Journal of Industrial Innovations 2(1) (2024) 56 – 67

Industrial Innovations

https://jii.araku.ac.ir

ISSN: 2980-8847

Applying Data Mining Techniques in order to assess the Violations of Life Insurance Customers

Hossein AmoozadKhalili [a*], Mirsaeid Hosseini Shirvani [b]

[a] Department of Industrial Engineering, Sari Branch, Islamic Azad University, Sari, Iran.
[b] Department of Computer Engineering, Sari Branch, Islamic Azad University, Sari, Iran.

Original Article

Use your device to scan and read the article online

Citation: AmoozadKhalili H, Hosseini Shirvani M. Applying Data Mining Techniques in order to assess the Violations of Life Insurance Customers. Industrial Innovations. 2024;2(1):56-67.

https://doi.org/10.61186/jii.2.1.56

KEYWORDS

Data mining;
Life insurance;
Fraud detection;
Decision tree;
Support Vector Machine (SVM).

ABSTRACT

The issue of fraud in insurance claims is one of the problems faced by insurance companies. Therefore, the issue of discovering such frauds in all types of insurances is one of the topics of interest for experts in various fields. Insurance fraud can be defined as taking damages from insurance companies by resorting to fraudulent means and documents. Losses caused through fraudulent activities affect the interests of insurers and potentially their financial stability. The current research uses data mining techniques to identify the fraudulent behavior of life insurance policyholders in insurance companies in order to identify the factors affecting these behaviors. The results of the article show that decision tree and support vector machine techniques are useful in identifying frauds and can be considered as the main center in business management to detect fraud. The results of the implementation of different methods on the studied dataset show the superiority of the neural network method over other methods. The neural network method has succeeded in classifying the desired classes in this research with an accuracy of 90.83, which is a good accuracy. Also, from the created decision tree, it is possible to detect frauds or the possibility of violations before issuing the insurance policy by using the data of the insurers under investigation, and if the violation is proven, it can be prevented from being issued.

Extended Abstract

1. Introduction

Insurance fraud can be defined as taking damages from insurance companies by resorting to fraudulent means and documents. Losses caused through fraudulent activities affect the interests of insurers and potentially their financial stability. The current research uses data mining techniques to identify the fraudulent behavior of life insurance policyholders in insurance companies in order to identify the factors affecting these behaviors. The results of the article show that decision tree and support vector machine techniques are useful in identifying frauds and can be considered as the main center in business management to detect fraud.

2. Modeling and formulation

The process of dividing the initial data set into training and testing data can be done with the help of cross-validation method so that there is more confidence in the results. In this type of validation, the data is divided into k subsets. From these k subsets, each time one is used for validation and another k-1 is used for training. This procedure is repeated k times and all data are used exactly once for training and once for validation. Finally, the average result of these k validation times is chosen as a final estimate. The x-mutual validation operator to apply the mutual evaluation method in Rapidminer can be used as shown in Figure 4.

* Corresponding author,
E-mail address: Amoozad92@yahoo.com
DOI: https://doi.org/10.61186/jii.2.1.56
Received: February 24, 2024; Received in revised form: May 31, 2024; Accepted: June 22, 2024.
Article type: Research Paper

3. Figures and Tables

Tree construction is done in top-down mode. It is during this phase that the tree is repeatedly split until all data items belong to the same category. It is laborious and computationally intensive to traverse the training dataset repeatedly. Tree pruning is a bottom-up mode and is used to improve the prediction and classification accuracy of the algorithm by minimizing the high proportion. High fitness in the decision tree algorithm leads to misclassification error Tree pruning requires less work compared to the tree growing step because the training dataset is scanned only once. The result of applying the decision tree method on the examined data set is shown in Figure 2.

accuracy: 89.17% +/- 7.50% (mikro: 89.17%)

	true false	true true	class precision
pred. false	84	5	94.38%
pred. true	8	23	74.19%
class recall	91.30%	82.14%	

Figure 2 The result of applying the decision tree method on the investigated data set

نوآوری‌های صنعتی، دوره ۲، شماره ۱، سال ۱۴۰۳، صفحات ۵۶ تا ۶۷

نوآوری‌های صنعتی

شاپا: ۸۸۴۷-۲۹۸۰

بکارگیری تکنیک‌های داده‌کاوی به‌منظور ارزیابی تخلفات مشتریان بیمه‌های زندگی

حسین عموزاد خلیلی [الف*]، میرسعید حسینی شیروانی [ب]

[الف] گروه مهندسی صنایع، دانشگاه آزاد اسلامی، واحد ساری، ساری، ایران. Amoozad92@yahoo.com

[ب] گروه مهندسی کامپیوتر، دانشگاه آزاد اسلامی، واحد ساری، ساری، ایران. Mirsaeid_hosseini@yahoo.com

واژگان کلیدی

داده‌کاوی؛
تقلبات بیمه‌ای؛
درخت تصمیم؛
ماشین بردار پشتیبان

تاریخ دریافت: ۱۴۰۲/۱۲/۰۵

تاریخ بازنگری: ۱۴۰۳/۰۳/۱۱

تاریخ پذیرش: ۱۴۰۳/۰۴/۰۲

چکیده

مساله بروز تقلب در ادعاهای بیمه یکی از مشکلات پیش روی شرکت‌های بیمه است. ازاین‌رو بحث کشف این‌گونه تقلبات در انواع بیمه‌ها یکی از موضوعات موردتوجه کارشناسان حوزه‌های مختلف است. کلاهبرداری بیمه‌ای را می‌توان گرفتن خسارت از شرکت‌های بیمه با توسل به وسایل و اسناد متقلبانه تعریف کرد. زیان‌هایی که از طریق فعالیت‌های کلاهبرداری ایجاد می‌شوند، منافع بیمه‌گران و به‌طور بالقوه ثبات مالی آنان را تحت‌تأثیر قرار می‌دهد. پژوهش حاضر، از تکنیک‌های داده‌کاوی در تشخیص رفتارهای متقلبانه بیمه‌گذاران بیمه‌های زندگی در شرکت‌های بیمه‌ای استفاده نموده تا عوامل مؤثر بر این رفتارها شناسایی گردد. نتایج مقاله نشان می‌دهد که تکنیک‌های درخت تصمیم و ماشین بردار پشتیبان، در شناسایی تقلبات صورت گرفته سودمند بوده و به‌عنوان مرکز اصلی در مدیریت کسب‌وکارها جهت تشخیص تقلب می‌تواند موردتوجه قرار گیرند. نتایج حاصل از پیاده‌سازی روش‌های مختلف بر روی مجموعه داده موردبررسی نشان‌دهنده برتری روش شبکه عصبی نسبت به سایر روش‌ها می‌باشد. روش شبکه عصبی با دقت ۹۰.۸۳ که دقت مناسبی است موفق به دسته‌بندی کلاس‌های موردنظر در این پژوهش شده است. همچنین از درخت تصمیم ایجادشده می‌توان جهت کشف تقلبات و یا احتمال تخلفات را قبل از صدور بیمه‌نامه با استفاده از داده‌های بیمه‌گذاران موردبررسی قرار داد و در صورت اثبات تخلف از صدور جلوگیری کرد.

۱- مقدمه

تمامی اموری که منجر به از بین رفتن باور و اعتماد عملکردی در بین ارکان صنعت بیمه گردد، تقلب در صنعت بیمه محسوب می‌شود. از بزرگترین نوع کلاهبرداری در بیمه می‌توان به ارائه اطلاعات نادرست اشاره کرد. شرکت‌های بیمه در سراسر جهان به‌طور مداوم منابع هنگفتی را در قالب جبران خسارت ادعاهای تقلبی از دست می‌دهند که در صورت صرف این منابع در فعالیت‌های تخصصی و سرمایه‌گذاری می‌توانستند خدمات مناسب‌تری به مشتریان خود ارائه نمایند. عوامل فروش می‌توانند داده‌کاوی را با اقدامات مربوط به کشف تقلب، ترکیب و دقت عملکرد خود را افزایش دهند که به این طریق نیاز به نیروی انسانی نیز کاهش می‌یابد [۱]. بدین معنی که بیمه‌گذاران اطلاعات نادرست به بیمه‌گر می‌دهند. مأخذ محاسبه تمامی قراردادهای بیمه و بیمه عمر، موضوع ریسک و برآورد میزان آن است که برحسب شدت آن بر سرنوشت قرارداد اثرگذار بوده و

* نویسنده مسئول؛
آدرس پست الکترونیک: Amoozad92@yahoo.com

ازاین‌رو بررسی میزان آن در قرارداد و تعیین حق بیمه بر اساس آن از اهمیت به سزایی برخوردار است. به‌بیان‌دیگر می‌توان کلاهبرداری در صنعت بیمه را گرفتن خسارت از شرکت‌های بیمه با استفاده از وسایل و اسناد متقلبانه تعریف کرد. میزان پول پرداخت‌شده ناشی از تقلب‌ها، مخارج شرکت‌های بیمه را افزایش می‌دهد و واضح است که منافع آنها را کاهش می‌دهد. درنتیجه ضریب خسارات، رشد می‌کند. افزایش ضریب خسارات توسط افزودن حق بیمه جبران می‌شود و منجر به کاهش علاقه مشتری به محصولات شرکت‌ها می‌شود. داده‌کاوی فرایند کشف الگوها در داده‌هاست. داده‌کاوی الگوهای حاوی اطلاعات را در داده‌های موجود جست‌وجو می‌کند. استفاده از مدل‌های ریاضی برای شناسایی تقلب، این امکان را به متخصصین شرکت‌های بیمه می‌دهد که با صرف زمان و هزینه کمتری تشخیص دهند که ادعای خسارت اعلام‌شده ازلحاظ آماری مشکوک به تقلب هست یا خیر.

بیمه‌گر صداقت و درستی اظهارات بیمه‌گذار را اساس کار خود قرار می‌دهد و بر اساس آن جبران خسارت را به عهده می‌گیرد. تقلب باعث تحمیل ضررهای مستقیم مالی به سازمان، تأثیر منفی بر خدمت‌رسانی به مشتریان، تأثیر منفی بر اعتبار و کارایی سازمان، سلب اعتماد عمومی نسبت به خدمات ارائه‌شده خواهد بود. کلاهبرداری بیمه‌ای و سوءاستفاده بیمه‌ای، نه‌تنها سودآوری بیمه گران را در معرض خطر قرار می‌دهد بلکه بر کل مجموعه صنعت بیمه تأثیر منفی می‌گذارد. از مصادیق بارز تقلب در بیمه‌های زندگی می‌توان به جعل درآمد و شغل، جعل امضا، جعل علت حادثه و فوت، جعل اطلاعات پزشکی و مرگ جعلی و عمدی نام برد. بعضی از کلاهبرداری‌ها در صنعت بیمه کاملاً آگاهانه و عمدی است. بیمه‌گذار ممکن است موجبات بروز خسارتی را فراهم آورد تا بدین طریق از محل بیمه‌نامه خود منفعتی کسب کند [۲]. یکی از راه‌های شناسایی تقلب در خسارت‌های اعلام‌شده، استفاده از اطلاعات تقلب‌های کشف‌شده در گذشته است. داده‌کاوی یک فناوری جدید قدرتمند برای کشف اطلاعات در حجم زیادی از داده‌ها است. داده‌کاوی زیرشاخه‌ای مهم در مدیریت دانش درنظر گرفته می‌شود. امروزه داده‌کاوی به سازمان‌های مختلف کمک می‌کند تا بر اطلاعات داده‌هایی که در مورد رفتار مشتریان خود جمع‌آوری کرده‌اند تمرکز کنند [۳]. حجم داده‌های تولیدشده در دنیا، هر دو سال یک بار، دو برابر می‌شود اما، تولید بیشتر داده، الزاماً به معنای تولید دانش بیشتر نیست. بیش از ۹۰ درصد داده‌های تولید شده در دنیا، داده‌های ساختارنیافته هستند. داده‌کاوی که به‌عنوان ابزاری برای کشف جرائم، ارزیابی میزان ریسک و فروش محصولات به کار می‌رود، دربرگیرنده ابزارهای تجزیه‌وتحلیل اطلاعات به‌منظور کشف الگوهای معتبر و ناشناخته در بین انبوهی از داده‌هاست.

۲- مرور ادبیات

سراسیا[1] و همکاران در سال (۲۰۲۲) [۴] پژوهش با عنوان "کشف تقلب در سیستم بیمه" را مورد بررسی قرارداده‌اند. در این مقاله به معرفی یک چارچوب مبتنی بر بلاک چین برای فعال کردن تراکنش‌های ایمن و تبادل داده‌ها در بین عوامل تعاملی مختلف در شبکه بیمه پرداخته شده و نشان داده شده است که چگونه می‌توان از این فناوری برای ایجاد سیستمی استفاده کرد که از انواع خاصی از تقلب در حوزه‌های خودرو، درمان تکمیلی و ادعاهای بیمه عمر و غیره جلوگیری کرد. جودیت سنتوس[2] و همکاران (۲۰۲۱) [۵] پژوهشی با عنوان "ابزارهای داده‌کاوی برتر در صنعت مراقبت‌های بهداشتی" انجام داده‌اند. مراقبت‌های بهداشتی به‌طور فزاینده‌ای چالش‌برانگیز شده‌اند و برای یافتن درمان نیاز به بازیابی اطلاعات از حجم زیادی داده‌های بسیار پیچیده دارند. در این مقاله به بررسی ابزارهای داده‌کاوی منبع باز رایج پرداخته‌شده که در آن معیارهای داده‌کاوی بر اساس الزامات کاربرد مراقبت‌های بهداشتی ارائه گردید که بهترین آنها با استفاده از معیارهای انتخاب پیشنهادی مشخص‌شده‌اند. این مطالعه نشان می‌دهد که رپیدماینر بیشترین پوشش نیازهای داده‌کاوی مراقبت‌های بهداشتی را ارائه می‌دهد. خالد[3] و همکاران (۲۰۲۱) [۶] در پژوهشی به "تشخیص تقلب مالی به‌وسیله داده‌کاوی در بازه زمانی ۲۰۰۹ تا ۲۰۱۹"

[1] Serasia
[2] Judith Santos
[3] Khaled

پرداختند. این مقاله یک بازنگری کلی از تحقیقات پیشرفته در تشخیص کلاهبرداری‌های مالی در این بازه زمانی و طبقه‌بندی آنها بر اساس انواع تقلب و فناوری داده‌کاوی را ارائه می‌دهد. سائل[1] و همکاران در سال (۲۰۱۹) [۷] در مقاله دیگری با مضمونی مشابه مقاله قبل عملکرد تکنیک‌های یادگیری ماشین در تشخیص کلاهبرداری‌های مالی را ارائه نمود. که مؤسسات مالی را مجبور می‌کند سیستم تشخیص خود را به‌طور مستمر بهبود ببخشند و یک فناوری پیش رفته در زمینه تشخیص تقلب و همچنین پیشگیری از آن را ارائه می‌دهد. کوچیتو وادیم[2] (۲۰۱۸) [۸] در پژوهشی به بررسی رویکردهای مختلف برای حل مشکلات داده‌کاوی پرداخته است. این مقاله به وظایف اصلی در تجزیه‌وتحلیل حجم زیادی از اطلاعات و مقایسه روش‌های حل آنها اختصاص داده‌شده است. به دلیل تنوع بسیار زیاد انواع داده‌ها و اشکال سازماندهی اطلاعات، داده‌های واقعی همیشه، با ابزارهای ماشین قابل تجزیه‌وتحلیل نیستند. وادیم مفهوم داده‌کاوی را در سه مورد داده‌کاوی، تجزیه‌وتحلیل داده‌ها و جمع‌آوری داده‌ها ترجمه کرده است. حبیب‌زاده و همکاران (۱۳۹۹) به تشخیص تجدیدنظرهای مالی با استفاده از داده‌کاوی پرداخته‌اند. مطالعات قبلی نشان داده است که اظهارات ناخواسته در مقیاس بزرگ می‌تواند به همان اندازه مضر باشد و اعتماد سرمایه‌گذاران را از بین ببرد. بنابراین جزء ضروریات است که به بیانات غیرعمدی قابل‌توجه نیز پرداخته شود. تکنیک‌های داده‌کاوی مورداستفاده در این مقاله شامل درخت تصمیم، شبکه‌های عصبی، بیز است. کاظمی و بهادر (۱۴۰۰) [۹] در مقاله‌ای به بررسی "ارائه یک مدل پیش‌بینی جهت شناسایی افراد مبتلا به دیابت با استفاده از درخت تصمیم" پرداخته‌اند. در این تحقیق از اطلاعات ۵۰۰ نفر از مراجعه‌کنندگان به مرکز بهداشت شهید بلندیان قزوین استفاده‌شده است. در این تحقیق با استفاده از روش‌های داده‌کاوی درخت تصمیم و شبکه عصبی و شبکه بیزین یک مدل پیش‌بینی‌شده انجام شده است؛ یافته‌ها: مدل درخت تصمیم بیش‌ترین دقت و شبکه بیزین کم‌ترین دقت را در تشخیص بیماران دیابت دارد و به‌تبع آن درخت تصمیم کم‌ترین خطا و شبکه بیزین بیشترین خطا را دارا است.

۳- مجموعه داده

مجموعه داده مورد استفاده در این پژوهش از یک شرکت بیمه‌ای داخلی جمع‌آوری‌شده است. این مجموعه داده متشکل از ۱۰۳ رکورد و ۲۶ ویژگی می‌باشد. در شکل‌های۱ و ۲ نمایی بخشی از این مجموعه داده نشان داده‌شده است. برای این مجموعه داده، یک ویژگی برچسب لحاظ شده است که بیانگر متقلب یا غیرنرمال بودن فرم بیمه‌گذار می‌باشد. برای افراد نرمال برچسب "صفر" و بر سایر افراد برچسب "یک" درنظر گرفته است. مبانی تعیین این برچسب‌ها بر اساس نظرات کارشناسان بیمه‌ای می‌باشد.

[1] N.Sael
[2] Kochetov Vadim

نسبت بیمه شده با بیمه گذار	محل زندگي	شغل	جنسیت	سن	محل تولد	کد رایانه (کد اصلي)
بيمه شده اصلي همان بيمه گذار	نواب	كارمند	مرد	?	قم	4269140
دختر	روستاي كربال	خانه دار	زن	20	زرقان	15436759
بيمه شده اصلي غير بيمه گذار/دختر	ابهر	كودك	زن	2	زنجان	29471110
بيمه شده اصلي غير بيمه گذار/همسر	ملكشهر	خانه دار	زن	47	اصفهان	30404508
بيمه شده اصلي همان بيمه گذار	زرندیه	مدير امور پخش و ...	مرد	24	ماموديه	30942582
دختر	بافت آباد	-	زن	3	تهران	32463028
بيمه شده اصلي همان بيمه گذار	تهرانسر	اپتوگرافي	مرد	44	تهران	33538566
دختر	خواجه عبدالله	دانشجو	زن	18	تهران	34315216
پدربزرگ	منظريه	-	زن	0	تهران	34587584
بيمه شده اصلي همان بيمه گذار	مشيريه	آرايشگر	زن	41	تهران	35665827
پسر	جنت آباد	دانش آموز	مرد	6	تهران	36741681
بيمه گذار	شهرک رزمندگان ن	راننده سدگين	مرد	41	اهواز	39507219
بيمه شده اصلي غير بيمه گذار/ مادر	?	دانش آموز	مرد	11	شهريار	39716282

شکل ۱ نمایی از مجموعه داده مورداستفاده (الف)

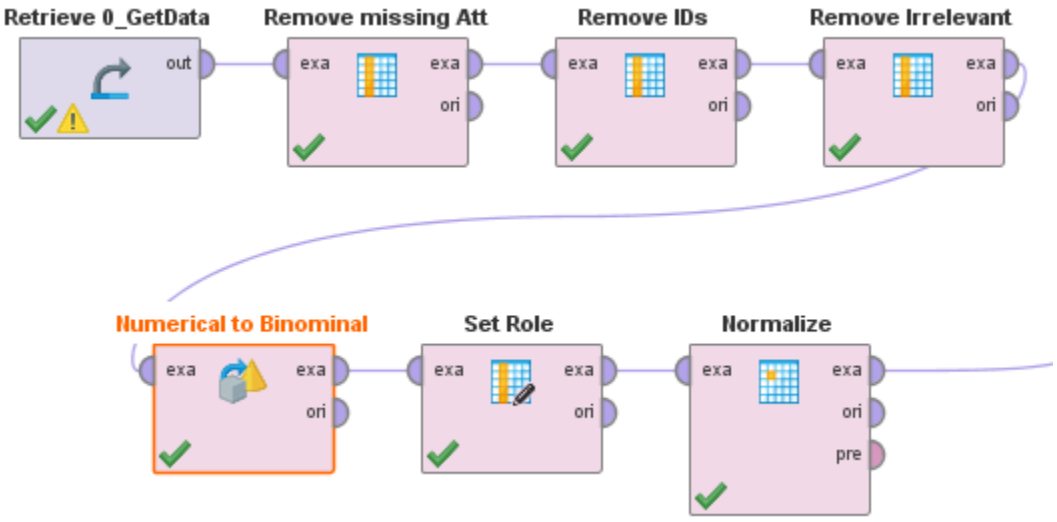
برای این مجموعه داده، یک ویژگی برچسب لحاظ شده است که بیانگر متقلب یا غیر نرمال بودن فرم بیمه‌گذار می‌باشد. برای افراد نرمال برچسب "صفر" و بر سایر افراد برچسب "یک" درنظر گرفته است. مبانی تعیین این برچسب‌ها بر اساس نظرات کارشناسان بیمه‌ای می‌باشد.

وضعيت بيمه نامه	حق بيمه سال اول	مدت زمان	روش پرداخت حق ...	نسبت استفاده كنندگان بيمه در شرايط فوت	نسبت استفاده كنندگان بيمه در شرايط حيات
به علت خودسوزي ...	3000000	20	اقساط سالانه	پسر 50 درصد / همسر 50 درصد	بيمه شده
فوت همراه با كتمان ...	3600000	30	اقساط ماهانه	بيمه گذار	بيمه گذار
فوت	6000000	25	اقساط دوماه	بيمه گذار	بيمه گذار
فوت	10680000	15	اقساط دوماهه	بيمه گذار	بيمه شده
خودكشي بيمه شده ...	10000000	20	اقساط سالانه	وراث قانوني	بيمه شده
فوت همراه با كتمان ...	6770000	30	اقساط دوماهه	بيمه گذار	بيمه گذار
جاري	30000000	30	اقساط سه ماهه	همسر 50درصد / پسر25درصد/دختر 25 ...	بيمه شده
جاري	20000000	30	اقساط سه ماه	مادر	بيمه شده
جاري	11000000	30	اقساط سالانه	50درصد مادر / 50درصد پدر	بيمه شده
جاري	20000000	30	اقساط سالانه	مادر	بيمه شده
جاري	50000000	30	اقساط سه ماهه	الويت اول بيمه گذار / الويت دوم مادر / ...	بيمه شده
جاري	20000000	30	اقساط سالانه	بيمه گذار	بيمه گذار
جاري	12000000	30	اقساط دو ماهه	بيمه گذار	بيمه گذار
جاري	30000000	30	اقساط دو ماهه	مادر	بيمه گذار

شکل ۲ نمایی از مجموعه داده مورداستفاده (ب)

۴- آماده‌سازی داده

در فرآیند آماده‌سازی داده، فعالیت‌های مختلفی می‌تواند صورت پذیرد. برای آماده‌سازی مجموعه داده موردنظر، فعالیت‌هایی نظیر حذف ویژگی‌های زائد، حذف ویژگی‌های دارای مقادیر گم شده، نرمال‌سازی و تبدیل ویژگی انجام‌شده است. شکل۳، فرآیندهای صورت گرفته برای آماده‌سازی داده‌ها در ابزار رپیدماینر نشان می‌دهد.



شکل ۳ فرآیندهای لازم برای اعمال قدم‌های لازم برای پیش‌پردازش در نرم‌افزار رپیدماینر

۵- مدل‌سازی

در این بخش به بیان روش‌های مدل‌سازی مورداستفاده در این مقاله به همراه بیان نتایج حاصل از اعمال آن‌ها پرداخته می‌شود. فرآیند تقسیم مجموعه داده اولیه به داده‌های آموزش و تست را می‌توان به کمک روش اعتبارسنجی متقابل[1] انجام داد تا نسبت به نتایج حاصل اطمینان بیشتری وجود داشته باشد. در این نوع اعتبارسنجی، داده‌ها به k زیرمجموعه افراز می‌شوند. از این k زیرمجموعه، هر بار یکی برای اعتبارسنجی و k-1 تای دیگر برای آموزش بکار می‌روند. این روال k بار تکرار می‌شود و همه داده‌ها دقیقاً یک بار برای آموزش و یک بار برای اعتبارسنجی بکار می‌روند. در نهایت میانگین نتیجه این k بار اعتبارسنجی به عنوان یک تخمین نهایی برگزیده می‌شود. البته می‌توان از روش‌های دیگر برای ترکیب نتایج استفاده کرد [ستایش و همکاران، ۱۳۹۳]. عملگر اعتبارسنجی x- متقابل برای اعمال روش ارزیابی متقابل در رپیدماینر را می‌توان به‌صورت شکل۴ استفاده نمود.

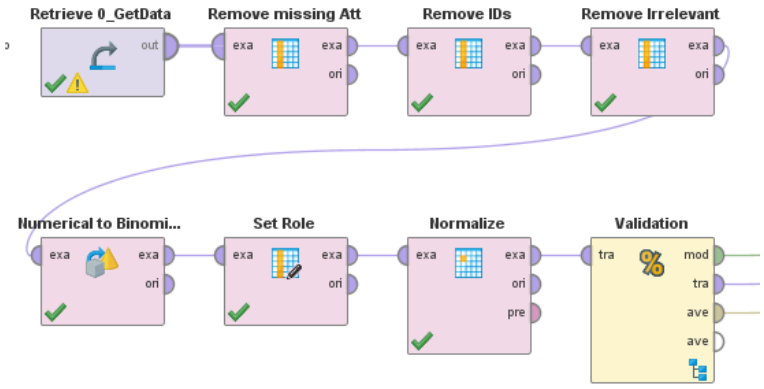
۶- درخت تصمیم

درخت‌های تصمیم یکی از پرکاربردترین نمایش‌ها برای طبقه‌بندی کننده‌ها هستند. حجم قابل توجه‌ای از مطالب به این فرم از طبقه‌بندی اختصاص دارند [۱۰]. الگوریتم درخت تصمیم یک تئوری مهم که در بسیاری از حوزه‌ها و رشته‌های کاربردی مزیت دارد [۱۱] و به‌طور مکرر یک مجموعه داده از اشیاء را با استفاده از روش حریصانه اول عمق یا روش اول عرض و وسعت، تجزیه و تقسیم می‌کند. تکنیک طبقه‌بندی درخت تصمیم در دو فاز اجرا می‌شود: ساخت درخت و هرس درخت.

ساخت درخت در حالت بالا به پایین انجام می‌شود. در طی این فاز یا مرحله است که درخت به‌طور مکرر تقسیم می‌شود تا زمانی که کل موارد داده به‌عنوان دسته یکسان تعلق داشته باشند. این‌که مجموعه داده آموزشی به‌طور مکرر پیمایش و قطع شود پرزحمت بوده و به لحاظ محاسباتی فشرده می‌باشد. هرس درخت یک حالت پایین به بالا است و در بهبود پیش‌بینی و دقت طبقه‌بندی الگوریتم به‌وسیله حداقل کردن تناسب زیاد به کار می‌رود. تناسب زیاد در الگوریتم درخت تصمیم به خطای طبقه‌بندی نادرست منجر می‌شود هرس درخت در مقایسه با مرحله رشد درخت به کار کمتری نیاز دارد زیرا مجموعه داده آموزشی فقط یک‌بار اسکن می‌شود [۱۲].

نتیجه اعمال روش درخت تصمیم بر روی مجموعه داده مورد بررسی در شکل۵ نشان داده شده است.

[1] Cross Validation

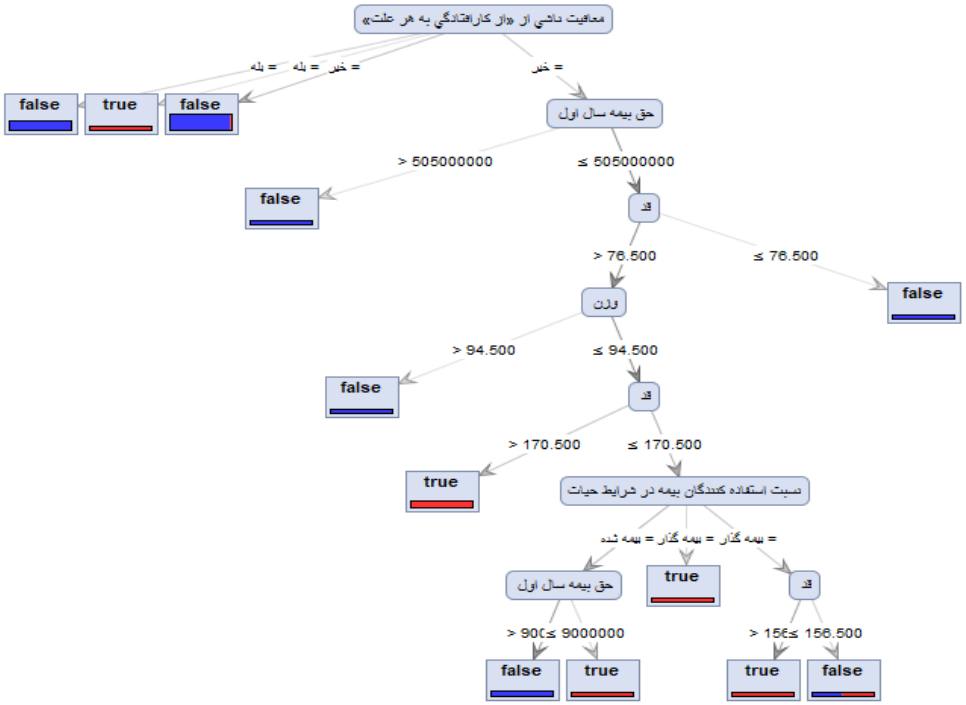


شکل ۴ اعمال روش ارزیابی متقابل

accuracy: 89.17% +/- 7.50% (mikro: 89.17%)

	true false	true true	class precision
pred. false	84	5	94.38%
pred. true	8	23	74.19%
class recall	91.30%	82.14%	

شکل ۵ نتیجه اعمال روش درخت تصمیم بر روی مجموعه داده مورد بررسی



شکل ۶ ساختار درخت تصمیم

با توجه به ساختار درخت تصمیم (شکل۶) می‌توان دید که از منظر این الگوریتم، مهم‌ترین ویژگی مجموعه داده مورد بررسی، ویژگی "معافیت ناشی ازکارافتادگی به هر علت" می‌باشد. دلیل این امر، قرار گرفتن این ویژگی در ریشه درخت می‌باشد. سایر ویژگی‌ها نیز به ترتیب اهمیت در سطوح بعدی قرار دارند. در برگ‌های درخت نیز وضعیت کلاس‌ها قابل‌مشاهده است. با مشاهده این درخت می‌توان دید که ویژگی‌هایی نظیر قد و وزن که به‌طور منطقی نمی‌توانند نقشی در موضوع تحقیق داشته باشند، در ساختار لحاظ شده‌اند. برای دستیابی به ساختار منطقی‌تر درخت تصمیم، می‌توان ویژگی‌های قد و وزن را از دیتای موردبررسی کنار گذاشت و ساختار جدید درخت تصمیم را تولید نمود. ساختار درخت تصمیم بدون درنظرگرفتن ویژگی‌های قد و وزن در شکل۷ نشان داده‌شده است.

۷- جنگل‌های تصادفی

جنگل تصادفی در واقع یک روش یادگیری تجمعی برای طبقه‌بندی، رگرسیون و وظایف دیگر می‌باشد. جنگل‌های تصادفی با ساخت تعداد زیادی درخت تصمیم در زمان آموزش عمل می‌کند و خروجی آن کلاس‌بندی است که می‌تواند به معنی ایجاد طبقات در نمونه‌ها (طبقه‌بندی) و یا به معنای پیش‌بینی نوع آن‌ها (رگرسیون) باشد. جنگل تصمیم تصادفی درواقع عادت پیش تناسب سازی[1] در درخت‌های تصمیم‌گیری را تصحیح می‌کند [۱۳].

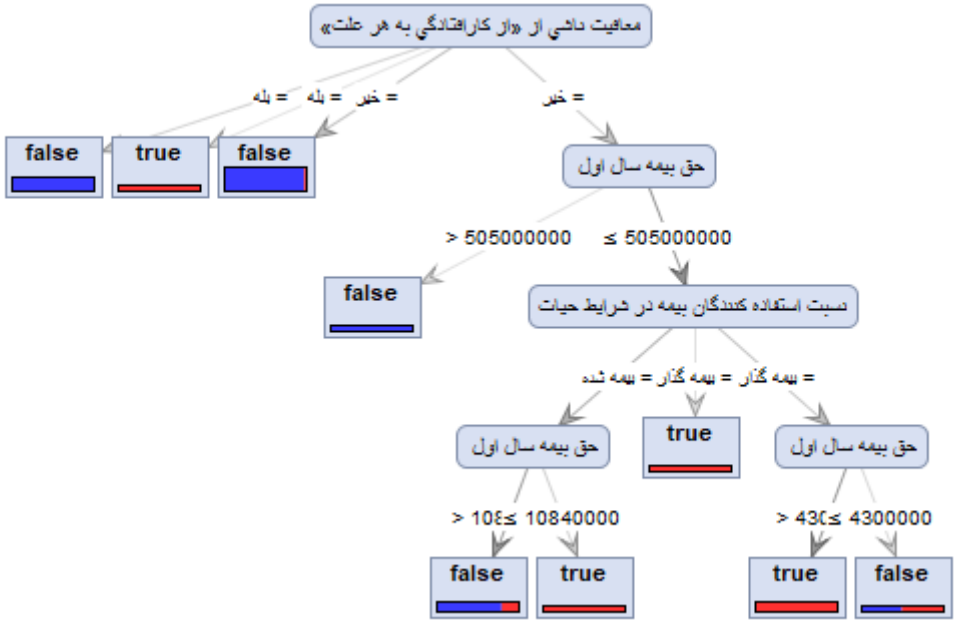
۸- شبکه عصبی

شبکه عصبی مصنوعی، مدل ساده‌شده‌ای از یک سیستم مرکزی است که ساختار مغز انسان را الگو قرار می‌دهد و با استفاده از ساختار محاسباتی پیچیده ارتباط درون نرون‌ها، توانایی واکنش به تغییرات و تطابق با محیط داده‌ها را به وجود می‌آورد. شبکه عصبی مصنوعی با پردازش داده‌های موجود، دانش یا قوانین نهفته در ورای داده‌ها را به ساختار شبکه منتقل می‌کند و بر اساس محاسبات بر روی داده‌های عددی یا مثال‌ها، قوانین کلی را فرا می‌گیرد. از مهمترین ویژگی‌های شبکه‌های عصبی، قابلیت یادگیری آنها است [۱۴].

۹- روش ماشین بردار پشتیبان

ماشین بردار پشتیبانی یکی از روش‌های یادگیری با نظارت است که از آن برای طبقه‌بندی خطی و غیرخطی و نیز رگرسیون چندبعدی استفاده می‌کنند. اساس کاری کلاس‌بندی کننده ماشین بردار پشتیبان کلاس‌بندی خطی داده‌ها است و باید در کلاس‌بندی خطی داده‌ها، خطی را انتخاب کنیم که حاشیه اطمینان بیشتری داشته باشد.

[1] Over-fitting



شکل ۷ ساختار درخت تصمیم بدون درنظر گرفتن ویژگی‌های قد و وزن

نتایج اعمال روش‌های درخت تصمیم، شبکه عصبی، ماشین بردار پشتیبان در شکل‌های ۸، ۹ و ۱۰ نمایش داده‌شده است.

accuracy: 85.00% +/- 6.24% (mikro: 85.00%)

	true false	true true	class precision
pred. false	91	17	84.26%
pred. true	1	11	91.67%
class recall	98.91%	39.29%	

شکل ۸ نتیجه‌ی اعمال روش درخت تصمیم بر روی مجموعه داده مورد بررسی

accuracy: 90.83% +/- 5.83% (mikro: 90.83%)

	true false	true true	class precision
pred. false	87	6	93.55%
pred. true	5	22	81.48%
class recall	94.57%	78.57%	

شکل ۹ نتیجه اعمال روش شبکه عصبی بر روی مجموعه داده مورد بررسی

accuracy: 88.33% +/- 6.67% (mikro: 88.33%)

	true false	true true	class precision
pred. false	88	10	89.80%
pred. true	4	18	81.82%
class recall	95.65%	64.29%	

شکل ۱۰ نتیجه ماشین بردار پشتیبان

۱۰- نتیجه‌گیری

با توجه به استفاده روزافزون از خدمات بیمه‌ای توسط طیف گسترده‌ای از مراجعان، چالش سو استفاده از خدمات بیمه‌ای نیز گسترش‌یافته است. شرکت‌های بیمه‌ای به‌طورمعمول، اطلاعات متعددی از مراجعان خود دریافت می‌نمایند و در پایگاه اطلاعات خود ذخیره می‌کنند. این حجم دیتای ذخیره‌شده درصورتی‌که به درستی موردبررسی و تجزیه‌وتحلیل قرار گیرد می‌تواند منجر به تولید دانش موردنیاز در زمینه‌های مختلف موردنیاز شرکت‌های بیمه‌ای شود.

جدول ۱ نتایج حاصل از اعمال روش‌های مختلف

نام روش	صحت
درخت تصمیم	۸۹.۱۷
جنگل تصادفی	۸۵.۰۰
شبکه عصبی	۹۰.۸۳
ماشین بردار پشتیبان	۸۸.۳۳

خلاصه نتایج حاصل از اعمال روش‌های مختلف با درنظر گرفتن معیار صحت در جدول۱ بیان‌شده است. نتایج حاصل از پیاده‌سازی روش‌های مختلف بر روی مجموعه داده موردبررسی نشان‌دهنده برتری روش شبکه عصبی نسبت به سایر روش‌ها می‌باشد. روش شبکه عصبی با دقت ۹۰.۸۳ که دقت مناسبی است موفق به دسته‌بندی کلاس‌های موردنظر در این پژوهش شده است. پژوهش حاضر، از تکنیک‌های داده‌کاوی در تشخیص رفتارهای متقلبانه بیمه‌گذاران بیمه‌های زندگی در شرکت‌های بیمه‌ای استفاده نموده تا عوامل مؤثر بر این رفتارها شناسایی گردد. نتایج مقاله نشان می‌دهد که تکنیک‌های درخت تصمیم و ماشین بردار پشتیبان، در شناسایی تقلبات صورت گرفته سودمند بوده و به‌عنوان مرکز اصلی در مدیریت کسب‌وکارها جهت تشخیص تقلب می‌تواند موردتوجه قرار گیرند. همچنین از درخت تصمیم ایجادشده می‌توان جهت کشف تقلبات و یا احتمال تخلفات را قبل از صدور بیمه‌نامه با استفاده از داده‌های بیمه‌گذاران موردبررسی قرارداد و در صورت اثبات تخلف از صدور جلوگیری کرد.

۱۱- منابع

[1] Gooderzi A, Tabatabai Manesh J. Evaluation of the risk of fraud in unemployment insurance benefits of the Social Security Organization. Insurance Journal. 2015;31:89-110. [In Persian]

[2] Firoozi M, Shokuri M, Kazemi L, Zahedi S. Identifying fraud in car insurance using data mining methods. Insurance Journal. 2013;26:103-28. [In Persian]

[3] Sharma M. Data mining: A literature survey. International Journal of Emerging Research in Management & Technology. 2014;3(2):21-37.

[4] Vyas S, Serasiya S. Fraud Detection in Insurance Claim System: A Review. 2022 Second International Conference on Artificial Intelligence and Smart Energy (ICAIS): IEEE; 2022. p. 922-7.

[5] Santos-Pereira J, Gruenwald L, Bernardino J. Top data mining tools for the healthcare industry. Journal of King Saud University-Computer and Information Sciences. 2022;34:4968-82.

[6] Al-Hashedi KG, Magalingam P. Financial fraud detection applying data mining techniques: A comprehensive review from 2009 to 2019. Computer Science Review. 2021;40:100402.

[7] Sadgali I, Sael N, Benabbou F. Performance of machine learning techniques in the detection of financial frauds. Procedia computer science. 2019;148:45-54.

[8] Vadim K. Overview of different approaches to solving problems of data mining. Procedia computer science. 2018;123:234-9.

[9] Kazemi A, Bahadur H. Presentation of a prediction model to identify people with diabetes using decision tree. Iranian Journal of Diabetes and Metabolism. 2022;21:151-64. [In Persian]

[10] Espejo PG, Ventura S, Herrera F. A survey on the application of genetic programming to classification. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews). 2010;40(2):121-44.

[11] Lin L-g, Shen P. An Algorithm of Multi-Variable Decision Tree Based on Genetic Programing. International Conference on Artificial Intelligence and Software Engineering. Phuket, Thailand. 2014. p. 314-8.

[12] Anyanwu MN, Shiva SG. Comparative analysis of serial decision tree classification algorithms. International Journal of Computer Science and Security. 2009;3:230-40.

[13] Ho TK. Random decision forests. Proceedings of 3rd international conference on document analysis and recognition: IEEE; 1995. p. 278-82.

[14] Zaranejad M, H. S. Forecasting the inflation rate in Iran's economy using dynamic artificial neural networks (time series perspective). Quantitative Economics Quarterly. 2008;6:145-67. [In Persian]