دقت و سرعت انجام کار شوند. این مقاله نخستین مطالعه‌ای است که خرید بیمه‌نامه‌های حرف و مشاغل آزاد تأمین اجتماعی را با استفاده از شیوه‌های یادگیری ماشینی و داده‌کاوی مورد بررسی قرار می‌دهد.

در مقاله حاضر به دو دسته از انواع بیمه‌نامه‌های حرف و مشاغل آزاد شامل پوشش بازنشستگی با حق‌بیمه ۱۲ درصد[2] و پوشش بازنشستگی و سایر ریسک‌ها با حق‌بیمه ۱۴ و ۱۸ درصدی (پوشش ازکارافتادگی علاوه بر بازنشستگی با حق‌بیمه ۱۴ درصد و پوشش ریسک ازکارافتادگی و مستمری بازماندگان علاوه بر بازنشستگی با حق‌بیمه ۱۸ درصد) پرداخته خواهد[3] شد و از آنجایی که تعداد افراد با نرخ ۱۴ درصد نسبت به سایر نرخ‌ها پایین تر است، آن را با ۱۸ درصد در یک دسته قرار دادیم. در ادامه تلاش می‌شود با تجزیه و تحلیل الگوهای خرید پوشش بیمه‌ای توسط افراد مختلف و بررسی ویژگی‌های افرادی که تحت پوشش این نوع بیمه‌نامه‌ها هستند به روشن شدن چرایی پایین بودن ضریب نفوذ این بیمه‌نامه‌ها کمک کرده و دانش جدیدی که از بررسی خصوصیات مشتریان به دست می‌آید به‌عنوان توصیه سیاستی در جهت بهبود فعالیت‌های سازمان برای جذب بیمه‌شده پیشنهاد شود. نتیجه این تجزیه و تحلیل می‌تواند باعث بهبود مدیریت روابط سازمان با مشتریان و همچنین افزایش جمعیت تحت پوشش سازمان شود. در این راستا، با پیش پردازش مناسب داده‌های

1. Machine Learning

۲. علاوه بر خدمات بازنشستگی، ریسک فوت در زمان بازنشستگی را نیز پوشش می‌دهد.

۳. علاوه بر این نرخ، ۲ درصد هم سهم دولت است که باید به سازمان پرداخت کند.

بیمه‌شدگان حرف و مشاغل آزاد که شامل اطلاعات شخصی آن‌ها مانند سن، جنسیت، میزان درآمد و سابقه کار است مدلی ساخته خواهد شد که به پیش‌بینی تقاضای بیمه‌نامه مناسب هر مشتری با توجه به ویژگی‌های فردی آن‌ها کمک می‌کند (متقاضیان بیمه‌نامه با خدمات مستمری بازنشستگی یا افرادی که علاوه بر بازنشستگی، متقاضی خدمات ازکارافتادگی و بازماندگان خواهند بود). طبقه‌بندی مشتریان در نهایت به مدیریت بهتر مستمری‌ها و هزینه‌ها منجر می‌شود.

این مقاله در پنج بخش سامان‌دهی شده است. در ادامه و در بخش دوم به ادبیات موضوع و پیشینه پژوهش پرداخته خواهد شد. بخش سوم مقاله به روش‌شناسی و توصیف داده‌ها اختصاص دارد. در این بخش وظایف و مراحل روش کریپس[1] که یکی از معروف‌ترین روش‌های داده‌کاوی و مورد استفاده در این مقاله است بیان خواهد شد و مجموعه‌ای از آمار توصیفی از نحوه توزیع مشخصه‌های افراد و مدلسازی شامل دو الگوریتم درخت تصمیم و جنگل تصادفی ارائه می‌شود. در بخش چهارم یافته‌های پژوهش به همراه ارزیابی نتایج مدلسازی و پیش‌بینی‌ها مطرح می‌شود. در پایان و در بخش پنجم جمع‌بندی و نتیجه‌گیری مقاله به همراه پیشنهادهایی کاربردی برای سازمان تأمین اجتماعی ارائه خواهد شد.

۲. ادبیات موضوع

سازمان تامین اجتماعی یک سازمان بیمه‌گر اجتماعی است که ماموریت اصلی آن پوشش کارگران مزد و حقوق‌بگیر به صورت اجباری و صاحبان حرف و مشاغل آزاد به صورت اختیاری است. این سازمان یک سازمان عمومی غیردولتی است که عمده منابع مالی آن از محل حق بیمه‌ها (با مشارکت بیمه‌شده، کارفرما و دولت) تامین می‌شود و ارائه خدمات بلندمدت (مستمری‌های بازنشستگی، ازکارافتادگی، بازماندگان)، کوتاه‌مدت (حمایت در برابر حوادث، بیماری‌ها و بارداری، غرامت دستمزد ایام بیماری، غرامت دستمزد ایام

1. Cross-Industry Standard Process for Data Mining (CRIPS-MD)

بارداری، پرداخت هزینه وسایل کمک پزشکی، کمک هزینه ازدواج، کمک هزینه کفن و دفن) و همچنین خدمات درمان و سلامت و مقرری بیمه بیکاری را بر عهده دارد. بر این اساس می‌توان سازمان تامین اجتماعی را در دسته سیستم‌هایی قرار داد که اغلب ریسک‌های مهم ذکرشده در استانداردهای سازمان بین‌المللی کار (ILO) را پوشش می‌دهد (برای مطالعه بیشتر ر.ک. مقاوله‌نامه ۱۰۲، با عنوان حداقل استانداردهای تأمین اجتماعی[1]).

بیمه حرف و مشاغل آزاد سازمان تأمین اجتماعی اختیاری است و شامل افرادی می‌شود که غیر مزد و حقوق‌بگیر (تحت عنوان کارفرما یا بدون کارفرما) هستند. این افراد اختیار دارند که انواع بیمه‌نامه‌ها در سه سطح حق بیمه ۱۲ درصد (خدمات مستمری بازنشستگی)، حق بیمه ۱۴ درصد (خدمات مستمری بازنشستگی و ازکارافتادگی) و حق بیمه ۱۸ درصد (خدمات مستمری بازنشستگی، ازکارافتادگی و فوت یا بازماندگان) را انتخاب و در هر مقطعی قرارداد خود را فسخ کنند. طبق آخرین آمار منتشر شده در سال ۱۳۹۹ حدود ۱ میلیون و ۱۰۰ هزار نفر از بیمه‌شدگان سازمان را بیمه‌شدگان حرف و مشاغل آزاد تشکیل می‌دهند که تنها ۸ درصد از بیمه‌شدگان اصلی سازمان هستند. این در حالی است که در همین سال طبق گزارش مرکز آمار ایران، خوداشتغالی (کارکن مستقل و کارفرما) سهمی ۴۷ درصدی از بازار کار داشته‌اند.

برای رفع مسئله پایین بودن ضریب نفوذ و خریداری بیمه‌نامه‌های حرف و مشاغل آزاد لازم است تا سازمان تأمین اجتماعی از طریق سازماندهی فرآیندهای بازاریابی و ارائه خدمات به مشتریان، باعث ایجاد یک سیستم تعاملی بین مشتریان و سازمان شود. از این طریق می‌توان ارتباط نزدیک‌تر و دقیق‌تری با خواسته‌های مشتریان ایجاد و آن‌ها را به طور کامل با خدمات سازمان آشنا کرد. این مفاهیم تحت عنوان مدیریت ارتباط با مشتری[2] شناخته می‌شود.

۲-۱. مبانی نظری

1. Social Security (Minimum Standards) Convention
2. Customer Relationship Management

مدیریت ارتباط با مشتری ترکیبی از افراد، فرآیندها و فناوری است که به دنبال درک مشتریان و ایجاد یک رویکرد یکپارچه جهت مدیریت روابط با تمرکز بر حفظ مشتری و توسعه روابط می‌باشد. مدیریت ارتباط با مشتری از پیشرفت‌های فناوری اطلاعات و تغییرات سازمانی در فرآیندهای مشتری محور تکامل یافته و شرکت‌هایی که مدیریت ارتباط با مشتری را با موفقیت پیاده‌سازی کنند می‌توانند از وفاداری مشتری و سودآوری بلندمدت بهره‌مند شوند (چن و پوپُویچ[1]، ۲۰۰۳).

مدیریت ارتباط با مشتری در بازاریابی در چهار بعد اهمیت بیشتری یافته که شامل: شناسایی مشتری[2]، جذب مشتری[3]، نگهداری مشتری[4] و توسعه مشتری[5] می‌باشد (حسینی و همکاران، ۲۰۱۰). این چهار بعد را می‌توان به صورت یک چرخه بسته از سیستم مدیریت مشتری در نظر گرفت که هدف مشترک آن‌ها ایجاد درک عمیق‌تر از مشتریان برای به حداکثر رساندن ارزش مشتری برای سازمان در درازمدت می‌باشد (ان‌گای[6] و همکاران، ۲۰۰۹). بُعد اول شناسایی مشتری است که شامل هدف قرار دادن جمعیتی است که به احتمال زیاد مشتری می‌شوند یا بیشترین سود را برای شرکت دارند. علاوه بر این شامل تجزیه و تحلیل مشتریانی است که در رقابت از دست می‌روند و اینکه چگونه می‌توان آن‌ها را به دست آورد (کراکلائور[7] و همکاران، ۲۰۰۴). پس از شناسایی مشتریان، سازمان می‌تواند منابع خود را به سمت جذب مشتریان هدف هدایت کند، یکی از عناصر جذب مشتری بازاریابی مستقیم است. مرحله سوم، نگهداری مشتری نیز از اهمیت بالایی برخوردار است و شامل عناصر بازاریابی تک به تک[8]، برنامه‌های وفاداری و مدیریت شکایت می‌باشد. مرحله چهارم توسعه مشتری نام دارد که به معنای افزایش با ثبات و مداوم تراکنش‌های مالی، ارزش

1. Chen & Popovich
2. Customer identification
3. Customer attraction
4. Customer retention
5. Customer development
6. Ngai
7. Kracklauer

۸. به کمپین‌های بازاریابی شخصی اطلاق می‌شود که با تجزیه و تحلیل، شناسایی و پیش‌بینی تغییرات در رفتارهای مشتری پشتیبانی می‌شود.

تراکنش‌ها و سودآوری مشتری است، و شامل تجزیه و تحلیل ارزش طول عمر مشتری[۱]، توسعه فروش و تجزیه و تحلیل سبد خرید می‌باشد (ان‌گای و همکاران، ۲۰۰۹).

این مراحل چهارگانه مدیریت ارتباط با مشتری باعث ایجاد ارزش برای مشتریان و سازمان تأمین اجتماعی شده و به حفظ مشتریان فعلی و جذب مشتریان جدید برای این سازمان منتج خواهد شد.[۲] البته مدیریت ارتباط با مشتری وابسته به تحلیل آمار و اطلاعات مشتریان است و به این منظور داده‌کاوی[۳] اطلاعات مشتریان و خریداران بیمه‌نامه‌ها لازم است. داده‌کاوی با کشف روندهای عمومی، دانش مستقیم و غیر مستقیم زیادی برای مدیران و تصمیم‌گیرندگان به همراه دارد و نتیجه آن اتخاذ تصمیم‌های درستی است که می‌تواند باعث افزایش کارایی و عملکرد سازمان شود.

داده‌کاوی در واقع فرآیندی با استفاده از ابزارهای توصیفی[۴] و تحلیلی برای جستجوی مدل‌ها، الگوها و رابطه بین داده‌های تاریخی و ایجاد مدل‌هایی است که کارکرد آن‌ها پیش-بینی[۵] است. بنابراین داده‌کاوی را می‌توان این‌گونه توصیف کرد: با توجه به هدف تعیین شده یک سازمان، روشی موثر و پیشرفته در کاوش و تجزیه و تحلیل و مدیریت حجم زیادی از داده‌های سازمانی و آشکارسازی قوانین ناشناخته و مدلسازی آن‌ها برای پیش‌بینی موارد آینده است (چن و هو[۶]، ۲۰۰۵). داده‌کاوی که مبتنی بر اصول آمار است، فرآیند کاوش، تجزیه و تحلیل مقادیر زیادی از داده‌ها را برعهده دارد تا الگوهای موجود در آن‌ها را کشف کند. از الگوریتم‌ها برای یافتن روابط و الگوها در داده‌ها استفاده می‌شود و سپس این اطلاعات در مورد الگوها برای پیش‌بینی[۷] و تخمین[۸] به کار می‌رود. به طور کلی، هدف

---

۱. تحلیل ارزش طول عمری مشتری به عنوان پیش‌بینی کل درآمد خالصی است که یک شرکت می‌تواند از مشتری انتظار داشته باشد.

۲. البته باید به این نکته توجه کرد که منظور از مشتری در اینجا بیمه‌شده است که ممکن است الزاماً ویژگی‌ها و مؤلفه‌های مشتری به معنای مرسوم را نداشته باشد.

3. Data mining
4. Descriptive
5. Predictive
6. Chen & Hu
7. Predictive
8. Forecasting

داده‌کاوی استخراج داده‌ها از یک مجموعه داده بزرگ‌تر برای اهداف طبقه‌بندی[1] یا پیش-بینی است. از نظر جهت‌گیری فرآیند، فعالیت‌های داده‌کاوی به سه دسته کلی تقسیم می‌شوند (ریجلسکی[2] و همکاران، ۲۰۰۲): کشف[3] به معنی فرآیند جستجو در پایگاه داده برای یافتن الگوهای پنهان بدون یک ایده یا فرضیه از پیش تعیین شده در مورد اینکه الگوها ممکن است چه باشند؛ مدلسازی پیش‌بینی کننده[4] که به فرآیند گرفتن الگوهای کشف شده از پایگاه داده و استفاده از آن‌ها برای پیش‌بینی آینده گفته می‌شود؛ و بررسی صحت مدل[5] که فرآیند بکارگیری الگوهای استخراج شده برای یافتن الگوهای داده غیرمعمول و بی شباهت است. داده‌کاوی کاربرد زیادی در تجزیه و تحلیل اطلاعات مربوط به مشتریان دارد و می‌تواند در تمامی مراحل ارتباط با مشتری از جمله شناسایی مشتری، جذب مشتری، نگهداری و توسعه مشتری استفاده شود.

۲-۲. پیشینه پژوهش

مطالعات بسیار کمی به بررسی بیمه حرف و مشاغل آزاد سازمان تأمین اجتماعی و انگیزه‌های افراد از پوشش اختیاری تحت این بیمه‌نامه‌ها پرداخته‌اند. عبدی (۱۳۸۵) به بررسی مشکلات بیمه‌شدگان حرف و مشاغل آزاد و اختیاری سازمان تأمین اجتماعی پرداخته است. در این مطالعه یک نمونه ۴۰۰ نفری از بیمه‌شدگان حرف و مشاغل آزاد و اختیاری مراجعه کننده به ۱۱ شعبه سازمان تأمین اجتماعی تهران و شهرری (از مجموع ۷۱۲۵۶ نفر بیمه‌شده اختیاری و ۴۷۸۴۴ نفر حرف و مشاغل آزاد تحت پوشش سازمان در ۳۰ شعبه) در پایان سال ۱۳۸۳ به‌صورت پرسشنامه‌ای مورد پرسش قرار گرفته‌اند و متغیرهای فردی (سن، جنسیت، تحصیلات، وضعیت تاهل، وضع درآمد و...) و متغیرهای بیمه‌ای آنان (سابقه بیمه، سن در زمان تقاضا، انواع پوشش انتخابی، مبلغ بیمه و...) بررسی شده است. نتایج به‌دست آمده از

1. Classification
2. Rygielski
3. Discovery
4. Predictive Modeling
5. Forensic Analysis

این مطالعه میدانی نشان می‌دهد که حدود ۹۰ درصد بیمه‌شدگان حرف و مشاغل آزاد را مردان تشکیل می‌دهند که فقط حدود ۶ درصد مجرد هستند. متوسط سن بیمه شدگان در مقطع پژوهش حدود ۴۵ سال بوده و افراد دارای تحصیلات بالاتر در مقایسه با جامعه آماری سهم بیشتری در نمونه داشته‌اند. سطوح تحصیلی زنان بیمه شده در مقایسه با سطوح تحصیلی مردان بیمه‌شده بیشتر بوده است. همچنین متوسط سطح درآمد افراد بیمه شده برابر با ۳۸۴ هزارتومان در ماه بوده است. اکثر بیمه شدگان اختیاری بوده که بر حسب آمار موجود از بیمه‌های اختیاری تهران استنباط شده (حدود ۶۰ درصد) و متوسط سابقه بیمه‌ای آن‌ها قبل از اولین تقاضا به طور متوسط حدود ۷/۲ سال بوده است. یافته‌های این پژوهش همچنین نشان می‌دهند که اولاً از مجموعه این بیمه‌شدگان حدود ۱۷ درصد از آنان سابقه قطع قرارداد خود را داشته‌اند. البته بیشتر فقط یک بار قطع قرارداد داشته‌اند اما پیش‌بینی می‌شود که حدود ۴۰ درصد به احتمال زیاد در دوسال آینده بیکار شوند و ۳۵ درصد از آن‌ها هم به احتمال زیاد قراردادشان در آینده به دلیل مشکلات مالی به‌طور دائم یا موقتی قطع خواهد شد. دوم این‌که آگاهی آنان از مسائل بیمه‌ای به شدت پایین بوده به‌گونه‌ای که حتی عده‌ای نوع پوشش بیمه‌نامه خود را نمی‌دانسته‌اند. و سوم این‌که انگیزه‌ها و اهداف اصلی آنان از بیمه به ترتیب بازنشستگی، درمان، ازکارافتادگی و فوت بوده است.

ملامحمدی و مستوفی (۱۳۹۳) به بررسی عوامل موثر بر موفقیت سازمان تامین اجتماعی در پوشش و برقراری مستمری بازنشستگی پرداخته‌اند. این مطالعه با استفاده از داده‌های یک نمونه ۱۹۱ نفری که بر اساس فرمول کوکران به‌طور تصادفی به روش پیمایشی و پرسشنامه از مجموعه ۸۱۲۳ نفری در بازه زمانی آبان ۱۳۹۱ تا شهریور ۱۳۹۳ از یک شعبه سازمان تامین اجتماعی در قم جمع‌آوری شده نشان می‌دهد که ۸۵/۷ درصد از نمونه بررسی شده بیش از ۴۱ سال سابقهٔ کار داشته و ۹۵/۳ درصد از نمونه‌های تحقیق را مردان متاهل تشکیل می‌دهند. همین‌طور ۸۰/۲ درصد از نمونه‌ها بیش از دو نفر را تحت کفالت خود دارند. ۷۲/۳ درصد سن بین ۶۱ تا ۶۵ سال دارند و بر همین اساس ۶۷ درصد از آنان سواد ابتدایی به پایین را

دارند. همچنین ۶۷ درصد نمونه سابقه بیمه کمتر از ۱۵ سال دارند. طبق نتایج به‌دست آمده، عواملی از جمله برنامه‌های حمایت از کارگران، رفتار بیمه‌ای کارفرما، آموزش و اطلاع رسانی قوانین سازمان تأمین اجتماعی و در نهایت توان اقتصادی بیمه‌شده و کارفرما از مهم ترین عوامل مؤثر در برقراری مستمری بازنشستگی بوده‌اند.

نجفی (۱۳۹۸) به بررسی و پیش‌بینی وفاداری و رویگردانی بیمه‌شدگان خویش‌فرمای سازمان تامین اجتماعی پرداخته و هدف خود را شناسایی میزان رویگردانی (قطع رابطه) بیمه‌شدگان و همچنین یافتن مدلی بر مبنای الگوریتم‌های هوش مصنوعی[1] قرار می‌دهد که میزان رویگردانی بیمه‌شدگان در سال‌های آینده را به درستی تخمین بزند. در این پژوهش اطلاعات و شاخص‌های مهم حدود ۲۱۴۰۷ نفر در قالب ۲۷ ویژگی از بانک‌های اطلاعاتی سازمان تامین اجتماعی استخراج گردیده است. سپس با استفاده از الگوریتم ژنتیک رتبه‌بندی نامغلوب[2]، تعداد ۷ ویژگی مهم که حداقل خطای طبقه‌بندی را داشته‌اند انتخاب شدند (ویژگی‌های جنسیت، نوع بیمه، تعداد روز سابقه ناشی از قرارداد، سال تولد بیمه‌شده، تاریخ قرارداد، تاریخ تولد بزرگترین فرزند، سال شروع قرارداد). در این مطالعه در مرحله انتخاب بهترین طبقه‌بندی کننده برای انجام پیش‌بینی‌های مورد نیاز نیز از سه الگوریتم شبکه عصبی چندلایه[3]، الگوریتم ماشین بردار پشتیبان[4] و الگوریتم کی-میانگین[5] استفاده شده که نهایتاً شبکه عصبی چندلایه، بهترین دقت طبقه‌بندی را با مقدار ۹۶/۸ درصد به دست آورد. سپس برای شبکه عصبی مذکور با استفاده از ۷ ویژگی مربوط به داده‌های سال‌های۱۳۶۷ تا ۱۳۹۵ آموزش انجام شد. شبکه عصبی آموزش دیده برای پیش‌بینی وفاداری و رویگردانی مشتریان سال‌های ۱۳۹۶ و ۱۳۹۷ که به تعداد ۸۳۶۴ رکورد می‌باشد، مورد استفاده قرار گرفت. با

1. Artificial Intelligence
2. Nsga-Ii
3. Multilayer Perceptron
4. Support Vector Machines
5. K-Nearest Neighbor Algorithm

توجه به نتایج به‌دست آمده، حدود ۲۷ درصد از بیمه‌شدگان سال‌های ۱۳۹۶ و ۱۳۹۷ در کلاس رویگردان دسته‌بندی می‌شوند.

بروفر و همکاران (۱۳۹۵) در مقاله‌ای با عنوان «شناسایی الگوی رفتاری مشتریان در بیمه عمر و تشکیل سرمایه با استفاده از داده‌کاوی» به ارائه یک مدل مناسب و کارآمد جهت بخش‌بندی مشتریان بر اساس برخی از مهم‌ترین ویژگی‌های مالی و جمعیت شناختی در قالب عوامل موثر بر شاخص‌های ارزش دوره عمر مشتری (آر.اف.ام)[1] پرداختند. پس از تعیین مقادیر شاخص‌های مدل آر.اف.ام شامل تازگی مبادله[2]، تعداد دفعات مبادله[3] و ارزش پولی[4] مبادله در ۱۸۰ هزار مشتری و وزن‌دهی آن‌ها با استفاده از فرآیند تحلیل سلسله مراتبی، تعداد خوشه بهینه بر اساس شاخص سیلوئت[5] و نرخ تاثیر شاخص‌های آر.اف.ام با استفاده از الگوریتم دو مرحله‌ای[6] تعیین شد و در مرحله بعد به خوشه‌بندی مشتریان با استفاده از روش کی-میانگین پرداخته شده است. آمار مورد مطالعه شامل اطلاعات خرید بیمه‌نامه عمر و تشکیل سرمایه مشتریان شرکت بیمه سامان در سال‌های ۱۳۹۰ تا ۱۳۹۳ می‌باشد که فیلدهای مختلفی از جمله مشخصات بیمه‌شده (تاریخ تولد، جنسیت، وضعیت تاهل، شغل، درآمد، میزان تحصیلات)، مشخصات بیمه‌نامه (حق‌بیمه سالیانه، سرمایه فوت، پوشش‌های تکمیلی، میزان خطرپذیری، روش پرداخت) و پوشش‌های تکمیلی (فوت بر اثر سانحه، از کار افتادگی، امراض صعب العلاج) را شامل می‌شود. در این تحقیق از روش‌شناسی کریپس استفاده شده است. پس از بکارگیری تکنیک‌های مختلف جهت آماده‌سازی داده‌ها (حذف داده‌های پرت و گم‌شده) در نهایت ۱۷۱۱۹۲ رکورد جهت خوشه‌بندی مورد استفاده قرار گرفت. پس از جلسات متعدد با خبرگان ۵ ویژگی موثر که اهمیت زیادی در تحقیق دارند

---

۱. ارزش دوره عمر مشتری (Recency Frequency Monetary) ارزشی است که مشتری در طول عمر خود برای سازمان ایجاد می‌کند. این مفهوم علاوه بر ارزش فعلی مشتری به ارزش بالقوه و ارزش آتی مشتری نیز اشاره دارد.

2. Recency
3. Frequency
4. Monetary
5. Silhouette
6. Two-step

انتخاب شدند که شامل: نسبت سرمایه اولیه به حق‌بیمه اولیه، نسبت سرمایه فوت حادثه به حق‌بیمه اولیه، نسبت سرمایه از کار افتادگی حادثه به حق‌بیمه اولیه، نسبت سرمایه امراض به حق‌بیمه اولیه و ماه شروع بیمه‌نامه می‌باشد. این مطالعه سپس با مقایسه دو الگوریتم خوشه‌بندی کی-میانگین و دو مرحله‌ای نتیجه می‌گیرد که الگوریتم کی-میانگین از نظر معیار شاخص سیلوئت با مقدار ۰/۷۰ برتری دارد که در نهایت این الگوریتم به عنوان بهترین الگوریتم برای خوشه‌بندی مشتریان انتخاب شد.

پرماه و همکاران (۱۳۹۹) در مقاله‌ای با عنوان؛ متغیرهای کلان اقتصادی و تقاضای بیمه‌های خویش فرمایی در سازمان تامین اجتماعی، به تاثیر متغیرهای کلان اقتصادی بر تقاضای بیمه‌های خویش فرما در ۳۱ استان کشور پرداخته‌اند. هدف خاص این مطالعه بررسی تاثیر تولید ناخالص داخلی، نرخ تورم، نرخ بیکاری و تحصیلات بر میزان تقاضای بیمه‌های اختیاری و حرف و مشاغل آزاد تامین اجتماعی بوده است. بدین منظور داده‌های مربوط به متغیرهای تحقیق در ۳۱ استان کشور در دوره زمانی ۱۳۹۵-۱۳۸۰ جمع‌آوری و از الگوی پویای پانلی به روش گشتاورهای تعمیم یافته(GMM) برای بررسی موضوع استفاده شده است. نتایج تحقیق بیانگر آن است که تاثیر نرخ تورم بر تقاضای بیمه‌های خویش‌فرمایی به صورت منفی و معنادار و متغیرهای نرخ بیکاری، تحصیلات و تولید ناخالص داخلی مثبت و معنادار بوده است.

عبدی و همکاران (۲۰۱۷) یک رویکرد داده‌کاوی سه مرحله‌ای را برای شناسایی مشتریان وفادار و برنامه‌ریزی فروش بیمه بر اساس ویژگی مشتریان در ایران را توسعه داده‌اند. مجموعه داده‌ای که در این مطالعه استفاده شده شامل ۴۷۸ مشتری بیمه می‌باشد. در مجموع ۷۰ مشتری یک پوشش بیمه خاص را خریداری و ۴۰۸ مشتری این نوع پوشش بیمه‌ای را خریداری نمی‌کنند که متغیرهای مورد استفاده در این مدل شامل سن، تحصیلات، شغل، سابقه کار، درآمد، تعداد اعضای خانواده و ... می‌باشد. این مطالعه از الگوریتم کی-نزدیک‌ترین همسایه[1] برای خوشه‌بندی[2] مشتریان بهره برده است. هدف این مطالعه پیش‌بینی

1. K-Nearest Neighbor
2. Clustering

این بوده که آیا یک پوشش بیمه‌ای خاص توسط مشتریان خریداری می‌شود یا خیر و همچنین در این مقاله تلاش شده تا از تکنیک‌های داده‌کاوی برای بهبود مدیریت ارتباط با مشتری و شناسایی مشتریان ارزشمند استفاده شود.

دامغانی و همکاران (۲۰۱۹) با استفاده از تکنیک‌های داده‌کاوی مدلی ترکیبی از طبقه‌بندی و خوشه‌بندی ساخته‌اند که بیمه‌نامه مناسب هر مشتری را بر اساس سابقه پوشش بیمه‌ای آن فرد پیشنهاد می‌دهد. در این مقاله یک مطالعه موردی واقعی شامل ۴۷۶ مشتری یک شرکت بیمه در ایران که شامل ۲۵ متغیر می‌باشد، مورد تجزیه و تحلیل قرار گرفته است که از این ۲۵ متغیر ۱۵ متغیر مربوط به ویژگی ثبتی مشتریان و ۱۰ متغیر مرتبط با پوشش بیمه‌ای مشتریان می‌باشد. متغیرها شامل سن، تحصیلات، میزان درآمد، تعداد اعضای خانواده و ... می‌باشد. در این مطالعه در نهایت الگوریتم خوشه‌بندی کی-میانگین انتخاب شد که طبق آن مشتریان به چهار خوشه تقسیم شده‌اند. برای طبقه‌بندی نیز در این مطالعه از الگوریتم کی- نزدیک-ترین همسایه استفاده شد که در آن معیار دقت برابر با ۸۴/۶۵ ، پوشش برابر با ۷۷/۶۰ و صحت نیز برابر با ۸۰/۸۲ رسید.

رحمان[1] و همکاران (۲۰۱۷) با تجزیه و تحلیل داده‌های شرکت بیمه زندگی پرایم اسلامی بنگلادش[2] در بازه زمانی ۲۰۱۱ تا ۲۰۱۴ و با استفاده از اطلاعات ۲۸۲۲۸۲ بیمه‌گذار به نحوه واکنش مشتریان به بیمه‌نامه‌های ارائه شده پرداخته‌اند. در این مطالعه از ۱۰ ویژگی (شرایط خط‌مشی[3]، سن، جنسیت، شغل، شهری یا روستایی، وضعیت تاهل، مبلغ تضمین شده[4]، بخش محل سکونت، حالت پرداخت حق بیمه سالانه یا ماهانه و منظم بودن در پرداخت) استفاده شد. هدف این مطالعه طبقه‌بندی مشتریان بر اساس ویژگی‌های آن‌ها بوده تا بتوانند برچسب کلاس را برای مشتریان آینده پیش‌بینی کنند. این هدف بر مبنای منظم بودن یا نبودن مشتری

1. Rahman
2. Prime Islamic Life Insurance Company Ltd., of Bangladesh
۳. خط مشی (Policy term) به بازه زمانی اشاره دارد که دارنده بیمه نامه از طریق آن حق بیمه خود را پرداخت می‌کند.
۴. مبلغ تضمین‌شده (Sum-Assured) مبلغ از پیش تعیین شده‌ای که بیمه‌گر متعهد به پرداخت آن در صورت فوت بیمه گذار می‌شود.

با استفاده از یک ویژگی به نام «نظم و قاعده»[1] مشخص می‌شود. از تکنیک‌های مختلف ارزیابی ویژگی برای تعیین موثرترین ویژگی‌های مورد نیاز طبقه‌بندی استفاده شده که اولین تکنیک «ارزیابی ویژگی کسب اطلاعات»[2] و دیگری «نسبت افزایش تکنیک‌های ارزیابی ویژگی»[3] می‌باشد. در نهایت از میان تمام ویژگی‌های انتخاب شده از دو تکنیک، ویژگی حالت پرداخت حق‌بیمه به عنوان موثرترین ویژگی شناخته شده است.

مائو[4] و همکاران (۲۰۱۸) به بررسی ارزش داده‌های غنی شده مشتریان برای مدیریت تحلیلی ارتباط با مشتری در بخش بیمه خودرو و خانه پرداخته‌اند. برای این منظور استعلام‌های آنلاین یک بیمه‌گر سوئیسی که دارای محصولات بیمه زندگی و غیرزندگی است را از سال ۲۰۱۲ تا ۲۰۱۵ مورد بررسی قرار داده‌اند. این داده‌ها شامل متغیرهای کمّی شخصی (روز تولد، جنسیت و غیره) و متغیرهای عمومی بیمه‌نامه (شناسه بیمه‌نامه، نسخه بیمه‌نامه، تاریخ آغاز و تاریخ انقضا) و قیمت‌گذاری متغیرهای کمکی مربوط به محصول بیمه خاص (ارزش خانوار بیمه‌شده برای بیمه‌نامه خانگی) بوده است. از مجموع ۲/۵ تا ۳ میلیون مشاهده تقریباً ۲۷۵ هزار مورد قیمت بیمه اتومبیل و ۹۰ هزار مورد قیمت بیمه خانه به عنوان نمونه انتخاب شده است. برای محصول بیمه اتومبیل متغیرهای کمّی مانند تاریخ تولد، کدپستی محل اقامت، جنسیت، تاریخ صدور گواهی‌نامه رانندگی وسیله نقلیه نیز ثبت شده است. در ادامه برای پیش‌بینی خرید و طبقه‌بندی مشتریان از مدل پیش‌بینی جنگل تصادفی استفاده شده که بر اساس دو معیار دقت[5] و امتیاز-F[6] دقت پیش‌بینی آزمون مربوطه بر روی داده‌های تست به ۰/۸۴۴ رسیده است.

1. Regularity
2. Information Gain Attribute Evaluation
3. Gain Ratio Attribute Evaluation Techniques
4. Mau
5. Accuracy
6. F-Score

عبدالرحمن[1] و همکاران (۲۰۲۱) در مطالعه‌ای با استفاده از خوشه‌بندی k-modes و طبقه‌بندی درخت تصمیم به تقسیم‌بندی[2] مشتری و پروفایل سازی[3] برای بیمه عمر پرداخته‌اند. الگوریتم درخت تصمیم برای طبقه‌بندی پروفایل مشتری با مقایسه دو شاخص آنتروپی و جینی استفاده شد و همچنین خوشه‌بندی k-modes مشتریان را به سه گروه تقسیم‌بندی کرد؛ «مشتریان بالقوه با ارزش بالا»، «مشتریان کم ارزش» و «مشتریان بی علاقه». در نهایت درخت تصمیم با شاخص جینی و با اعتبار سنجی متقابل ۱۰ برابری[4] به عنوان بهترین مدل برازش با دقت ۸۱/۳ درصد تعیین شد. آتزون و همکاران (۲۰۲۲) با استفاده از الگوریتم جنگل تصادفی و رگرسیون لجستیک به پیش‌بینی مدت زمان سپری شده بیمه‌نامه‌های عمر می-پردازند. نتایج بدست آمده نشان می‌دهند که عوامل غیراقتصادی (مدت زمان مانده تا سررسید، شرکت بیمه و رویکرد بازاریابی آن) در تعیین مدت زمان سپری شده نقش مهمی دارند و در مقابل عوامل اقتصادی و مالی بی‌تأثیر بوده‌اند (البته به غیر از نرخ رشد درآمد که تأثیر زیادی داشت). این مطالعه همچنین نشان می‌دهد که مدل‌های خطی مانند رگرسیون لجستیک در مدل‌سازی تصمیمات مالی پیچیده مانند تقاضای بیمه‌نامه‌ها عملکرد مناسبی ندارند.

کونگ و همکاران (۲۰۲۲) با طراحی یک فیلتر همکارانه (CF) با استفاده از الگوریتم‌های یادگیری ماشین به پیش‌بینی رفتار خریداران بیمه‌نامه‌های زندگی در کره‌جنوبی می‌پردازد. این مقاله سپس عملکرد این الگوریتم‌ها را با مدل رگرسیون لجستیک مقایسه می‌کند و نشان می‌دهد مدل‌های مبتنی بر یادگیری ماشین و فیلترهای همکارانه (CF) مبتنی بر آن‌ها در پیش‌بینی و توصیه‌های سیاستی برای طبقه‌بندی مشتریان بسیار مناسب هستند به‌ویژه زمانی‌که مشخصه‌های قراردادی و اطلاعات خرید بیمه‌نامه نیز به مشخصه‌های خریداران افزوده شود. سیورینو و پنگ[5]

1. Abdul-Rahman
2. Segmentation
3. Profiling
4. 10-Fold Cross Validation
5. Severino & Peng

(۲۰۲۱) در مطالعه‌ای تحت عنوان الگوریتم‌های یادگیری ماشین برای پیش‌بینی تقلب در بیمه اموال با استفاده از شواهد تجربی و داده‌های دنیای واقعی، به پیش‌بینی تقلب در ادعاهای بیمه دارایی در یک شرکت بزرگ بیمه‌ای در برزیل پرداخته‌اند، که مجموعه داده‌های آن مربوط به خسارت‌های ثبت شده برای بیمه‌های مسکونی و تجاری از سال ۲۰۰۹ تا ۲۰۱۸ می‌باشد و متغیرهای آن شامل روزهای بین پایان قرارداد و ادعا[1]، روزهای بین شروع قرارداد و ادعا[2]، زمان تایید[3]، مبلغ بیمه شده، حق بیمه، تعداد اقساط، سن و اداعای قبلی می‌باشد. در این مطالعه ۹ مدل به صورت بازگشتی آزمایش شدند و میانگین نتایج پیش‌بینی‌کننده با کنترل مثبت کاذب[4] و منفی کاذب[5] مقایسه شدند. نتایج با استفاده از معیارهای مختلف ارزیابی شدند و مشخص شد که روش‌های مبتنی بر مجموعه[6] (جنگل تصادفی[7] و تقویت گرادیان[8]) و شبکه‌های عصبی عمیق[9] بهترین نتایج را به همراه داشته‌اند و عملکرد متوسط بالاتری را در مقایسه با سایر طبقه‌بندی‌کننده‌ها از جمله رگرسیون لجستیک[10] رایج نشان می‌دهند.

به‌طورکلی، مرور مطالعات مربوطه در ایران نشان می‌دهد که عمده پژوهش‌ها در خصوص شناسایی ویژگی‌ها و انگیزه‌های مشتری در صنعت بیمه مربوط به بیمه‌های بازرگانی است و تعداد مطالعاتی که به شناسایی ویژگی‌های بیمه‌شدگان (به‌نوعی مشتریان) اختیاری بیمه‌های اجتماعی از طریق الگوریتم‌های یادگیری ماشینی پرداخته‌اند بسیار کم است. از طرفی، معدود مطالعات انجام‌شده نشان می‌دهند که انعقاد انواع بیمه‌نامه بازنشستگی یا قطع قراردادهای بیمه می‌تواند به دلایل مختلفی اتفاق بیافتد که بررسی آن‌ها برای درک چرایی پایین بودن ضریب نفوذ بیمه‌های حرف و مشاغل آزاد در ایران ضروری است. اما درک

1. Days Between Contract End And Claim
2. Days Between Contract Start And Claim
3. Time Of Approval
4. False Positive
5. False Negative
6. Ensemble
7. Random Forest
8. Gradient Boosting
9. Deep Neural Networks
10. Logistic Regression

نیازها و تقاضای مشتریان و شناخت ویژگی‌های آن‌ها به سادگی امکان پذیر نیست و ناگزیر بسیاری از شرکت‌های بازنشستگی و بیمه عمر به استفاده از ابزارهای جدید مانند یادگیری ماشین روی آورده‌اند. الگوریتم‌های یادگیری ماشینی به‌راحتی و بدون مداخله مستقیم انسانی در روند کار می‌توانند با حجم زیادی از داده‌های چند بعدی و چند متغیر کار کرده و الگوها و روندهای موجود را شناسایی کنند که باعث افزایش دقت و سرعت انجام کار می‌شود. مطالعات زیادی از این روش‌های جدید در حوزه بیمه و شناخت ویژگی‌های مشتریان آن‌ها استفاده کرده‌اند و توانسته‌اند به نتایجی دست یابند که در ارائه محصولات جدید بیمه‌ای متناسب با نیاز مشتریان مفید واقع گردند.

۳. روش شناسی و توصیف داده‌ها

داده‌کاوی یکی از مهم‌ترین ابزارها در جهت شناسایی الگوهای رفتاری مشتری می‌باشد. بر همین اساس، در این مقاله از روش کریپس که یکی از معروف‌ترین روش‌های داده‌کاوی است استفاده شده است. این روش اولین بار توسط چمپان[1] و همکاران (۲۰۰۰) مطرح شد که نقطه عطفی در روش‌های داده‌کاوی و مدل‌های فرآیند بود چراکه یک نمای کلی از چرخه عمر یک پروژه داده‌کاوی شامل شش مرحله را در اختیار قرار می‌دهد (ویرث و هیپ[2]، ۲۰۰۰).

- مرحله اول، درک کسب و کار[3]، متمرکز است بر فهم اهداف و الزامات پروژه از دیدگاه تجاری، تبدیل این دانش به تعریف مسئله داده‌کاوی و سپس طرح اولیه پروژه که برای دستیابی به اهداف طراحی شده است. در واقع در این مرحله تلاش بر این است که مشکل و مسئله مورد نظر در فضای کسب و کار تامین اجتماعی تعیین و مشخص شود.
- مرحله دوم، درک داده‌ها[4]، با جمع‌آوری داده‌های اولیه شروع می‌شود و با فعالیت‌هایی به‌منظور آشنایی با داده‌ها، شناسایی مشکلات مربوط به کیفیت داده‌ها، کشف اولین بینش در داده‌ها یا شناسایی زیرمجموعه‌های جالب برای تشکیل فرضیه‌هایی به جهت پی بردن به

1. Champan et al
2. Wirth & Hipp
3. Business Understanding
4. Data Understanding

اطلاعات پنهان ادامه می‌یابد. بین درک کسب وکار و درک داده ارتباط نزدیکی وجود دارد، چون درک مسئله داده‌کاوی حداقل نیاز به درک کمی از داده‌های موجود دارد.

- مرحله سوم، آماده‌سازی داده‌ها[1] است که همه فعالیت‌ها برای تبدیل داده‌های خام اولیه به مجموعه داده‌های نهایی که در مدل وارد می‌شوند را در بر می‌گیرد. برای این منظور متغیرهایی که قرار است تحلیل شوند و یا برای تحلیل مناسب هستند انتخاب می‌شوند (انتخاب ویژگی[2]). این مرحله کمّی‌سازی متغیرها مثلاً تبدیل متغیر ترتیبی[3] به عددی[4] و همچنین حذف داده‌های پرت و نامرتبط از مجموعه داده را در بر می‌گیرد.
- مرحله چهارم مدلسازی[5] است. در این مرحله تکنیک‌های مدل‌سازی مختلف انتخاب و اعمال می‌شوند و پارامترهای آن‌ها تا مقادیر بهینه کالیبره[6] می‌شوند، به عنوان مثال با تغییر نسبت تعداد داده‌های مجموعه تست[7] و داده‌های مجموعه آموزش[8] و یا تغییر پارامترهای اصلی[9] الگوریتم که تعیین آن‌ها با توجه به شرایط به دست خبره یا کاربر صورت می‌گیرد. به طور معمول، برای یک نوع مسئله داده‌کاوی چندین تکنیک وجود دارد. برخی از تکنیک‌ها به فرمت‌های داده خاصی نیاز دارند. همچنین ارتباط نزدیکی بین آماده‌سازی داده و مدل‌سازی وجود دارد. در اغلب موارد هنگام مدل‌سازی است که شخص متوجه مشکلات داده‌ها می‌شود یا ایده‌هایی برای ساخت داده‌های جدید به دست می آورد.
- مرحله پنجم اعتبارسنجی[10] است. در واقع، مدلی که ساخته شده ممکن است از دیدگاه تحلیل داده‌ها کیفیت بالایی داشته یا نداشته باشد و لذا پیش از اقدام به استقرار نهایی

1. Data Preparation
2. Feature Selection
3. Ordinal
4. Nominal
5. Modeling
6. Calibrated
7. Test Set
8. Train Set
9. Hyper Parameter
10. Evaluation

مدل، ارزیابی دقیق مدل و بررسی مراحل اجرا شده در ساخت مدل برای اطمینان از دستیابی مناسب به اهداف تجاری ضروری است.

- مرحله ششم، پیاده‌سازی[1] مدل است. معمولاً دانش به دست آمده باید به گونه‌ای سازماندهی و ارائه شود که سازمان بتواند از آن استفاده کند. در بسیاری از موارد این کاربر، و نه تحلیلگر داده خواهد بود، که مراحل استقرار را انجام خواهد داد.

مرحله اول یعنی درک کسب‌وکار و شناخت مسئله در بخش‌های پیشین مقاله با جزئیات بررسی شد. در این بخش و در ادامه به درک داده‌ها، آماده‌سازی داده‌ها و معرفی روش مدلسازی پرداخته می‌شود. یافته‌های حاصل از مدلسازی و اعتبارسنجی آن در بخش چهارم مقاله ارائه خواهد شد.

۱-۳. توصیف و درک داده‌ها

برای پیش‌بینی خرید بیمه متناسب با خصوصیات مشتری و بررسی علت پایین بودن ضریب نفوذ بیمه‌نامه حرف و مشاغل آزاد می‌بایست مشخصات جمعیت شناختی را مورد بررسی قرار داد. بررسی پیشینه پژوهش نیز نشان داد که متغیرهای جمعیت‌شناختی از جمله عوامل موثر بر تقاضا یا پیش‌بینی خرید بیمه‌نامه‌ها هستند. در این پژوهش از مجموعه داده‌های مربوط به بیمه‌شدگان حرف و مشاغل آزاد سازمان تامین اجتماعی در سال ۱۳۹۹ استفاده شده که شامل ۱۲۸۶۱۷۴ نفر بیمه شده است. جدول ۱ آمار توصیفی مربوط به متغیرها و مشخصه‌های اصلی مورد مطالعه را نمایش می‌دهد.

جدول ۱. مشخصات جمعیت شناختی بیمه شده‌های حرف و مشاغل آزاد سازمان تأمین اجتماعی

مشخصات جمعیت شناختی	جنسیت	زن	۵۷۳۹۴۷ (۴۴/۶ درصد)
		مرد	۷۱۲۲۲۷ (۵۵/۴ درصد)
	سن	کمترین	۱۴

1. Deployment

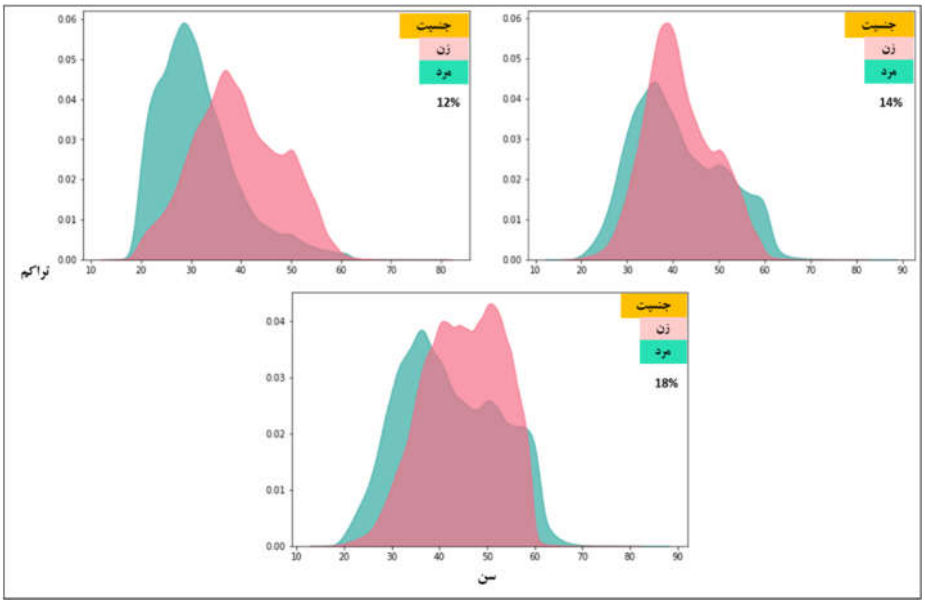
		متوسط	۴۰
		بیشترین	۸۶
	سابقه کار (روز)	کمترین	۱
		متوسط	۲۷۱۰
		بیشترین	۱۶۱۲۵
	میانگین درآمد ماهیانه (میلیون تومان)	کمترین	۱/۹۱
		متوسط	۱/۹۷
		بیشترین	۱۸/۶
نوع بیمه‌نامه	مستمری بازنشستگی (۱۲ درصد)		۴۱۹۹۶۴
	مستمری بازنشستگی و ازکارافتادگی (۱۴ درصد)		۱۸۲۲۵۵
	مستمری بازنشستگی، ازکارافتادگی و بازماندگان (۱۸ درصد)		۶۸۳۹۵۵

منبع: سازمان تامین اجتماعی (۱۴۰۰)

هدف این مطالعه بررسی و پیش‌بینی خرید بیمه حرف و مشاغل آزاد سازمان تأمین اجتماعی است. قرار است مدلی ساخته شود که بتواند بر اساس چهار ویژگی سن، جنسیت، سابقه و متوسط درآمد افراد را در دو طبقه متقاضی بیمه‌نامه با مستمری بازنشستگی (حق‌بیمه ۱۲ درصد) یا متقاضی بیمه‌نامه پوشش دهنده ریسک ازکارافتادگی و فوت علاوه بر مستمری بازنشستگی (حق‌بیمه ۱۴ و ۱۸ درصد) طبقه‌بندی کند. به‌عبارت دیگر، هدف این است که بر اساس چهار ویژگی دردسترسِ ثبت‌شده[1] (جنسیت، سن، سابقه‌کار و میانگین درآمد بیمه-شده) به‌عنوان متغیرهای ورودی مدل، پیش‌بینی شود که افراد کدام‌یک از بیمه‌نامه‌ها را انتخاب خواهند کرد. ابتدا به بررسی ارتباط متغیر هدف یعنی نوع بیمه‌نامه خریداری شده با متغیرهای توضیحی یا ورودی به تفکیک پرداخته خواهد شد که قطعاً موجب روشن‌تر شدن کار پیش از ورود به یافته‌های پژوهش خواهد شد.

۱. متأسفانه ویژگی‌ها و اطلاعات بیشتری از بیمه‌شدگان حرف و مشاغل آزاد در دسترس محققان قرار نگرفته است.

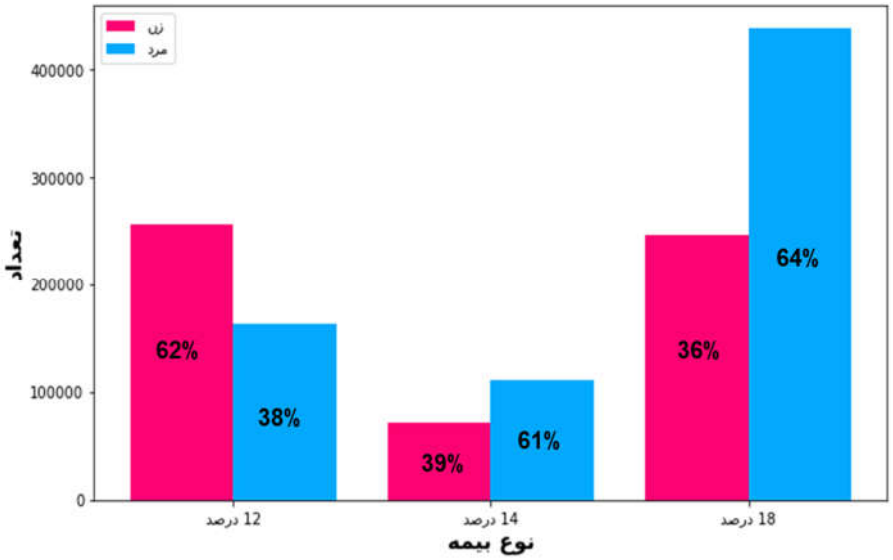
یکی از متغیرهای ورودی مهم سن بیمه‌شده است. نمودار ۲ توزیع سنی زنان و مردان را به تفکیک نوع بیمه‌نامه نمایش می‌دهد. نکته قابل توجه در این نمودار مربوط به بیمه ۱۲ درصد پوشش دهنده بازنشستگی است که متمایز از توزیع سنی در سایر انواع بیمه‌نامه‌های ۱۴ و ۱۸ درصد است. همانطور که می‌توان دید در میان بیمه‌شدگان مرد، فقط جوانان هستند که این نوع بیمه ۱۲ درصدی را تقاضا می‌کنند در حالی‌که مردان در سنین بالاتر خریدار بیمه‌نامه‌های ۱۴ و ۱۸ درصدی هستند که علاوه بر بازنشستگی، ریسک فوت و ازکارافتادگی را نیز پوشش می‌دهد. در نمودار می‌توان دید که برای زنان اما این موضوع برقرار نیست و زنان در همه سنین متقاضی بیمه ۱۲ درصدی پوشش دهنده بازنشستگی هستند و توزیع سن زنان در بیمه‌نامه ۱۲ درصدی متقارن است.



نمودار ۱. وضعیت توزیع سنی جمعیت زنان و مردان در بیمه‌های مختلف حرف و مشاغل آزاد

منبع: یافته‌های پژوهش

همانطور که در بررسی سن نیز دیده شد، متغیر ورودی و توضیحی جنسیت بیمه‌شدگان بسیار اهمیت دارد و از این رو نگاهی به آمار توصیفی خرید بیمه‌نامه‌ها بر اساس جنسیت می‌تواند مفید باشد. برای فهم بهتر می‌توان به سهم مردان و زنان از هر نوع بیمه‌نامه ۱۲ درصد (بازنشستگی)، ۱۴ درصد (بازنشستگی و ازکارافتادگی) و ۱۸ درصد (بازنشستگی، ازکارافتادگی و فوت) که به تفکیک در نمودار ۳ نشان داده شده توجه کرد.



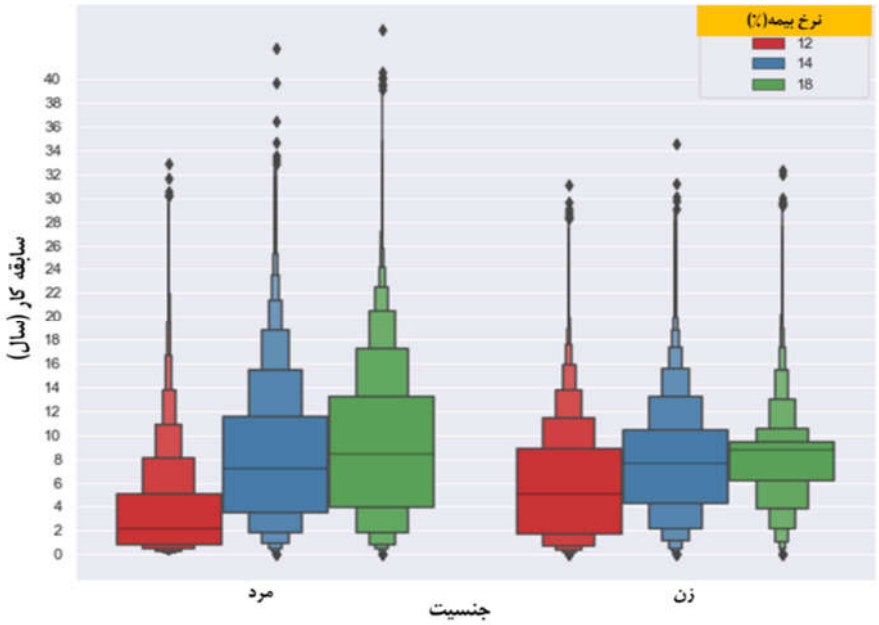
نمودار ۳. سهم مردان و زنان در هر یک از بیمه‌های حرف و مشاغل آزاد

منبع: یافته‌های پژوهش

همانطور که در نمودار ۳ می‌توان دید، مشابه حالت کلی همچنان سهم مردان در بیمه‌نامه‌های ۱۴ و ۱۸ درصدی بیشتر و حدوداً دو برابر زنان است. در حالی‌که به‌وضوح می‌توان دید این موضوع در بیمه‌نامه‌های با نرخ ۱۲درصد برعکس است و سهم زنان حدود ۲ برابر مردان است. این نمودار ساده به‌خوبی نشان می‌دهد که زنان از بیمه ۱۲ درصد و مردان از بیمه‌های ۱۴ و ۱۸ درصد استقبال می‌کنند. یک دلیل مهم این امر را باید در قوانین تکفل در ایران

جستجو کرد که تکفل فرزندان و زن را بر عهده شوهر یا پدر قرار داده است. بر این اساس، در صورت فوت بیمه‌شده مرد مستمری بازماندگان به همسر و فرزندانش تعلق می‌گیرد در حالی‌که در صورت فوت بیمه‌شده زن، در حالت عادی مستمری به همسر و فرزندان تعلق نخواهد گرفت. همین امر باعث شده تا زنان دلیل محکمی برای خرید بیمه‌نامه‌های پوشش دهنده ریسک ازکارافتادگی و فوت نداشته باشند و با پرداخت حق‌بیمه کمتر، بیمه‌نامه ۱۲ درصدی را انتخاب کنند که صرفاً بازنشستگی را پوشش می‌دهد. در مقابل، مردان که عموماً افرادی (شامل همسر و فرزندان) را تحت تکفل خود دارند و غالباً با ریسک بیشتری برای ازکارافتادگی مواجه هستند، ترجیح داده‌اند با پرداخت حق‌بیمه‌های بالاتر بیمه‌نامه‌هایی را انتخاب کنند که علاوه بر بازنشستگی، ازکارافتادگی و فوت را نیز ارائه دهد.

یکی دیگر از متغیرهای توضیحی و مهم و تأثیرگذار سابقه کار است که در نمودار ۴ به تفکیک جنسیت و برای بیمه‌نامه‌های مختلف نشان داده شده است. می‌توان دید اکثر مردان با سابقه کاری بیشتر، تمایل به استفاده از بیمه‌نامه با نرخ ۱۸ درصد و پس از آن بیمه‌های با نرخ ۱۴ درصد را دارند. اما برای زنان کمی متفاوت است و ارتباط قوی بین سابقه کار و نوع بیمه‌نامه برای زنان وجود ندارد. می‌توان دید اکثر جمعیت زنان با سابقه کاری ۲ تا ۹ سال متقاضی بیمه‌نامه با نرخ ۱۲ درصد هستند و پس از آن، اکثر جمعیت با سابقه کاری ۴ تا ۱۰ سال متقاضی بیمه‌نامه با نرخ ۱۴ درصد هستند.

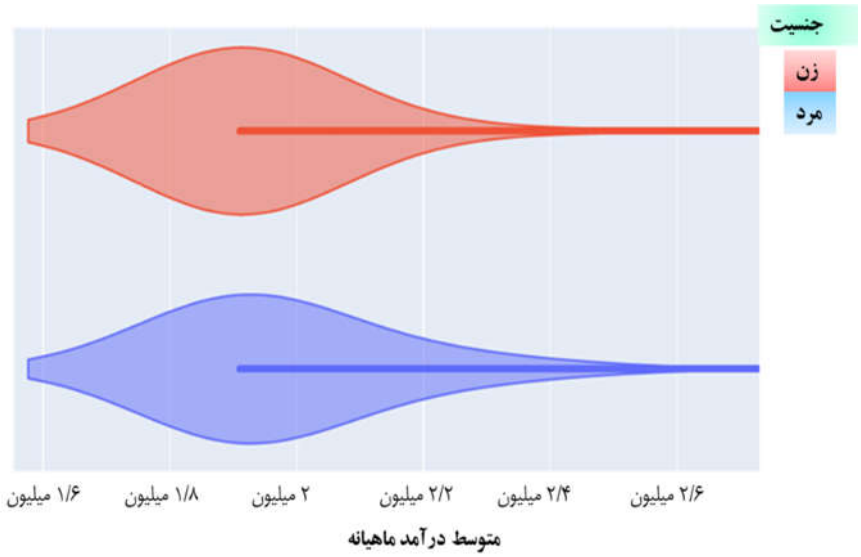


نمودار ٤. وضعیت سابقه کار نسبت به جنسیت در بیمه‌های مختلف حرف و مشاغل آزاد

منبع: یافته‌های پژوهش

یکی دیگر از ویژگی‌های ثبت شده از بیمه‌شدگان متوسط درآمد ماهانه آن‌هاست که می‌توان از آن برای توضیح و پیش‌بینی خرید بیمه‌نامه‌ها استفاده کرد. نمودار ۵ توزیع درآمد افراد بر اساس جنسیت آن‌ها را نشان می‌دهد. از این نمودار نیز می‌توان دریافت که اکثر افرادی که از بیمه‌های حرف و مشاغل آزاد استفاده می‌کنند در سال ۱۳۹۹ درآمدی بین ۱/۹ میلیون تومان تا ۲/۲ میلیون تومان داشته‌اند که بسیار نزدیک به حداقل دستمزد آن سال معادل ۱۹۱۰۴۲۷تومان بوده است (خط ممتد در نمودار بازه درآمدی را نشان می‌دهد و در قسمت-هایی که نمودار برآمدگی دارد نشان دهنده وجود افراد بیشتر با آن درآمد است). این موضوع البته به احتمال زیاد به دلیل عدم گزارش صحیح درآمد توسط بیمه‌شده برای پرداخت حق‌بیمه کمتر است. طبق قوانین بازنشستگی در ایران تنها درآمد دو سال آخر در تعیین

مستمری بازنشستگی تأثیرگذار است و به همین دلیل افراد در طول دوران کاری (به جز دو سال پایانی) انگیزه دارند تا حداقل درآمد را گزارش کنند و تنها سوابق خود را تکمیل کنند.



نمودار ۵. وضعیت توزیع درآمد مردان و زنان تحت بیمه حرف و مشاغل آزاد

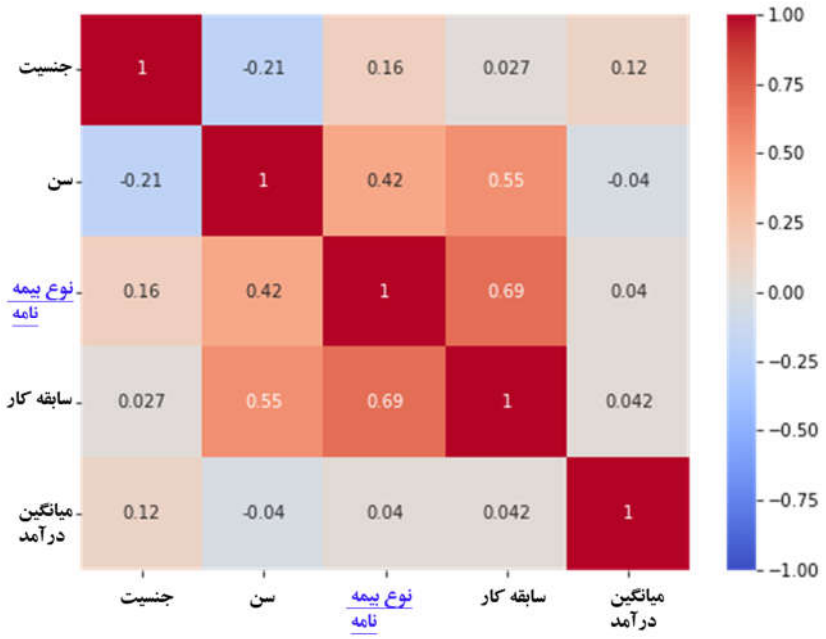
منبع: یافته‌های پژوهش

۲-۳. آماده‌سازی داده‌ها

آماده‌سازی داده‌ها و انتخاب ویژگی‌های مناسب شامل سه مرحله است که در هر قسمت عملیات‌های مختلفی بر روی داده‌ها انجام می‌شود. مرحله اول پیش پردازش داده‌هاست. در این مرحله متغیر جنسیت با جایگذاری مقدار ۰ برای زنان و مقدار ۱ برای مردان به مشخصه با مقادیر عددی تبدیل شد. یکی دیگر از اقدامات مهم در این مرحله شناسایی مقادیر گم شده و مقادیر تکراری است که با رصد تمامی داده‌ها و مشاهدات مشخص شد هیچ مقدار گم‌شده‌ای برای هیچکدام از مشخصه‌ها وجود ندارد. همچنین در پیش پردازش داده‌ها باید داده‌های پرت نیز مورد بررسی قرار گیرد که البته این مهم بعد از مرحله دوم از آماده‌سازی داده‌ها یعنی انتخاب موثرترین مشخصه‌ها انجام خواهد گرفت. برای اینکه مشخص شود

کدام یک از ویژگی‌های افراد (سن، جنسیت، درآمد و سابقه کار) در تمایل آن‌ها برای انتخاب نوع بیمه‌نامه موثرتر است، از میزان همبستگی بین متغیرهای ورودی با متغیر هدف یا نوع بیمه‌نامه انتخاب‌شده توسط افراد استفاده شده است. معمولاً هدف از انتخاب موثرترین ویژگی‌ها این است که تنها از ویژگی‌های مهم در ساخت مدل و الگوریتم استفاده شود که البته در اینجا این امر موضوعیت ندارد و همانطور که با مرور آمار توصیفی دیده شد هر چهار ویژگی بر انتخاب بیمه‌نامه مؤثرند و مورد استفاده قرار خواهند گرفت. در اینجا هدف از انتخاب موثرترین ویژگی‌ها صرفاً مرتب‌سازی داده‌هاست تا بتوان داده‌های پرت را حذف کرد.

با توجه به نمودار حرارتی ۶ می‌توان میزان همبستگی بین مشخصه‌های بیمه‌شدگان و متغیر هدف یا نوع بیمه‌نامه (۱۲ درصدی و بیمه‌نامه‌های (۱۴ و ۱۸درصدی) با پوشش ریسک بیشتر) را مشاهده کرد. میزان همبستگی بین متغیر «سابقه کار» با متغیر هدف «نوع بیمه‌نامه» تقریباً ۷۰ درصد است که بسیار بیشتر از همبستگی سایر متغیرهای ورودی با متغیر هدف می‌باشد. در نتیجه، از لحاظ آماری می‌توان گفت که سابقه کار افراد بیشترین تاثیر را نسبت به سایر متغیرهای ورودی (سن، جنسیت و میانگین درآمد) در انتخاب بیمه‌نامه دارد. البته تمامی متغیرهای دیگر از جمله جنسیت یا سن نیز تأثیرگذارند.



نمودار ٦. وضعیت همبستگی متغیرهای موردبررسی با یکدیگر

منبع: یافته‌های پژوهش

از روش فیلتر که از معروف‌ترین روش‌ها در انتخاب ویژگی در مسائل طبقه‌بندی به‌شمار می‌رود نیز برای انتخاب مؤثرترین مشخصه استفاده شد. این روش رابطه‌ی بین متغیرهای ورودی و متغیر هدف را ارزیابی و نتیجه آن، انتخاب آن دسته از متغیرهای ورودی است که قوی‌ترین رابطه را با متغیر هدف دارند. در این روش برای انتخاب ویژگی از معیارهای آماری از جمله آزمون مربع کای $Chi - square = (\frac{(Actual-Expected)^2}{Expected})^{\frac{1}{2}}$ بین متغیرهای ورودی برای امتیاز دادن استفاده می‌شود. طبق نتایج این روش نیز مشاهده شد که سابقه کار بیشترین امتیاز را دارد و در نتیجه تاثیر آن بر متغیر هدف به مراتب بیشتر از بقیه مشخصه‌ها می‌باشد. بر این اساس، مجموعه داده موجود بر اساس سابقه کار تقسیم‌بندی شد و پس از حذف داده‌های پرت ۷۲۶۵۸۲ رکورد باقی ماند که در مراحل بعدی و مدلسازی مورد

استفاده قرار می‌گیرد. مرحله سوم از آماده‌سازی داده‌ها، هم‌مقیاس‌سازی ویژگی‌ها است. منظور از هم‌مقیاس‌سازی این است که باید بازه تغییرات در هر ویژگی (متغیر) یکسان شود تا الگوریتم از واحد و مقیاس متغیر متأثر نشود. در اینجا از روش نرمالایز کردن مطابق رابطه $y = \frac{x-min}{max-min}$ استفاده می‌شود تا مقادیر همه متغیرها در محدوده جدید ۰ و ۱ قرار گیرند.

۳-۳. مدلسازی

برای مدلسازی در این مقاله از روش‌های مبتنی بر یادگیری ماشین استفاده می‌شود که کاربرد زیادی در حوزه‌های بیمه و بانکی پیدا کرده‌اند (متدین و همکاران ۱۴۰۰، متفکرآزاد و همکاران ۱۳۹۰). یادگیری ماشین شاخه‌ای از هوش مصنوعی[1] و علوم کامپیوتر[2] است که در آن کامپیوترها از مجموعه‌ای از الگوریتم‌ها استفاده می‌کنند تا الگو و روند داده‌های تاریخی را یاد بگیرند و بتوانند رفتار و مقادیر داده‌های آینده را پیش‌بینی کنند، در واقع رفتاری همانند انسان در یادگیری را دارند که به تدریج آن را بهبود می‌بخشند (هرویتز و کریش[3]، ۲۰۱۸). یادگیری ماشین در حال حاضر یکی از حوزه‌های رو به رشد در علم داده[4] است که با استفاده از روش‌های آماری، الگوریتم‌ها را برای طبقه‌بندی یا پیش‌بینی آموزش می‌دهد تا بینش‌هایی کلیدی در پروژه‌های داده‌کاوی را آشکار کنند. به طور کلی می‌توان یادگیری ماشین را به دو دسته یادگیری تحت نظارت[5] و یادگیری بدون نظارت[6] تقسیم بندی کرد. یادگیری تحت نظارت معمولاً با مجموعه‌ای از داده‌ها و درک خاصی از نحوه طبقه‌بندی آن داده‌ها آغاز می‌شود، و برای یافتن الگوهایی در داده‌ها در نظر گرفته شده است که می‌تواند در یک فرآیند تحلیلی اعمال شود. یادگیری بدون نظارت زمانی مناسب است که با حجم عظیمی از

1. Artificial Intelligence (Ai)
2. Computer Science
3. Hurwitz & Kirsch
4. Data Science
5. Supervised Learning
6. Unsupervised Learning

داده‌ها سروکار داشته باشیم که برچسب گذاری[1] نشده باشند (ویژگی هدف مشخص نباشد) که در آن یک فرآیند تکراری از تجزیه و تحلیل داده‌ها بدون دخالت انسان انجام می‌شود.

در این مطالعه با توجه به آمار قابل اطمینان که از دفتر آمار، اطلاعات و محاسبات سازمان تامین اجتماعی گرفته شده، برای مدلسازی از دو الگوریتم درخت تصمیم و جنگل تصادفی استفاده می‌شود که هر دو جزء الگوریتم‌های طبقه‌بندی از یادگیری ماشین می‌باشند. طبقه-بندی زمانی استفاده می‌شود که داده‌ها بر اساس یک ویژگی هدف به گروه‌های مختلف طبقه‌بندی شوند و یا احتمال نتیجه برچسب هدف را بر اساس سوابق تاریخی بتوان پیش‌بینی کرد.

درخت تصمیم در یادگیری ماشین جزء الگوریتم‌های یادگیری تحت نظارت است که برای نمایش طبقه‌بندی نمونه‌ها به کار می‌رود. این الگوریتم با تقسیم داده‌های آموزش به گره‌های[2] مجزا و در نظر گرفتن ویژگی‌ها به صورت یک به یک ساخته و دارای ساختار درختی سلسله مراتبی[3] است که اجزای اصلی آن شامل: یک گره ریشه، شاخه‌ها، گره‌های داخلی و گره‌های برگ می‌باشد. شروع یک درخت تصمیم با یک گره ریشه (مهم‌ترین ویژگی که در اینجا سابقه‌کار شناخته شد) است که هیچ شاخه ورودی ندارد. سپس شاخه-های خروجی از گره ریشه وارد گره‌های داخلی شده که به عنوان گره‌های تصمیم نیز شناخته می‌شوند. گره‌های برگ همه نتایج ممکن را در مجموعه داده نشان می‌دهند. یادگیری درخت تصمیم با انجام یک جستجوی حریصانه برای شناسایی نقاط تقسیم بهینه در یک درخت، از استراتژی تقسیم[4] و غلبه[5] استفاده می‌کند، که این فرآیند تقسیم به صورت بازگشتی از بالا به پایین تکرار می‌شود تا زمانی که همه یا اکثر رکوردها تحت برچسب‌های کلاس خاصی طبقه‌بندی شوند. معیارهای متعددی برای تصمیم‌گیری در مورد محل تقسیم و میزان تقسیم وجود دارند که از میان این روش‌ها و معیارها دو روش کسب اطلاعات[6] و

1. Unlabeled
2. Node
3. Hierarchical
4. Divide
5. Conquer
6. Information Gain

ناخالصی جینی[1] به عنوان دو معیار محبوب تقسیم شناخته می‌شوند. این روش‌ها به ارزیابی کیفیت شرایط هر آزمون و اینکه چگونه می‌توان نمونه‌ها را در یک کلاس طبقه‌بندی کرد کمک می‌کند.

توضیح روش کسب اطلاعات به مفهوم آنتروپی[2] مرتبط است که ناخالصی مقادیر نمونه را اندازه می‌گیرد و برای محاسبه آن از رابطه (۱) استفاده می‌شود.

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

که عبارت‌های S مجموعه داده، C کلاس‌های مجموعه S و P(c) نسبت نقاط داده‌ای متعلق به کلاس c به تعداد کل نقاط داده در مجموعه S است. مقادیر آنتروپی می‌تواند بین ۰ و ۱ قرار گیرد. اگر تمام نمونه‌های مجموعه داده S متعلق به یک کلاس باشند، آنتروپی برابر صفر خواهد بود. اگر نیمی از نمونه‌ها به عنوان یک کلاس و نیمی دیگر در کلاس دیگری طبقه‌بندی شوند، آنتروپی به بالاترین حد خود یعنی ۱ خواهد رسید. برای انتخاب بهترین ویژگی جهت تقسیم و یافتن درخت تصمیم بهینه نیز باید از ویژگی (مشخصه یا متغیر) با کمترین مقدار آنتروپی استفاده شود. به تفاوت آنتروپی قبل و بعد از تقسیم در یک ویژگی معین نیز اطلاعات بدست آمده گفته می‌شود. ویژگی با بالاترین بهره اطلاعات بهترین تقسیم را ایجاد می‌کند زیرا بهترین کار را در طبقه‌بندی داده‌های آموزشی بر اساس طبقه‌بندی هدف خود انجام می‌دهد. کسب اطلاعات معمولاً با رابطه (۲) نشان داده می شود.

$$Information\ Gain(S, \alpha) = Entropy(S) - \sum_{v \in vcalues(\alpha)} \frac{|S_v|}{|S|} Entropy(S_v) \quad (۲)$$

1. Gini Impurity
2. Entropy

که α یک ویژگی یا برچسب کلاس خاص بوده و $\frac{|S_v|}{|S|}$ نسبت مقادیر در مجموعه داده S_v به تعداد مقادیر در مجموعه داده S است. ابتدا آنتروپی گره والد محاسبه می‌شود. سپس آنتروپی هر گره جداگانه در تقسیم (انشعاب) فرموله و در انتها میانگین وزنی تمام گره‌ها محاسبه می‌شود. در روش دوم یا شاخص جینی، دو آیتم از یک جامعه به طور تصادفی انتخاب و در یک کلاس طبقه‌بندی می‌شوند. این شاخص با متغیر هدف طبقه‌بندی مانند «موفقیت» یا «شکست» کار می‌کند و فقط برای تقسیم (انشعاب) متغیرهای باینری (دودویی) کاربرد دارد. مجموع مجذور احتمال موفقیت و شکست $(p^2 + q^2)$ برای محاسبه جینی زیرگره‌ها استفاده می‌شود. هرچه مقدار شاخص جینی بیشتر باشد، ارزش همگنی بیشتر است (مائمون و روکاچ[1]، ۲۰۱۴).

الگوریتم دیگری که در این مطالعه استفاده خواهد شد، الگوریتم جنگل تصادفی مجموعه‌ای از n درخت تصمیم است. هر درخت تصمیم در جنگل بر روی زیرمجموعه‌های مختلف مجموعه داده‌های آموزشی، آموزش داده می‌شود. جنگل تصادفی از انتخاب ویژگی تصادفی در حالی که درخت در حال رشد است استفاده می‌کند. در مورد مجموعه داده‌های چند بعدی این ویژگی بسیار مهم است، به‌ویژه زمانی که صدها یا هزاران ویژگی وجود دارد، به عنوان مثال در تشخیص پزشکی و اسناد، بسیاری از ویژگی‌های ضعیف مرتبط ممکن است اصلاً در یک درخت تصمیم ظاهر نشوند. این الگوریتم یک راه حل تخصصی برای اکثر مشکلات است و به عنوان یک روش یادگیری گروهی[2] شناخته می‌شود زیرا گروهی از مدل‌های ضعیف با هم ترکیب می‌شوند تا این مدل قدرتمند را تشکیل دهند (سولیوان[3]،

1. Maimon & Rokach

۲. یادگیری گروهی (Ensemble method) تشکیل شده از مجموعه‌ای از طبقه‌بندی‌کننده‌ها - به عنوان مثال درخت‌های تصمیم و پیش‌بینی‌های آن‌ها برای شناسایی محبوب‌ترین نتیجه جمع‌آوری می‌شوند.

3. Sullivan

۲۰۱۷)، فرضیه نهایی در جنگل تصادفی با استفاده از روش رای اکثریت[1] در بین درختان تولید می‌شود (تنها[2] و همکاران، ۲۰۱۷).

٤. یافته‌های پژوهش

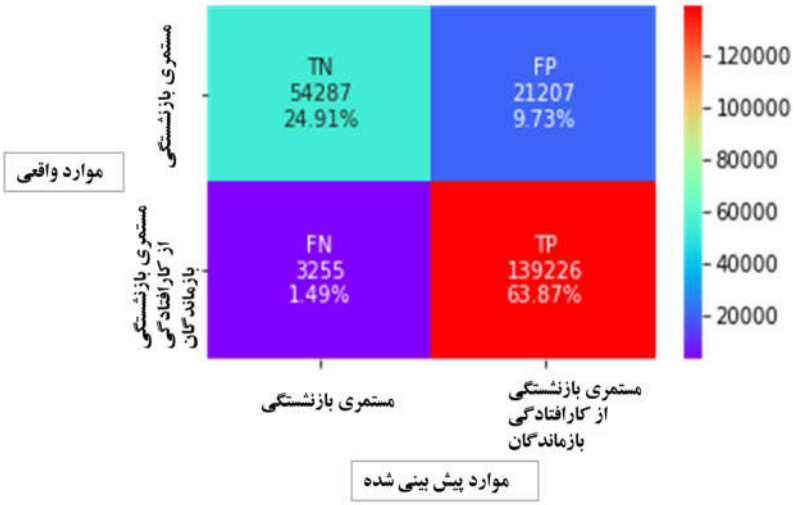
این مقاله با استفاده از روش کریپس و الگوریتم‌های طبقه‌بندی درخت تصمیم و جنگل تصادفی به بررسی رفتار و پیش‌بینی خرید دو نوع بیمه‌نامه حرف مشاغل آزاد پوشش دهنده بازنشستگی (۱۲ درصد) و پوشش دهنده ریسک‌های ازکارافتادگی و بازماندگی علاوه بر بازنشستگی (۱۴ و ۱۸ درصد) پرداخت.

٤-۱. نتایج پیش‌بینی الگوریتم درخت تصمیم و جنگل تصادفی

در این بخش و در ابتدا نتایج حاصل از بکارگیری الگوریتم درخت تصمیم ارائه خواهد شد. در اجرای این الگوریتم ۷۰ درصد از داده‌ها به عنوان داده‌های آموزش[3] در نظر گرفته شد و از پارامتر اصلی[4] حداکثر عمق[5] برابر با ۶ و شاخص انشعاب جینی[6] استفاده شده است. نتایج حاصل از این الگوریتم در ماتریس اغتشاش نمودار ۷ نمایش داده شده است. ماتریس اغتشاش به عنوان شاخصی از مشخصه‌های یک قانون طبقه‌بندی (تفکیک کننده) استفاده می‌شود. این ماتریس شامل تعداد عناصری است که به درستی یا نادرستی در هر کلاس طبقه‌بندی شده‌اند. تعداد مشاهداتی که به درستی برای هر کلاس طبقه‌بندی شده در قطر اصلی ماتریس و تعداد مشاهداتی که به اشتباه طبقه‌بندی شده در خارج از قطر اصلی قرار می‌گیرند. فرض کنید دو کلاس مثبت و منفی وجود دارد. شیوه کار به این صورت است که برای هر مشاهده در مجموعه تست، کلاس واقعی را با کلاسی که توسط طبقه‌بندی‌کننده

1. Voting
2. Tanha
3. Training Set
4. Hyperparameter
5. Maximum Depth
6. Gini Index

آموزش‌دیده و تخصیص داده شده مقایسه می‌کنند. برای مشاهداتی که در کلاس مثبت طبقه‌بندی شده‌اند، اگر به درستی طبقه‌بندی شده باشند مثبت واقعی[1] (TP) و مشاهده منفی که به اشتباه طبقه‌بندی شده مثبت کاذب (FP) نامیده می‌شود. به‌طور مشابه برای مشاهداتی که در کلاس منفی طبقه‌بندی شده‌اند، اگر مشاهده مثبتی باشد و به اشتباه طبقه‌بندی شده باشد منفی کاذب (FN) و بر عکس، یک مشاهده منفی که به درستی طبقه‌بندی شده باشد منفی واقعی[2] (TN) نامیده می‌شود.



نمودار ۷. نتایج ماتریس اغتشاش در الگوریتم درخت تصمیم

منبع: یافته‌های پژوهش

عملکرد الگوریتم را می‌توان با معیارهای مختلفی ارزیابی کرد. نخستین معیار، معیار دقت[3] است که طبق رابطه (۳) بیان می‌شود.

1. True Positive
2. True Negative
3. Accuracy Score

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

این معیار درصد طبقه‌بندی صحیح در هر دو بیمه‌نامه را اندازه‌گیری می‌کند. مطابق نمودار ۷ می‌توان دید که ۱۳۹۲۲۶ نفر به‌درستی متقاضی بیمه‌نامه ۱۴ و ۱۸ درصدی و ۵۴۲۸۷ نفر نیز به درستی متقاضی بیمه‌نامه ۱۲ درصدی طبقه‌بندی شده‌اند. بنابراین، از مجموع ۲۱۷۹۷۵ رکورد داده‌های تست، ۸۸ درصد آن‌ها به درستی طبقه‌بندی شده‌اند. نرخ دقت ۸۸ درصد نشان می‌دهد که مدل می‌تواند خروجی یا خریداری بیمه‌نامه ۱۲ درصدی در مقابل بیمه‌نامه ۱۴ و ۱۸ درصدی را با دقت ۸۸ درصد پیش‌بینی کند. معیار دوم، معیار پوشش[1] است که طبق رابطه (۴) به ما می‌گوید از میان افرادی که واقعاً متقاضی بیمه ۱۴ و ۱۸ درصدی بوده‌اند (سطر پایین ماتریس اغتشاش)، چند درصد درست تشخیص داده شده‌اند. در واقع تمرکز اصلی معیار پوشش بر روی داده‌هایی است که واقعاً متقاضی بیمه ۱۴ و ۱۸ درصدی (با خدمات مستمری بازنشستگی، از کارافتادگی و بازماندگان) بوده‌اند و هدف آن پوشش حداکثری این گروه است تا هیچ یک از آن‌ها به اشتباه در دسته دیگر قرار نگیرند. از نتایج ماتریس اغتشاش می‌توان دید که از مجموع ۱۴۲۴۸۱ نفری که متقاضی بیمه ۱۴ و ۱۸ درصدی بوده‌اند، تعداد ۱۳۹۲۲۶ نفر یا ۹۷/۷ درصد به درستی تشخیص داده شده‌اند.

$$Recall = \frac{TP}{TP+FN} \quad (۴)$$

معیار سوم، معیار صحت[2] است که طبق رابطه (۵) تمرکز اصلی آن بر روی مواردی است که توسط الگوریتم در دسته ۱۴ و ۱۸ درصدی طبقه‌بندی و متقاضی این بیمه‌نامه پیش‌بینی شده‌اند (ستون دوم ماتریس اغتشاش) و به ما می‌گوید صحت این پیش‌بینی چقدر بوده است. از نتایج ماتریس اغتشاش می‌توان دید که از مجموع ۱۶۰۴۳۳ نفری که پیش‌بینی‌شده متقاضی بیمه ۱۴ و ۱۸ درصدی باشند، تعداد ۱۳۹۲۲۶ نفر یا ۸۶/۷ درصد واقعاً متقاضی بوده و این پیش‌بینی در مورد آن‌ها صحیح است.

1. Recall
2. Precision

$$Precision = \frac{TP}{TP + FP} \quad (۵)$$

بین معیارهای صحت و پوشش یک بده‌بستان[1] وجود دارد و تلاش برای بهبود یک معیار اغلب منجر به بدتر شدن معیار دوم می‌گردد. ساده‌ترین راه حل این مشکل این است که در یک تصمیم‌گیری چند معیاره (MCDM) هر دو را ترکیب کرده و جمع وزنی این معیارها را مدنظر قرار داد. این معیار ترکیبی امتیاز-F است که طبق رابطه (۶) مقدار آن برابر ۹۱/۹ درصد محاسبه می‌شود. این معیار ترکیب متعادلی از دو معیار صحت و پوشش است. هر چه معیار F بیشتر باشد نشان دهنده تایید دقت مدل است به‌این معنی که حد آستانه در بهترین حالت خود قرار گرفته و پیش‌بینی با کمترین خطا همراه است.

$$F = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (۶)$$

در محاسبه معیارهای یادشده تمرکز بر بیمه‌نامه ۱۴ و ۱۸ درصدی بود و دقت، صحت و پوشش الگوریتم در طبقه‌بندی این گروه ارزیابی شد. به‌عبارت دیگر، عملکرد مدل و الگوریتم در پیش‌بینی افرادی که متقاضی خرید بیمه‌نامه پوشش ریسک بالاتر (بیمه ۱۴ و ۱۸ درصدی) هستند مدنظر بوده است. اما از آنجائی‌که هر دو مقدار متغیر هدف یا هر دو نوع بیمه‌نامه برای ما به یک اندازه اهمیت دارند، می‌توان توجه را به سمت گروه دیگر یعنی بیمه ۱۲ درصدی گرداند و با استفاده از اطلاعات ماتریس اغتشاش عملکرد الگوریتم در پوشش این گروه را بررسی کرد. برای این منظور، این بار، بر افرادی که واقعاً متقاضی بیمه ۱۲ درصدی بوده‌اند (سطر بالای ماتریس اغتشاش) تمرکز می‌شود تا درستی تشخیص الگوریتم برای این گروه تعیین شود. این معیار در اصطلاح «تشخیص پذیری[2]» خوانده می‌شود و همانطور که از ماتریس اغتشاش می‌توان دید، از مجموع ۷۵۴۹۴ نفر که واقعاً متقاضی بیمه‌نامه ۱۲ درصدی (با خدمات مستمری بازنشستگی) بوده‌اند، تعداد ۵۴۲۸۷ یا ۷۱/۹ درصد به درستی طبقه‌بندی و پیش‌بینی

1. Trade-off
2. Specificity

شده‌اند. با توجه را به مواردی که متقاضی بیمه‌نامه ۱۲ درصدی پیش‌بینی‌شده‌اند (ستون اول ماتریس اغتشاش) متمرکز کرد و صحت این پیش‌بینی را سنجید. این شاخص «ارزش پیش بینی شده منفی[1]» نام دارد و می‌توان دید که طبق نتایج ماتریس اغتشاش، از ۵۷۵۴۲ نفری که پیش‌بینی شده متقاضی بیمه‌نامه ۱۲ درصدی باشند تعداد ۵۴۲۸۷ یا ۹۴/۳ درصد واقعاً متقاضی خرید این بیمه‌نامه بوده و این پیش‌بینی در مورد آن‌ها صحیح است.

در جدول ۲ به طور خلاصه نتایج حاصل از معیارهای ارزیابی الگوریتم درخت تصمیم ارائه شده است.

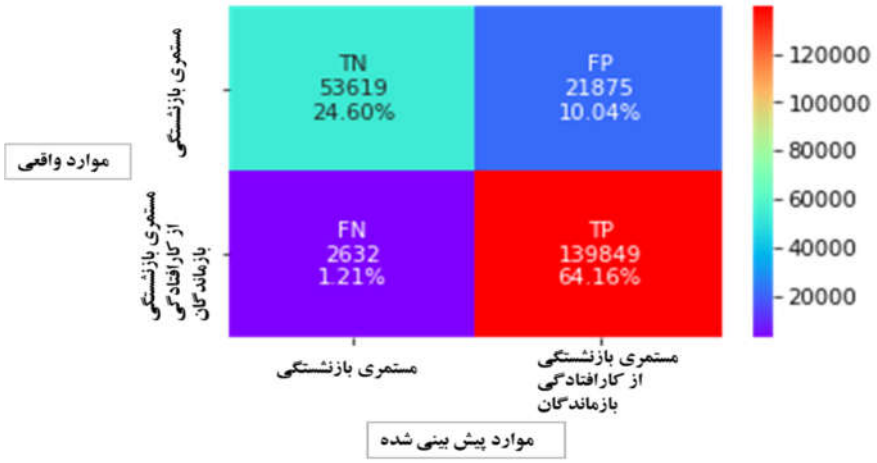
جدول ۲. نتایج حاصل از الگوریتم درخت تصمیم

نام معیار	معیار (درصد)
دقت	۸۸/۷
پوشش	۹۷/۷
صحت	۸۶/۷
امتیاز-F	۹۱/۹
تشخیص پذیری	۷۱/۹
ارزش پیش‌بینی شده منفی	۹۴/۳

منبع: یافته‌های پژوهش

علاوه بر درخت تصمیم، در این مقاله از الگوریتم جنگل تصادفی نیز برای پیش‌بینی خرید بیمه‌نامه‌ها استفاده شده است. نتایج بدست آمده از بکارگیری الگوریتم جنگل تصادفی با پارامترهای اصلی، حداکثر عمق ۸ و شاخص جینی و تعداد ۱۰۰ درخت مورد نیاز برای تخمین[2]، در نمودار ۸ نشان داده شده است.

1. Negative Predictive Value
2. N-Estimators



نمودار ۸. نتایج ماتریس اغتشاش در الگوریتم جنگل تصادفی

موارد واقعی و پیش‌بینی‌شده انتخاب دو نوع بیمه‌نامه را می‌توان در ماتریس اغتشاش دید. مانند قبل عملکرد الگوریتم را می‌توان با معیارهای مختلفی ارزیابی کرد. معیار دقت درصد طبقه‌بندی صحیح در هر دو بیمه‌نامه را اندازه‌گیری می‌کند. مطابق نمودار ۸ می‌توان دید که ۱۳۹۸۴۹ نفر به‌درستی متقاضی بیمه‌نامه ۱۴ و ۱۸ درصدی و ۵۳۶۱۹ نفر نیز به درستی متقاضی بیمه‌نامه ۱۲ درصدی طبقه‌بندی شده‌اند. بنابراین، از مجموع ۲۱۷۹۷۵ رکورد داده‌های تست، ۸۸ درصد آن‌ها به درستی طبقه‌بندی شده‌اند. معیار پوشش نشان می‌دهد از میان افرادی که واقعاً متقاضی بیمه ۱۴ و ۱۸ درصدی بوده‌اند (سطر پایین ماتریس اغتشاش)، چند درصد درست تشخیص داده شده‌اند. در واقع تمرکز اصلی معیار پوشش بر روی داده‌هایی است که واقعاً متقاضی بیمه ۱۴ و ۱۸ درصدی (با خدمات مستمری بازنشستگی، از کارافتادگی و بازماندگان) بوده‌اند و هدف آن پوشش حداکثری این گروه است تا هیچ یک از آن‌ها به اشتباه در دسته دیگر قرار نگیرند. از نتایج ماتریس اغتشاش می‌توان دید که از مجموع ۱۴۲۴۸۱ نفری که متقاضی بیمه ۱۴ و ۱۸ درصدی بوده‌اند، تعداد ۱۳۹۸۴۹ نفر یا ۹۸/۱ درصد به درستی تشخیص داده شده‌اند. معیار سوم، معیار صحت است که بر روی مواردی که توسط الگوریتم در دسته ۱۴ و ۱۸ درصدی طبقه‌بندی و متقاضی این بیمه‌نامه پیش‌بینی شده‌اند (ستون دوم ماتریس اغتشاش) تمرکز دارد و نشان می‌دهد

صحت این پیش‌بینی چقدر بوده است. از نتایج ماتریس اغتشاش می‌توان دید که از مجموع ۱۶۱۷۲۴ نفری که پیش‌بینی شده متقاضی بیمه ۱۴ و ۱۸ درصدی باشند، تعداد ۱۳۹۸۴۹ نفر یا ۸۶/۴ درصد واقعاً متقاضی بوده و این پیش‌بینی در مورد آن‌ها صحیح است. معیار امتیاز-F که ترکیبی است از معیارهای صحت و پوشش نیز برابر ۹۱/۹ درصد محاسبه می‌شود.

اگر این بار توجه را معطوف به گروه دیگر یعنی بیمه ۱۲ درصدی کرده و با استفاده از اطلاعات ماتریس اغتشاش عملکرد الگوریتم در پوشش این گروه را بررسی کنیم می‌توان معیار «تشخیص پذیری» را محاسبه کرد که نشان می‌دهد چند درصد این افراد توسط الگوریتم درست تشخیص داده شده‌اند. همانطور که از ماتریس اغتشاش می‌توان دید، از مجموع ۷۵۴۹۴ نفر که واقعاً متقاضی بیمه‌نامه ۱۲ درصدی (با خدمات مستمری بازنشستگی) بوده‌اند، تعداد ۵۳۶۱۹ یا ۷۱ درصد به درستی طبقه‌بندی و پیش‌بینی شده‌اند. یا می‌توان به ستون اول ماتریس اغتشاش توجه کرد و صحت این پیش‌بینی یعنی شاخص «ارزش پیش بینی شده منفی» را سنجید. طبق نتایج ماتریس اغتشاش، از ۵۶۲۵۱ نفری که پیش‌بینی شده متقاضی خرید بیمه‌نامه ۱۲ درصدی باشند تعداد ۵۳۶۱۹ یا ۹۵/۳ درصد واقعاً متقاضی این بیمه‌نامه بوده و این پیش‌بینی در مورد آن‌ها صحیح است. در جدول ۳ به طور خلاصه نتایج حاصل از معیارهای ارزیابی جنگل تصادفی ارائه شده است.

جدول ۳. نتایج حاصل از الگوریتم جنگل تصادفی

نام معیار	دقت معیار(درصد)
دقت	۸۸/۷
پوشش	۹۸/۱
صحت	۸۶/۴
امتیاز-F	۹۱/۹
تشخیص پذیری	۷۱
ارزش پیش‌بینی شده منفی	۹۵/۳

منبع: یافته‌های پژوهش

٤-۲. تحلیل و تفسیر ویژگی‌های فردی مؤثر بر انتخاب بیمه‌نامه‌های حرف و مشاغل آزاد

با اتکا به نتایج به دست آمده می‌توان گفت که سابقه کار افراد از موثرترین عوامل در انتخاب بیمه‌نامه افراد است. الگوریتم درخت تصمیم برای همه افراد از هر جنسیتی در صورتی که سابقه کار کمتر از ۴/۵ سال داشته باشند پیش‌بینی کرده که متقاضی بیمه ۱۲ درصدی خواهند بود و همچنین این الگوریتم نشان می‌دهد افراد تنها با افزایش سابقه کار تمایل به بیمه‌های ۱۴ و ۱۸ درصدی با پوشش بیشتر از جمله ازکارافتادگی و ریسک فوت پیدا می‌کنند. البته در اینجا جنسیت تاثیرگذار است و طبق نتایج بدست آمده مردان بسیار زودتر به خریداری بیمه‌های با پوشش بیشتر تمایل پیدا می‌کنند. جداول ۴ و ۵ به تفکیک میزان تمایل زنان و مردان در سوابق کاری، سن و سطوح درآمدی مختلف برای انواع بیمه‌نامه‌ها را نمایش می‌دهند.

همانطور که از جدول ۴ پیداست، زنان با سابقه کمتر از ۵/۵ سال پیش‌بینی شده که متقاضی بیمه‌نامه ۱۲ درصدی با پوشش بازنشستگی و زنان با سابقه‌کار بالاتر از ۹/۵ سال پیش‌بینی‌شده که خریدار بیمه‌نامه‌های ۱۴ و ۱۸ درصدی باشند. اما پیش‌بینی تقاضا برای زنانی که سابقه کاری بین ۵/۵ تا ۹/۵ سال دارند کمی سخت‌تر بوده و به دیگر متغیرها بستگی دارد. با توجه به نتایج به دست آمده می‌توان گفت زنانی که ۵/۵ تا ۷/۵ سال سابقه کاری، سن کمتر از ۳۷/۵ سال و درآمد کمتر از ۲ میلیون و ۲۰۰ هزارتومان دارند، با احتمال بیشتر متقاضی بیمه‌نامه ۱۲ درصدی خواهند بود اما برای زنان با سن و درآمد بیشتر احتمال خریداری بیمه ۱۴و۱۸ درصدی بیشتر خواهد بود. نتیجه مهمی که می‌توان گرفت این است که سابقه‌کار، سن و درآمد تاثیر مثبتی بر انتخاب بیمه‌نامه ۱۴ و ۱۸ درصدی با پوشش از کار افتادگی و فوت علاوه بر بازنشستگی توسط زنان دارند. در واقع، در سنین و سوابق بالا تقاضا برای بیمه‌نامه با خدمات بیشتر و در سنین و سوابق پایین تقاضا برای بیمه‌نامه با خدمات کمتر اما ارزان‌تر وجود دارد.

جدول٤.خلاصه نتایج برآورد درخت تصمیم برای زنان بیمه‌شده حرف و مشاغل آزاد

سابقه (سال)	سن	متوسط درآمد (میلیون ریال)	پیش‌بینی (درصد احتمال)

۰ تا ۵/۵	در هر سن	در هر سطحی از درآمد	بیمه‌نامه ۱۲ درصد (احتمال ۱۰۰)
۵/۵ تا ۷/۵	کوچکتر از ۳۷/۵	کمتر از ۲۲	بیمه‌نامه ۱۲ درصد (احتمال ۵۵)
		بیشتر از ۲۲	بیمه‌نامه ۱۴ و ۱۸ درصد (احتمال ۶۳)
	بزرگتر از ۳۷/۵	کمتر از ۲۱/۴۷	بیمه‌نامه ۱۴ و ۱۸ درصد (احتمال ۵۷)
		بیشتر از ۲۱/۴۷	بیمه‌نامه ۱۴ و ۱۸ درصد (احتمال ۷۹)
۷/۵ تا ۸/۵	کوچکتر از ۳۹/۵	کمتر از ۲۱/۹۷	بیمه‌نامه ۱۴ و ۱۸ درصد (احتمال ۶۱)
		بیشتر از ۲۱/۹۷	بیمه‌نامه ۱۴ و ۱۸ درصد (احتمال ۷۹)
	بین ۳۹/۵ تا ۵۹/۵	در هر سطحی از درآمد	بیمه‌نامه ۱۴ و ۱۸ درصد (احتمال ۶۵)
	بزرگتر از ۵۹/۵	در هر سطحی از درآمد	بیمه‌نامه ۱۲ درصد (احتمال ۵۹)
۸/۵ تا ۹/۵	کوچکتر از ۳۹/۵	در هر سطحی از درآمد	بیمه‌نامه ۱۴ و ۱۸ درصد (احتمال ۷۳)
	بزرگتر از ۳۹/۵	در هر سطحی از درآمد	بیمه‌نامه ۱۴ و ۱۸ درصد (احتمال ۸۷)
بیشتر از ۹/۵	در هر سن	در هر سطحی از درآمد	بیمه‌نامه ۱۴ و ۱۸ درصد (احتمال ۱۰۰)

منبع: یافته‌های پژوهش

برای مردان گرایش به انتخاب بیمه‌نامه‌های ۱۴ و ۱۸ درصدی با پوشش کامل بسیار سریع‌تر اتفاق می‌افتد. آنچنان که در جدول ۵ نشان داده شده خریداری بیمه‌نامه ۱۲ درصدی تنها به مردان با سابقه کمتر از ۴/۵ محدود بوده و در سطوح سابقه بالاتر از ۶/۵ مردان بطور کامل پیش‌بینی می‌شود که متقاضی بیمه‌نامه ۱۴ و ۱۸ درصدی باشند. برای مردان با سابقه کاری بین ۴/۵ تا ۶/۵ نیز طبق نتایج به دست آمده می‌توان دید که خریداری بیمه‌نامه ۱۲ درصدی به افراد جوان با سن کمتر از ۲۴/۵ محدود خواهد بود و با افزایش سن تقاضا برای بیمه‌نامه ۱۴ و ۱۸ درصدی بالاتر رفته و الگوریتم احتمال بیشتری برای انتخاب این بیمه‌نامه پیش‌بینی می‌کند. طبق نتایج بدست آمده، پیش‌بینی شده که مردان سنین بین ۲۴/۵ سال تا ۲۷/۵ به احتمال ۶۱ درصد متقاضی بیمه ۱۴ و ۱۸ درصد باشند و این احتمال برای سنین ۲۷/۵ سال تا ۳۰/۵ به ۷۰ درصد، برای سنین ۳۰/۵ سال تا ۳۳/۵ به ۷۴ درصد، برای سنین ۳۳/۵ سال تا ۳۶/۵ به ۸۰ درصد، برای سنین بین ۳۶/۵ سال تا ۳۹/۵

به ۸۴ درصد، برای سنین ۳۹/۵ سال تا ۴۹/۵ به ۸۹ درصد و برای سنین بیشتر از ۴۹/۵ سال نیز به بیش از ۹۳ درصد می‌رسد.

در تمامی این نتایج و با توجه به ساختار الگوریتم درخت تصمیم، می‌توان به کم اهمیت بودن درآمد افراد در تعیین نوع بیمه‌نامه آن‌ها اشاره کرد. اما به‌خوبی می‌توان دید که سه عامل دیگر یعنی سن، جنسیت و سابقه از توضیح دهندگی و قابلیت پیش‌بینی بالایی برخوردارند. به‌طور کلی، یافته‌ها نشان می‌دهند که برای افراد با سن و سابقه کم بیمه‌نامه ۱۲ درصد و برای افراد با سن و سابقه بالا بیمه‌نامه ۱۴ و ۱۸ درصد جذابیت بیشتری خواهد داشت.

جدول۵. خلاصه نتایج برآورد درخت تصمیم برای مردان بیمه‌شده حرف و مشاغل آزاد

سابقه (سال)	سن	متوسط درآمد (میلیون ریال)	پیش‌بینی (درصد احتمال)
۰ تا ۴/۵	در هر سن	در هر سطحی از درآمد	بیمه‌نامه ۱۲ درصد (احتمال ۱۰۰)
۴/۵ تا ۶/۵	کوچکتر از ۲۴/۵	در هر سطحی از درآمد	بیمه‌نامه ۱۲ درصد (احتمال ۵۵)
	۲۴/۵ تا ۲۷/۵	در هر سطحی از درآمد	بیمه‌نامه ۱۴ و ۱۸ درصد (احتمال ۶۱)
	۲۷/۵ تا ۳۰/۵	در هر سطحی از درآمد	بیمه‌نامه ۱۴ و ۱۸ درصد (احتمال ۷۰)
	۳۰/۵ تا ۳۳/۵	در هر سطحی از درآمد	بیمه‌نامه ۱۴ و ۱۸ درصد (احتمال ۷۴)
	۳۳/۵ تا ۳۶/۵	در هر سطحی از درآمد	بیمه‌نامه ۱۴ و ۱۸ درصد (احتمال ۸۰)
	۳۶/۵ تا ۳۹/۵	در هر سطحی از درآمد	بیمه‌نامه ۱۴ و ۱۸ درصد (احتمال ۸۴)
	۳۹/۵ تا ۴۹/۵	در هر سطحی از درآمد	بیمه‌نامه ۱۴ و ۱۸ درصد (احتمال ۸۹)
	بزرگتر از ۴۹/۵	در هر سطحی از درآمد	بیمه‌نامه ۱۲ درصد (احتمال ۹۳)

بیشتر از ۶/۵	در هر سن	در هر سطحی از درآمد	بیمه‌نامه ۱۴ و ۱۸ درصد (احتمال ۱۰۰)

منبع: یافته‌های پژوهش

۵. جمع‌بندی و نتیجه‌گیری

در این مطالعه سعی شد با استفاده از تجزیه و تحلیل داده‌های مربوط به بیمه‌شدگان حرف و مشاغل آزاد سازمان تامین اجتماعی در سال ۱۳۹۹، درکی از مشخصه‌های افراد این گروه حاصل شود. این درک و فهم از داده‌های مربوط به افراد بیمه‌شده می‌تواند در جهت ایجاد برنامه‌هایی به منظور افزایش نرخ نفوذ این نوع بیمه‌ها مفید باشد زیرا با توجه به نتایج حاصل شده می‌توان برای هر وضعیت یک استراتژی مخصوص را در پیش گرفت و با هدف قرار دادن گروه مد نظر به افزایش پوشش این نوع بیمه‌ها کمک کرد. در بین چهار مشخصه اصلی از افراد، نتایج بدست آمده نشان دادند که سه ویژگی سابقه کار، سن افراد و جنسیت به ترتیب بیشترین تاثیر را بر انتخاب بیمه‌نامه با نرخ‌های متفاوت دارند و پس از آن درآمد هم می‌تواند تا حدودی تاثیرگذار باشد.

مهمترین نتایج بدست آمده عبارتند از:

- جنسیت: از بررسی رابطه بین تعداد و جنسیت بیمه‌شدگان مشخص شد که مردان با ۵۵/۴ درصد بیشترین متقاضیان انواع بیمه‌نامه‌های حرف و مشاغل آزاد هستند که البته بیش از همه بیمه‌نامه با نرخ ۱۸ درصد پوشش دهنده ریسک ازکارافتادگی و فوت (مستمری بازماندگان) علاوه بر بازنشستگی برای آن‌ها جذابیت دارد. در مقابل اما زنان به طور عمده متقاضی بیمه‌های ۱۲ درصد هستند. برای توضیح و درک این نتیجه باید به این واقعیت توجه کرد که در ایران بار تکفل همسر و فرزندان بر عهده مردان است و به همین دلیل است که مردان بیشتر متقاضی بیمه‌های با پوشش بالاتر ۱۴ و ۱۸ درصدی هستند و در مقابل این بیمه‌نامه‌ها با خدمات کامل‌تر از جمله مستمری

بازماندگان جذابیت زیادی برای زنان ندارد چون بازماندگان زنان بیمه‌شده عموماً واجد شرایط دریافت مستمری بازماندگی نیستند. این نتیجه با نتایج مطالعه عبدی (۱۳۸۵) تطابق دارد که مردان را عمده مشتریان بیمه‌نامه‌های حرف و مشاغل آزاد و اختیاری معرفی می‌کند که انگیزه و اهداف اصلی آنان از بیمه به ترتیب بازنشستگی، درمان، ازکارافتادگی و فوت بوده است.

- سن: جوانان به طور عمده متقاضی بیمه‌های ۱۲ درصد هستند که این یعنی آن‌ها نگرانی از بابت ریسک فوت و مستمری بازماندگان ندارند و درنتیجه متقاضی آن نیستند. این امر تا حدودی به خصلت نزدیک‌بین بودن[1] افراد نیز برمی‌گردد. با این وجود مشاهده شد که افزایش سن خرید بیمه‌نامه‌های ۱۴ و ۱۸ درصدی را به‌طور ویژه برای مردان جذابتر می‌کند. مشاهده شد که مردان با سابقه کاری بالا و افزایش سن تمایل به استفاده از خدمات کامل بیمه‌ای یعنی نرخ ۱۴و۱۸ درصد دارند اما در بین زنان متغیر سن به طور عمده تاثیر چندانی بر تمایل آن‌ها ندارد و زنان با افزایش سن و در هر گروه سنی همچنان متقاضی بیمه ۱۲ درصد هستند. این یعنی چون در ایران طبق قوانین بازنشستگی همسر و فرزندان تحت تکفل مردان هستند، در نتیجه بیمه‌نامه ۱۴ و ۱۸ درصدی با پوشش مستمری بازماندگان و ریسک فوت با افزایش سن برای مردان جذابیت پیدا می‌کند اما برای زنان در هر گروه سنی جذابیت ندارد چراکه همانطور که گفته شد بازماندگان زنان بیمه‌شده عموماً واجد شرایط دریافت مستمری بازماندگی نخواهند بود.

با توجه به این نتایج می‌توان گفت که در ارتباط با مردان، اولویت سازمان در بیمه حرف و مشاغل آزاد باید برجسته سازی و تبلیغ خدمات بیمه‌نامه با نرخ ۱۸ درصد و در ارتباط با زنان باید بر تبلیغ خدمات بیمه‌نامه با نرخ ۱۲ درصد تمرکز کند. علاوه بر این موارد، نتایجی که از برآورد الگوریتم درخت تصمیم حاصل شد نشان می‌دهند که:

1. Myopic

- سابقه کاری: افراد با سابقه کاری کمتر از ۴/۵ سال با هر سطحی از درآمد، سن یا جنسیت متقاضی بیمه‌های ۱۲ درصد با خدمات مستمری بازنشستگی هستند، اما مردان زمانی که سابقه کاری آن‌ها از ۶/۵ سال بیشتر می‌شود و زنان وقتی سابقه کاری آن‌ها بالاتر از ۹/۵ سال است به سمت بیمه‌های ۱۴و۱۸ درصدی (خدمات کامل‌تر) روی می‌آورند.
- درآمد: طبق نتایج به‌دست آمده اگرچه متغیر درآمد به دلیل عدم اظهار مقادیر واقعی توسط بیمه‌شدگان چندان اثر گذار نبود، اما باز هم مشاهده شد که افراد با درآمد بالا متقاضی بیمه‌نامه ۱۴ و ۱۸ درصدی هستند و در درآمدهای پایین متقاضی بیمه با خدمات ارزان‌تر (۱۲درصد) هستند. البته اثر درآمد چندان تعیین کننده نیست و در بازه‌های بالا یا پایین سوابق یا سن، دیگر قدرت پیش‌بینی کنندگی ندارد و در هر سطحی از درآمد، در سنین و سوابق بالا تقاضا برای بیمه‌نامه با خدمات بیشتر و در سنین و سوابق پایین تقاضا برای بیمه‌نامه با خدمات کمتر اما ارزان‌تر مشاهده می‌شود.

دقت حاصل از الگوریتم درخت تصمیم و جنگل تصادفی به ۸۸ درصد رسید، یعنی به احتمال ۸۸ درصد به درستی می‌توان پیش‌بینی کرد که افراد با توجه به مشخصه‌هایی که دارند مانند سن، جنسیت، سابقه کار و درآمد، متقاضی بیمه حرف و مشاغل آزاد با خدمات مستمری بازنشستگی هستند یا متقاضی بیمه با خدمات مستمری بازنشستگی، از کار افتادگی و بازماندگان خواهند بود. در کل می‌توان گفت که دو الگوریتم به خوبی عمل کرده‌اند و دقت مشابهی در پیش‌بینی تقاضای بیمه‌نامه دارند.

نتایج این پژوهش، ازطریق ارائه تصویری از مشخصه‌های کلیدی بیمه‌شدگان حرف و مشاغل آزاد و انگیزه آن‌ها برای انتخاب هر یک از انواع این بیمه‌ها، می‌تواند به سازمان تأمین اجتماعی در گسترش پوشش بیمه و بهبود مدیریت ارتباط با این دسته از بیمه‌شدگان کمک کند. برای مثال، با توجه به انگیزه پایین زنان و جوانان برای انتخاب بیمه‌های با خدمات

گسترده، می‌توان ازطریق ارائه مشوق‌ها یا خدمات کوتاه‌مدت، جذابیت این نوع بیمه‌ها را برای این گروه از افراد افزایش داد.

۶. تقدیر و تشکر

نویسندگان وظیفه خود می دانند از کارشناسان سازمان تأمین اجتماعی که در اختیار قراردادن داده‌ها همکاری داشته‌اند و همچنین داوران محترم به سبب ارائه نقطه نظرات ارزشمند که به غنای مقاله کمک شایانی داشته است تشکروقدردانی نمایند.

منابع

- Abdi, A. (2006). Examining the issues and problems of insured persons and the free and optional jobs in interaction with the social security organization. *Social Security Quarterly, 8(1).* 255-282. (In Persian)
- Abdi, F., Khalili-Damghani, K., & Abolmakarem, S. (2017). Solving customer insurance coverage sales plan problem using a multi-stage data mining approach. *Kybernetes.1, 2-19.*
- Abdul-Rahman, S., Arifin, N. F. K., Hanafiah, M., & Mutalib, S. (2021). Customer Segmentation and Profiling for Life Insurance using K-Modes Clustering and Decision Tree Classifier. *International Journal of Advanced Computer Science and Applications (IJACSA), 12(9)*
- Azzone, M.; Barucci, E.; Mancayo, G.G.; Marazzina, D. (2022). A Machine Learning Model for Lapse Prediction in Life Insurance Contracts. *Expert Systems with Applications,* 191.
- Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). CRISP-DM 1.0: Step-by-step data mining guide. *SPSS inc, 9,13.*
- Chen, I. J., & Popovich, K. (2003). Understanding customer relationship management (CRM): People, process and technology. *Business process management journal. 672-688.*
- Chen, Y., & Hu, L. (2005). Study on data mining application in CRM system based on insurance trade. *In Proceedings of the ۷th international conference on Electronic commerce. 839-841.*
- Hurwitz, J., & Kirsch, D. (2018). Machine learning for dummies. *IBM Limited Edition, 75.*
- Hosseini, S. M. S., Maleki, A., & Gholamian, M. R. (2010). Cluster analysis using data mining approach to develop CRM methodology to assess the customer loyalty. *Expert Systems with Applications,37(7), 5259-5264.*
- Kracklauer, A. H., Mills, D. Q., & Seifert, D. (2004). Customer management as the origin of collaborative customer relationship management. *In Collaborative Customer Relationship Management (pp.3-6). Springer, Berlin, Heidelberg*
- Khalili-Damghani, K., Abdi, F., & Abolmakarem, S. (2019). Solving customer insurance coverage recommendation problem using a two-stage

clustering-classification model. *International Journal of Management Science and Engineering Management, 14(1), 9-19.*

- Kong, H.; Yun, W.; Joo, W.; Kim, J.H.; Kim, K.K.; Moon, I.C.; & Kim, W.C. (2022). Constructing a personalized recommender system for life insurance products with machine-learning techniques. *Intelligent Systems in Accounting, Finance and Management, 29 (4).* 242-253.
- Maimon, O. Z., & Rokach, L. (2014). Data mining with decision trees*: theory and applications (Vol. 81). World scientific*
- Mau, S., Pletikosa, I., & Wagner, J. (2018). Forecasting the next likely purchase events of insurance customers: A case study on the value of data-rich multichannel environments. *International Journal of Bank Marketing, 1123-1144.*
- Mollamohammadi, R., & Mostofi, M.R. (2014). The Factors Affecting the Success of the Social Security Organization in Paying Retirement Pension to Those Insured by Qom'First Branch of Social Security Organization. *Organizational Culture Management, 12(2),* 299-323. (In Persian)
- Motdin, N.; Nazarian, R.; Daman-Ksheideh, M., & Seifipour, R. (2021). Designing a Comparative Model of Bank Credit Risk Using Neural Network Models, Survival Probability Function and Support Vector Machine. *Journal of Economic Modeling Research, 11 (45),* 199-230. (In Persian)
- Motafakkerazad, M.A.; & Ghafarnejad Mehraban, A. (2011). Intelligent Modeling of Asymmetric Effects of Monetary Shocks on Output in Iran(Neural Network Application). *Journal of Economic Modeling Research, 2 (4),* 83-102. (In Persian)
- Najafi, A. (2019). Predictability of loyalty and separation of self-insurance Persons of Social Security Organization based on data mining method. *Social Security Quarterly, 15(1).* 88-109. (In Persian)
- Ngai, E. W., Xiu, L., & Chau, D. C. (2009). Application of data mining techniques in customer relationship management: A literature review and classification. *Expert systems with applications,36(2), 2592-2602.*
- Parmah, S.; Mardomdar, S., & Heidari, A. (2020). Macroeconomic Variables and Demand for Self-employment Insurance in the Social Security Organization. *Social Security Quarterly, 16(1).* 41-59. (In Persian)
- Rahman, S., Arefin, K. Z., Masud, S., Sultana, S., & Rahman, R. M. (2017, April). Analyzing Life Insurance Data with Different Classification

Techniques for Customers' Behavior Analysis. *In Asian Conference on Intelligent Information and Database Systems.15-25.*

- Rygielski, C., Wang, J. C., & Yen, D. C. (2002). Data mining techniques for customer relationship management. *Technology in society, 24(4), 483-502.*
- Severino, Matheus Kempa, and Yaohao Peng. "Machine learning algorithms for fraud prediction in property insurance: Empirical evidence using real-world microdata." *Machine Learning with Applications 5 (2021): 100074.*
- Shokohyar, S.; Rezaeian, A., & Boroufar, A. (2017). Identifying the customer behavior model in life insurance Sector using data mining. *Management Research in Iran, 20(4).* 65-94. (In Persian)
- Sullivan, W. (2017). Machine Learning For Beginners Guide Algorithms: Supervised & Unsupervsied Learning. Decision Tree & Random Forest Introduction. *Healthy Pragmatic Solutions Inc*
- Tanha, J., Van Someren, M., & Afsarmanesh, H. (2017). Semi-supervised self-training for decision tree classifiers. *International Journal of Machine Learning and Cybernetics, 8(1), 355-370.*
- Wirth, R., & Hipp, J. (2000, April). CRISP-DM: Towards a standard process model for data mining. *In Proceedings of the ٤th international conference on the practical applications of knowledge discovery and data mining. 29-40.*