Iran J. Insur. Res., 9(1): 15-37, Winter 2020, Serial #31

Iranian Journal of Insurance Research
(IJIR)
Homepage: https://ijir.irc.ac.ir/?lang=en

ORIGINAL RESEARCH PAPER

The application of data mining using machine learning algorithms to investigate the impact of vehicle characteristics in predicting the risk of material damage in the field of third party insurance

M. Asghari Oskoei[1,*], F. Khanizadeh[2], A. Bahador[3]

[1] *Faculty of Mathematical and Computer Sciences, Allameh Tabatabai University, Tehran, Iran*
[2] *Insurance Research Institute and responsible for the specialized desk of algorithm design and machine learning, Tehran, Iran*
[3] *Insurance Research Institute and head of specialized car insurance desk, Tehran, Iran*

ARTICLE INFO

Article History
Received: 22 April 2020
Revised: 30 May 2020
Accepted: 03 October 2020

Keywords
Insurance Customer Classification; Decision Tree; Support Vector Machine; Naïve Bayes; Neural Networks.

ABSTRACT

Objective: Classifying the risk of policyholders based on observable characteristics can help insurance companies to reduce losses, identify customers more accurately, and prevent adverse selection in the insurance market. The purpose of this article is to examine the financial losses caused by third party insurance and to predict the risk of policyholders in the event of an accident.

Methodology: using decision tree algorithms, support vector machine, Naive Bayes and neural network; The hidden data patterns have been discovered in order to classify third party insurance policyholders. Also, the unbalanced distribution of data in two groups of damaged and undamaged causes an important challenge in the application of machine learning and data mining methods, which is considered in this article.

Findings: The data set belongs to one of the insurance companies and contains more than four hundred thousand samples registered in five years and includes four independent variables of car type, car group, license plate type and car age and a dependent and two-valued variable of financial damage. According to the obtained results, the best performance and prediction accuracy (with accuracy F1=0.72±0.01) is related to the decision tree model.

Conclusion: The impact of variables on the occurrence of damage in order of priority are: car type, license plate type, car age and car group. The evaluation results show that more data related to the driver's characteristics is needed for more accurate prediction of damage and high-risk customers.

***Corresponding Author:**
Email: *oskoei@atu.ac.ir*
DOI: 10.22056/ijir.2020.01.02

نشریه علمی پژوهشنامه بیمه دوره 9، شماره 1، زمستان 1398، شماره پیاپی 31، ص 15-37

نشریه علمی پژوهشنامه بیمه

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

مقاله علمی

کاربرد داده‌کاوی با استفاده از الگوریتم‌های یادگیری ماشین برای بررسی تاثیر ویژگی‌های خودرو در پیش‌بینی ریسک خسارت مالی در رشته بیمه شخص ثالث

محمدرضا اصغری اسکویی[1,*]، فربد خانی‌زاده[2]، آزاده بهادر[3]

[1] *دانشکده علوم ریاضی و رایانه، دانشگاه علامه طباطبایی، تهران، ایران*

[2] *پژوهشکده بیمه و مسئول میز تخصصی طراحی الگوریتم و یادگیری ماشین، تهران، ایران*

[3] *پژوهشکده بیمه و مسئول میز تخصصی بیمه‌های اتومبیل، تهران، ایران*

اطلاعات مقاله

تاریخ دریافت: 03 اردیبهشت 1399

تاریخ داوری: 10 خرداد 1399

تاریخ پذیرش: 12 مهر 1399

کلمات کلیدی

دسته‌بندی مشتریان

درخت تصمیم

ماشین بردار پشتیبان

نایو بیز و شبکه‌های عصبی

چکیده:

هدف: طبقه‌بندی ریسک بیمه‌گذاران بر مبنای ویژگی‌های قابل مشاهده می‌تواند به شرکت‌های بیمه جهت کاهش زیان، شناخت دقیق‌تر مشتریان و جلوگیری از وقوع انتخاب نامساعد در بازار بیمه کمک شایانی کند. هدف این مقاله، بررسی خسارت‌های مالی ایجاد شده در بیمه شخص ثالث و پیش‌بینی ریسک بیمه‌گذاران در احتمال وقوع حادثه می‌باشد.

روش‌شناسی: با استفاده از الگوریتم‌های درخت تصمیم، ماشین بردار پشتیبان، نایو بیز و شبکه عصبی؛ به کشف الگوهای پنهان داده‌ها، در راستای طبقه‌بندی بیمه‌گذاران بیمه شخص ثالث پرداخته شده است. همچنین توزیع نامتعادل داده‌ها در دو گروه خسارت‌دیده و خسارت‌ندیده سبب یک چالش مهم در کاربرد روش‌های یادگیری ماشین و داده‌کاوی است که در این مقاله مورد توجه قرار گرفته است.

یافته‌ها: مجموعه داده متعلق به یکی از شرکت‌های بیمه و حاوی بیش از چهارصد هزار نمونه ثبت شده در پنج سال و شامل چهار متغیر مستقل نوع خودرو، گروه خودرو، نوع پلاک و سن خودرو و یک متغیر وابسته و دو ارزشی خسارت مالی است. با توجه به نتایج بدست آمده بهترین کارکرد و دقت پیش‌بینی (با دقت F1=۰,۷۲±۰,۰۱) مربوط به مدل درخت تصمیم می‌باشد.

نتیجه‌گیری: میزان تاثیرگذاری متغیرها در وقوع خسارت به ترتیب اولویت عبارتنداز: نوع خودرو، نوع پلاک، سن خودرو و گروه خودرو. نتایج ارزیابی نشان می‌دهد برای پیش‌بینی دقیق‌تر خسارت و مشتریان پر ریسک به داده‌های بیشتری مرتبط با ویژگی‌های راننده نیاز می‌باشد.

***نویسنده مسئول:**

ایمیل: *oskoei@atu.ac.ir*

DOI: 10.22056/ijir.2020.01.02

مقدمه

بیمه یکی از اصلی‌ترین ابزارهای مدیریت ریسک است و شرکت‌های بیمه با قبول ریسک باعث ایجاد آرامش درجامعه می‌گردند. به همین خاطر لازم است که شرکت‌های بیمه به ابزارهای تحلیل ریسک قدرتمندی دسترسی داشته باشند تا بتوانند ریسک دریافتی را به خوبی مدیریت کنند. هر بیمه‌گذار یا هر مورد بیمه‌شده، سطح متفاوتی از ریسک را به شرکت بیمه تحمیل می‌نماید. برای اطمینان از اینکه هر بیمه‌گذار حق بیمه منصفانه‌ای را پرداخت می‌کند، بیمه‌گران سطح ریسک بیمه‌گذار را تعیین و او را در یکی از طبقات ریسک قرار می‌دهند که بالتبع هر چه ریسک بیشتر باشد، حق بیمه بیشتر خواهد بود.

در شرایط فعلی، ارزیابی ریسک در صنعت بیمه کشور، براساس تجربیات سایر کشورها صورت می‌گیرد و صنعت بیمه از فقدان الگوریتم‌ها و سیستم‌های خودکاری که با حساسیت قابل قبولی بتوانند میزان ریسک مشتریان مختلف را بررسی و ارزیابی کند رنج می‌برد. این موضوع در رشته بیمه شخص ثالث که یک بیمه اجباری است، شرایط را حادتر می‌کند و در همین راستا ارزیابی ریسک فرد در رشته بیمه شخص ثالث بسیار حائز اهمیت است. به همین دلیل استفاده از ابزارهای داده‌کاوی می‌تواند در سنجش و پیش‌بینی ریسک بیمه‌گذاران بسیار راه‌گشا باشد.

در صنعت بیمه، داده‌کاوی می‌تواند به شرکت‌ها جهت کسب مزیت تجاری کمک کند. به عنوان مثال با به‌کارگیری تکنیک‌های داده‌کاوی، شرکت‌ها می‌توانند با استفاده از داده‌ها در مورد الگوهای خرید و رفتار مشتری، به کسب دانش پرداخته و همچنین درک خود را برای کمک به کاهش تقلب، ارتقای بیمه‌گری و بالابردن مدیریت ریسک افزایش دهند. (حاجی‌حیدری و همکاران، 1390)

شناسایی مشتریان مستلزم تحلیل مشتریان هدف و دسته‌بندی کردن مشتریان است که منجر به یافتن گروه‌هایی از مشتریان سودآور براساس ویژگی‌های آن‌ها می‌شود. طبقه‌بندی ریسک در حقیقت به معنای گروه‌بندی مشتریان با خصوصیات ریسک مشابه است که احتمال بروز خسارت‌های مشابهی دارند. طبقه‌بندی ریسک بیمه‌گذاران بر مبنای ویژگی‌های قابل مشاهده می‌تواند به شرکت‌های بیمه جهت کاهش زیان، افزایش نرخ پوشش بیمه و جلوگیری از وقوع انتخاب نامساعد در بازار بیمه کمک شایانی کند. (حنفی‌زاده و رستخیز پایدار، 1390)

عدم دسته‌بندی مشتریان باعث شده که مشتریان کم‌ریسک‌تر، خسارات مالی مشتریان پُر ریسک را جبران کنند. از این رو تفاوت چندانی بین مشتریان پُر ریسک و کم‌ریسک وجود ندارد. در واقع در کشور ما به جای فرد، اتومبیل بیمه می‌شود و این امر موجب شده تا بیشتر شرکت‌های بیمه در زمینه بیمه اتومبیل، متحمل زیان شوند (ترکستانی و همکاران، 1395). با استفاده از ابزارهای دسته‌بندی به منظور جداسازی مشتریان و شناخت آن‌ها، می‌توان سیاست‌گذاری‌های مناسبی را برای آن‌ها در نظر گرفت (ایزدپرست، 1390). در این مقاله از روش‌های درخت تصمیم[1]، ماشین‌بردار پشتیبان[2]، نایو بیز (بیز ساده)[3] و همچنین شبکه عصبی[4]، برای تحلیل داده‌های بیمه شخص ثالث و طبقه‌بندی نمونه‌های خسارت دیده و خسارت نادیده استفاده نمودیم و دقت مدل‌ها در این چهار روش را مقایسه کردیم.

مروری بر پیشینه پژوهش

در کشور ما بیمه اتومبیل و به‌ویژه بیمه شخص ثالث از مهمترین رشته‌های بیمه‌ای است که سهم عمده‌ای را در پرتفوی صنعت بیمه به خود اختصاص داده است و از طرفی به دلیل داشتن ضریب خسارت بالا، توجه بیش از پیش به این رشته بیمه‌ای را ضروری می‌نماید. این نوع بیمه در اکثر کشورهای جهان، یکی از مهمترین نوع فعالیت بیمه‌ای محسوب می‌شود و دست‌کم در ایران، حدود نیمی از صنعت بیمه را در اختیار دارد. ضمن اینکه به دلیل وجود این سهم عمده، فرصت مناسبی برای کاوش اطلاعات و استخراج الگوهای ناشناخته جهت تصمیمات کلان در صنعت بیمه از این طریق فراهم می‌شود. (کریم‌زادگان مقدم و بهروان، 1394)

---

[1]. Decision Tree
[2]. Support Vector Machine
[3]. Naïve Bayes
[4]. Neural Networks

عدم توجه به سطح ریسک مشتریان از سوی شرکت‌های بیمه، موجب شده تا مشتریان با سطوح ریسکی متفاوت، حق بیمه یکسانی را پرداخت کنند که از طرفی موجب نارضایتی بیمه‌گذاران و از طرفی دیگر باعث افزایش روزافزون ضریب خسارت و زیان‌ده شدن در این رشته شده است. از نوآوری‌های این تحقیق به نسبت تحقیقات گذشته، می‌توان به تعداد حجم داده‌های استفاده‌شده (415.687) در این تحقیق اشاره نمود و همچنین بهره‌مندی از دو روش داده کاوی[1] و الگوریتم‌های یادگیری ماشین[2] که موجب ارائه ارزیابی دقیقی از میزان تأثیر مشخصات خودرو در ایجاد خسارت شده است. تاکنون در هیچ تحقیقی در داخل کشور خودروهای سواری براساس کیفیت‌شان مورد تحلیل قرار نگرفته بودند و اکثر تحقیقات داخلی در خصوص بیمه بدنه اتومبیل انجام شده و تحقیقات اندکی به بیمه شخص ثالث اختصاص یافته است. در ادامه به برخی پژوهش‌های مرتبط انجام شده در داخل و نیز در جدول(1) به پژوهش‌های خارجی اشاره می‌شود.

در پژوهش ترکستانی و همکاران (1395)، از شبکه عصبی در راستای پیش‌بینی میزان خسارت بالقوه بیمه‌گذاران و تعیین نرخ بهینه استفاده شده است و نتایج پژوهش نشان می‌دهد که مدل ارائه شده می‌تواند با دقت 91 درصد طبقه خسارتی را تخمین بزند و با دقت 87 درصد میزان خسارت بالقوه بیمه‌گذاران را پیش‌بینی کند. پژوهش کریم‌زادگان مقدم و بهروان (1394)، در خصوص تعرفه‌گذاری پویا در رشته بیمه شخص ثالث می‌باشد که در این مقاله از شبکه‌های عصبی، درخت تصمیم و خوشه‌بندی استفاده شده است که نتایج به‌دست آمده از مدل‌ها با استفاده از ماتریس آشفتگی[3] و نسبت خسارت، مورد اعتبارسنجی قرار گرفته که نتایج حاکی از امکان استفاده از روش ارائه‌شده در تعرفه‌گذاری پویا در خصوص بیمه شخص ثالث به صورتی کارآمد را نشان می‌دهد، به نحوی که نسبت خسارت، کاهش می‌یابد و ماتریس آشفتگی، صحت ارزیابی را نشان می‌دهد.

در پژوهش حاجی‌حیدری و همکاران (1390)، با در نظر گرفتن همزمان مشخصه‌های بیمه‌گذار و اتومبیل در رشته بیمه بدنه، چند مدل (درخت تصمیم، شبکه‌های عصبی، شبکه‌های بیزین، ماشین بردار پشتیبان، رگرسیون لجستیک و تحلیل تمایزی) را به منظور پیش‌بینی طبقه خسارتی بیمه‌گذاران مقایسه کردند. طبق نتایج این پژوهش، مدل‌های شبکه‌های عصبی و درخت تصمیم با حدود 82 درصد، بیشترین دقت را در پیش‌بینی داشتند. حنفی‌زاده و رستخیزپایدار (1390)، ابتدا عوامل موثر بر ایجاد خسارت در بدنه اتومبیل را در ایران بررسی کردند. پس از مشخص شدن عوامل با استفاده از شبکه‌های عصبی خودسازمان‌ده[4]، به خوشه‌بندی بیمه‌گذاران براساس ریسک بالقوه آن‌ها پرداختند. در پژوهشی دیگر، فتح‌نژاد و ایزدپرست (1390)، با استفاده از تکنیک خوشه‌بندی k-means و درخت تصمیم بیمه‌گذاران را خوشه‌بندی کردند و نتیجه گرفتند که علاوه بر مشخصات اتومبیل، مشخصات رفتاری مشتری نیز در پیش‌بینی سطح خسارت مشتریان بیمه بدنه اتومبیل تأثیرگذار است. دقت مدل‌های استفاده‌شده در این پژوهش حدود 60 درصد بوده است. در پژوهش اصغری‌اسکوئی و قاسم‌زاده (1395) و اصغری اسکوئی (1394) رویکرد انتخاب ویژگی[5] براساس الگوریتم تکاملی[6] و کاربرد شبکه عصبی برای پیش‌بینی سری زمانی به کار رفته است. در ادامه خلاصه‌ای از برخی تحقیقات خارجی صورت‌گرفته در حوزه مرتبط با این مقاله در جدول (1) قابل مشاهده می‌باشد.

1. Data Mining
2. Machine Learning
3. Confusion Matrix
4. Self-Organization Map (SOM)
5. Feature Selection
6. Evolutionary Algorithm

جدول 1: خلاصه‌ای از تحقیقات خارجی گذشته

مقالات خارجی				
ردیف	مدل استفاده‌شده	حوزه بیمه‌ای	هدف نهایی	نام پژوهش
1	درخت تصمیم و شبکه عصبی	بیمه اتومبیل	ارزیابی ریسک براساس مدل‌سازی پیش‌گویی‌کننده و تحلیل الگوی سطح ریسک	(Wuyu and Cerna, 2019)
2	درخت تصمیم، رگرسیون لجستیک و شبکه عصبی	بیمه اتومبیل	انتخاب بیمه‌گذاران مناسب با توجه به سطح ریسک آن‌ها	(Baecke and Bocca, 2017)
3	درخت تصمیم	بیمه اتومبیل	استفاده از درخت تصمیم به عنوان مدل پیش‌بینی برای خسارت‌های اتومبیل	(Frempong, Nicholas and Boateng 2017)
4	خوشه‌بندی، رگرسیون بردار پشتیبان[1] و رگرسیون لجستیک کرنل (هسته)[2]	بیمه اتومبیل	طبقه‌بندی ریسک و پیش‌بینی میزان خسارت در راستای محاسبه حق بیمه	(Kašćelan, et al., 2016)
5	شبکه عصبی	بیمه اتومبیل	مدل پیش‌بینی برای خسارت‌های بیمه اتومبیل با استفاده از شبکه‌های عصبی	(Yunos, Ali, Shamsyuddin and Ismail 2016)
6	GLM	بیمه اتومبیل	محاسبه حق بیمه اتومبیل با استفاده از روش GLM	(David, 2015)
7	درخت تصمیم	بیمه اتومبیل	طبقه‌بندی بیمه‌گذاران و تعمیم این مدل به بیمه‌گذاران جدید	(Thakur and Sing, 2013)

مبانی نظری پژوهش

در این بخش به بررسی مبانی نظری مرتبط با این مقاله می‌پردازیم:

داده‌کاوی

در صنعت بیمه، اطلاعات و استفاده از آن بسیار حائز اهمیت است، به‌طوری‌که می‌توان موفقیت در بیمه را در گرو توانایی شرکت‌ها در تبدیل داده‌های خام به اطلاعات کاربردی دانست.

داده‌کاوی موجب بهبود روند تصمیم‌گیری در یک سازمان از طریق استخراج اطلاعات مهم از داده‌های موجود و جستجو روابط و الگوهای پنهان و آشکار از مجموعه داده‌های جمع‌آوری‌شده توسط سازمان خواهد شد و نهایتاً به بهینه‌سازی تصمیمات کسب‌وکار، ارتباطات و بهبود رضایتمندی مشتریان کمک می‌کند(حنفی‌زاده و رستخیز پایدار، 1390). داده‌کاوی بر طبق تعریف موسسه سیستم تحلیل آماری، فرایند انتخاب، اکتشاف، مدل‌سازی و شفاف‌سازی الگوهای مفید و ناشناخته در حجم زیادی از داده می‌باشد.

[1]. Support Vector Regression (SVR)
[2]. Kernel Logistic Regression (KLR)

روش‌ها و تکنیک‌های داده‌کاوی

بر حسب اینکه در فرایند داده‌کاوی، استنتاج چه نوع دانشی از مجموعه آموزشی مورد نظر است، از روش‌های مختلف داده‌کاوی می‌توان بهره جست. به‌طور کلی الگوریتم‌های یادگیری ماشین از نظر شیوه یادگیری به دو دسته اصلی الگوریتم‌های یادگیری با نظارت و الگوریتم‌های یادگیری بدون نظارت تقسیم می‌شوند که به صورت مختصر اینجا معرفی می‌شوند.

یادگیری با نظارت

در این نوع از الگوریتم‌ها، با دو نوع از متغیرها سروکار داریم. نوع اول که متغیرهای مستقل نامیده می‌شوند، یک یا چند متغیر هستند که براساس مقادیر آن‌ها، متغیر دیگری را پیش‌بینی خواهیم نمود. نوع دوم هم متغیرهای وابسته یا هدف یا خروجی هستند که مقادیر آن‌ها را به کمک این الگوریتم‌های یادگیری با نظارت پیش‌بینی خواهیم نمود. برای این منظور باید تابعی ایجاد کنیم که ورودی‌ها (متغیرهای مستقل) را گرفته و خروجی مورد نظر (متغیر وابسته یا هدف) را تولید کند.

نمونه‌هایی از این الگوریتم‌ها عبارتند از رگرسیون، درخت‌های تصمیم، جنگل‌های تصادفی، N نزدیک‌ترین همسایه، نایو بیز، ماشین بردار پشتیبان، شبکه عصبی و ... مسائل یادگیری با نظارت، به دو گروه طبقه‌بندی[1] (برای پیش‌بینی پاسخ‌های گسسته) و رگرسیون[2] (برای پیش‌بینی پاسخ‌های پیوسته) تقسیم می‌شوند.

یادگیری بدون نظارت

در این نوع از الگوریتم‌ها، متغیر هدف نداریم و خروجی آلگوریتم براساس الگوی درون داده‌ها مشخص می‌شود. بهترین مثال برای این نوع از الگوریتم‌ها، خوشه‌بندی یک جمعیت با داشتن اطلاعات شخصی و خریدهای مشتریان می‌باشد که به صورت خودکار آن‌ها را به گروه‌های همسان و هم‌ارز تقسیم کنیم.

در این دسته از یادگیری، تنها ورودی (x) را داریم و خروجی از پیش تعیین شده نیست. در واقع اینجا ناظری وجود ندارد تا به الگوریتم در یادگیری کمک کند. هدف اصلی یادگیری بدون نظارت، مدل کردن توزیع داده می‌باشد تا بتوان اطلاعات بیشتری درباره داده را بدست آورد. برعکس یادگیری با نظارت، هیچ ناظری وجود ندارد و مدل مجبور است خودش ساختار مخفی داده بدون برچسب را پیدا کند. الگوریتم K-Means از این دسته هستند.

یادگیری عمیق

یکی از چالش‌های مهم در یادگیری ماشین، انتخاب بهینه ویژگی‌های مؤثر در فرایند یادگیری است. معمولاً ویژگی‌ها به صورت انتخاب مستقیم متغیرهای ورودی و یا به صورت ترکیب خطی یا غیر خطی آن‌ها حاصل می‌شوند. یادگیری عمیق فرایند انتخاب یا استخراج ویژگی را همزمان در طول فرایند یادگیری با هدف رسیدن به حداکثر کارآمدی و حداقل خطا آموزشی انجام می‌دهد. این نوع یادگیری از مباحث نوین و پرکاربرد در علوم کامپیوتر می‌باشد که قابلیت یادگیری الگوهای پیچیده را نیز دارد و به خاطر قدرت و دقت بالایشان در بسیاری از مسائل دنیای واقعی به کار گرفته شده‌است.

[1]. Classification
[2]. Regression

درخت تصمیم

درخت تصمیم، نوعی روش یادگیری با نظارت است که با کمک یک ساختار درختی نتایج دسته‌بندی را ارائه می‌دهد. در این درخت هر گره نشانگر یک آزمون برای یک تصمیم بر روی یک متغیر مستقل است و هر شاخه، خروجی آزمون را نمایش می‌دهد. برگ‌های درخت نیز نمایانگر تصمیم نهائی و کلاس‌ها است. به‌طور عادی، پیچیدگی یک درخت تصمیم با افزایش تعداد ویژگی‌ها افزایش می‌یابد. اگر چه در بعضی از شرایط، تنها تعداد کمی از ویژگی‌ها می‌توانند یک کلاس را تعیین کند و بقیه ویژگی‌ها کم‌تأثیر یا بی‌تأثیر می‌باشد (ایزدپرست، 1390).

ماشین بردار پشتیبان

ماشین بردار پشتیبان، روش به نسبت جدیدی در حوزه داده‌کاوی می‌باشد که در بسیاری از مسائل طبقه‌بندی به‌طور موفقیت‌آمیزی عمل کرده است. ماشین بردار پشتیبان، یک طبقه‌بندی‌کننده دوتایی است که با استفاده از نگاشت داده‌ها از فضای ورودی اصلی به فضایی با بعد بالاتر برای جداسازی آن‌ها عمل می‌کند. این مدل ابر صفحه‌ای را جستجو می‌کند که فاصله‌اش با داده‌های دو طبقه حداکثری است. با امکان تعریف مرزهای انعطاف‌پذیر قدرت تعمیم‌پذیری نسبت به داده‌های جدید را افزایش می‌دهد. ماشین بردار پشتیبان می‌تواند با استفاده از داده‌های آموزشی کمتر نسبت به روش‌های رقیب، مرزهای سیستم را با دقت مناسبی تخمین بزند، بدون آنکه تعمیم‌پذیری سیستم را مخدوش کند. (حاجی‌حیدری و همکاران، 1390)

نایو بیز

اغلب به عنوان یک راه‌کار ساده آماری برای دسته‌بندی و تشخیص برچسب اشیا یا نقاط از روش نایو بیز استفاده می‌شود. الگوریتم نایو بیز، مبتنی بر مشاهدات آماری و احتمال با فرض استقلال ویژگی‌ها نسبت به یکدیگر عمل می‌کنند. در بیشتر مدل‌های نایو بیز از روش حداکثرسازی تابع درست‌نمایی استفاده می‌شود. هر چند تکنیک نایو بیز دارای فرضیات محدود و قابل دسترس است؛ ولی به خوبی می‌تواند از عهده حل مسائل واقعی برآید. یکی از مزایای قابل توجه در الگوریتم نایو بیز، امکان برآورد پارامترهای مدل با اندازه نمونه کوچک به عنوان مجموعه «داده آموزشی» می‌باشد(Salma, et al. 2019).

شبکه‌های عصبی

شبکه‌های عصبی، ساختارهای شبکه‌ای بسیار سازمان‌یافته‌ای هستند و دارای سه نوع لایه می‌باشند؛ لایه‌های ورودی، لایه‌های خروجی و لایه‌های میانی (یا لایه‌های پنهان). هر کدام از گره‌ها (که به نام نرون شناخته می‌شود)، در لایه‌های پنهان و لایه‌های خروجی دارای یک کلاسه‌بند[1] هستند. نرون‌های ورودی، ابتدا اطلاعات ویژگی‌های شیء را دریافت و سپس به نرون‌های لایه پنهان ارسال می‌کنند. لایه پنهان این اطلاعات را پردازش کرده و نتایج را به لایه پنهان بعدی می‌فرستد. این پروسه ادامه می‌یابد تا اطلاعات به نرون‌های لایه خروجی برسد. در آنجا مقدار[2] به‌دست‌آمده، تعیین‌کننده احتمال دسته قرارگیری شی می‌باشد. مجموعه این پروسه به عنوان پیش‌انتشار[3] شناخته می‌شود. نمره به‌دست‌آمده در خروجی، نشانگر دسته‌ای است که مجموعه ورودی‌ها به آن تعلق دارند. به این نوع شبکه عصبی، پرسپترون چندلایه[4] یا MLP گفته می‌شود. خروجی یک نرون، از جمع وزن‌دار ورودی‌ها و اعمال تابع فعالیت نرون برآن حاصل می‌شود. ضرایب وزنی که به عنوان وزن[5] و بایاس[6] شناخته می‌شوند، در طول فرایند یادگیری شبکه، بر اساس مجموعه نمونه‌های آموزشی و آلگوریتم پس‌انتشار خطا تنظیم می‌شوند. قبل از

1. Classifier
2. Value
3. Forward Propagation
4. Multi-Layer Perceptron (MLP)
5. Weight
6. Bias

یادگیری، انتخاب پیکره‌بندی مناسب شبکه از جمله تعداد لایه‌ها، تعداد نرون در هر لایه و تابع فعالیت نرون‌ها که بستگی به نوع مسئله و پیچیدگی آن دارد، باید انجام شود.

از کاربردهای شبکه‌های عصبی می‌توان طبقه‌بندی، شناسایی و تشخیص الگو، پیش‌بینی سری‌های زمانی، بهینه‌سازی، سیستم‌های خبره و فازی، مسائل مالی، بیمه، امنیتی، بازار بورس و وسایل سرگرم‌کننده و ساخت وسایل صنعتی، پزشکی و امور حمل و نقل را نام برد. (عمرانی نوش‌آبادی، 1390)

روش‌شناسی پژوهش

این پژوهش یک پژوهش کاربردی می‌باشد و تلاش شده است تا با استفاده از الگوریتم‌های درخت تصمیم، ماشین بردار پشتیبان، نایو بیز و شبکه عصبی؛ به کشف الگوهای پنهان داده‌ها، در راستای طبقه‌بندی بیمه‌گذاران بیمه شخص ثالث پرداخته شود. مراحل اجرایی در این تحقیق به صورت زیر می‌باشد:

جمع‌آوری داده‌ها از پایگاه داده بیمه‌گذاران بیمه شخص ثالث یکی از شرکت‌های بیمه؛

پیش‌پردازش[1] و پالایش[2] داده‌ها و تعیین شاخص‌هایی برای تعریف طبقات ریسک بیمه‌گذاران؛

بررسی آماری و تقسیم داده‌ها به زیر مجموعه‌های متعادل و تصادفی در دو دسته داده‌های آزمایشی و داده‌های آموزشی؛

استخراج الگوها با استفاده از الگوریتم‌های درخت تصمیم، ماشین بردار پشتیبان، نایو بیز و شبکه عصبی و مقایسه نتایج حاصله از این الگوریتم‌ها؛

ارائه الگوی کشف‌شده از طبقه‌بندی بیمه‌گذاران و شناسائی ویژگی‌های تعیین‌کننده؛

ارزیابی نتایج طبقه‌بندی و اعتبارسنجی مدل.

برای ساخت مدل لازم است که ابتدا تکنیک مدل‌سازی انتخاب شود که در این مقاله چهار روش (درخت تصمیم، ماشین بردار پشتیبان، نایو بیز و شبکه عصبی) بررسی شده است و ابزار مورد استفاده در این مقاله کتابخانه سای کیت لرن[3] در زبان برنامه‌نویسی پایتون می‌باشد.

داده‌های پژوهش

جامعه آماری تحقیق شامل کلیه بیمه‌گذاران یکی از شرکت‌های بیمه در رشته بیمه شخص ثالث در بازه زمانی ابتدای سال 1392 تا انتهای سال 1396 می‌باشد. از این میان، تعدادی از بیمه‌گذاران خسارت دریافت کرده و تعدادی دیگر، خسارتی از این شرکت بیمه دریافت ننموده‌اند. در مجموع بیش از چهارصد هزار رکورد در بانک اطلاعاتی بیمه‌گذاران بیمه شخص ثالث ثبت شده بود که پس از مورد ارزیابی و تحلیل قرار گرفتن، متغیرهای موجود در این مجموعه در جدول (2) ارائه شده است.

شایان ذکر است که در متغیرهای اشاره شده در جدول فوق، دسته‌بندی‌های مربوط به گروه خودرو و نوع پلاک توسط شرکت بیمه صورت گرفته است. در مورد نوع خودرو، طبقه‌بندی ارائه شده با ایده تیم نویسندگان انجام شده که براساس طبقه‌بندی‌های موجود در بازار و یا میزان فراوانی نوع خودروی مدنظر در مجموعه داده‌ها می‌باشد. به عنوان مثال، به طور عرف، دسته‌بندی موتورسیکلت‌ها بر اساس میزان قدرت موتورسیکلت بر حسب سی‌سی صورت می‌گیرد. همچنین برای گروه اتوکار، اتوبوس و ون بیشترین فراوانی را در بین سایر موارد به خود اختصاص داده بودند.

[1]. Pre-processing
[2]. Cleaning
[3]. Sci-Kit Learn Library in Python

جدول 2: متغیرهای به‌کار رفته در مدل

متغیر		نام متغیر/ نوع متغیر		مقادیر متغیر
مشخصات اتومبیل	نوع خودرو	CarType/ متغیر مستقل	سواری/ Car	خودروهای ژاپنی، کره‌ای، آمریکایی و اروپایی/ High خودروهای ایرانی و چینی/ LM
			موتورسیکلت‌ها/ Motor	کمتر از 125 سی‌سی/ M125 بین 125 تا 200 سی‌سی/ M200 بین 200 تا 300 سی‌سی/ M300 سایر/ MotOther
			باركش/ Container	وانت/ Truck کامیون/ Lorry تریلر/ Trailer سایر/ ConOther
			اتوکار/ Autocar	اتوبوس/ Bus ون/ Van سایر/ AutOther
	گروه خودرو	Car Group/ متغیر مستقل		سواری/ Car موتورسیکلت‌ها/ Motor باركش/ Container اتوکار/ Autocar ماشین‌آلات‌کشاورزی/ Farming
	نوع پلاک	Plate Type/ متغیر مستقل		شخصی/ Private فاقد پلاک/ Lack عمومی/ Public دولتی/ Gov ترانزیت/ Transit گذر موقت/ Temp معلولان/ Disabled منطقه آزاد/ Free نظامی/ Military سیاسی/ Politics
مشخصات راننده	سن وسیله نقلیه	Car Age/ متغیر مستقل		از 0 تا 45 سال
	خسارت مالی	Property Damage/ متغیر وابسته		بیمه‌گذار خسارت مالی ندیده است: 0 بیمه‌گذار خسارت مالی دیده است: 1

پیش‌پردازش داده‌ها

پیش‌پردازش داده‌ها از گام‌های مهم فرایند داده‌کاوی است که میزان دقت نتایج به‌دست آمده، تا حد زیادی به اجرای درست آن بستگی دارد. یک تعریف ساده می‌تواند این باشد که پیش‌پردازش داده‌ها مجموعه عملیات و روش‌هایی است برای تبدیل داده‌های خام جمع‌آوری‌شده از منابع متنوع به اطلاعات پاک‌سازی‌شده‌ای که برای انجام تحلیل‌ها مناسب باشد.

در واقع کیفیت داده‌ها و اطلاعات مفیدی که از آن حاصل می‌شود؛ به‌طور مستقیم بر توانایی مدل برای یادگیری تأثیر می‌گذارد. بنابراین بسیار مهم است که ما داده‌های خود را قبل از ارائه به مدل، مورد پیش‌پردازش قرار دهیم. در همین راستا برای پیش‌پردازش داده‌ها، اقدامات به شرح زیر انجام شد:

حذف متغیرهای نامناسب

برخی از متغیرهای موجود در پایگاه داده مانند شماره بیمه‌نامه، تاریخ صدور بیمه‌نامه و... به دلیل بی‌ارتباط بودن با هدف پژوهش از مجموعه متغیرها حذف شدند.

تبدیل کلیه مقادیر به مقادیر عددی

از بین متغیرهای مستقل نوع خودرو، گروه خودرو، نوع پلاک و سن وسیله نقلیه؛ تنها متغیر سن وسیله نقلیه، متغیر عددی می‌باشد و سایر متغیرها، متغیر اسمی می‌باشند و لذا برای این متغیرهای اسمی، کدگذاری انجام شد.

بررسی صلاحیت داده‌ها جهت ورود به مدل نهایی

در بین حجم انبوهی از داده‌ها، تمام داده‌ها از کیفیت لازم برخوردار نبودند. لذا معیوب‌بودن داده‌ها از لحاظ خطاهای اندازه‌گیری بررسی شد؛ به عنوان مثال در متغیر سن خودرو، اعداد نامتعارفی (اعداد سه رقمی) وارد شده بود که رکوردها از لحاظ وجود داده‌های نامرتبط مورد بررسی قرار گرفت و رکوردهایی که قابلیت اصلاح را داشتند، اصلاح شده و در صورتی که این قابلیت را نداشتند، حذف شدند.

حذف رکوردهای ناقص

پیش‌پردازش در خصوص رکوردهای ناقص نیز صورت گرفت، به این معنا که چنانچه اکثر متغیرهای یک رکورد، گمشده باشند، آن رکورد حذف شده است که این موضوع به کاهش داده‌های نهایی منجر شد.

حذف داده‌های دارای نوفه[1] (نویز)

در برخی از متغیرها، یک داده به دو دسته متفاوت تعلق داشت؛ به‌عنوان مثال در برخی موارد خودرو وانت یک‌بار به عنوان سواری و یک‌بار به عنوان بارکش ثبت شده بود که تا جای امکان، این‌گونه اطلاعات دارای نویز از داده‌ها حذف گردید.

آمار توصیفی

از آمار توصیفی به منظور سازمان‌دهی، خلاصه‌کردن و توصیف اطلاعات استفاده می‌شود و معمولاً قبل از آنالیز داده‌ها، سازمان‌دهی داده‌ها، می‌تواند منجر به آشکارشدن نکات پنهان داده‌ها شود. لذا به‌همین منظور در این بخش، به بررسی آمار توصیفی مربوط به متغیرهای مستقل (نوع پلاک، گروه خودرو، نوع خودرو و سن وسیله نقلیه) و متغیر وابسته (خسارت مالی داشتن یا نداشتن) به شرح جدول (3) پرداخته شده است.

[1]. Noise

جدول 3: نوع پلاک

نوع پلاک	تعداد	درصد فراوانی	درصد خسارت‌دیده مالی
خصوصی	292149	٪70	٪87/60
فاقد پلاک	16676	٪4	٪70
عمومی	6707	٪2	٪4/14
دولتی	3750	٪1	٪1/66
ترانزیت	95575	٪23	٪2/35
عبور موقت	94	---	٪0/02
معلولان	80	---	٪0/03
منطقه آزاد	651	---	٪0/2
نظامی	1	---	---
سیاسی	4	---	---

مأخذ: یافته‌های پژوهش

همانطور که در جدول (3) مشاهده می‌شود متغیر «نوع پلاک» به ۱۰ زیر گروه تقسیم می‌شود و بیشترین و کمترین فراوانی به ترتیب متعلق به پلاک‌های خصوصی و نظامی می‌باشد. خودروهای با پلاک ترانزیت با وجود درصد فراوانی بالا پس از پلاک‌های خصوصی، لیکن با توجه به مقادیر ستون چهارم خسارت بالایی به بار نیاورده‌اند. به طور مشابه جداول (4) و (5) فهرست ویژگی‌ها و دسته‌بندی آن‌ها و آمار توصیفی متغیرهای «گروه خودرو» و«نوع خودرو» را نمایش می‌دهد.

جدول 4: گروه خودرو

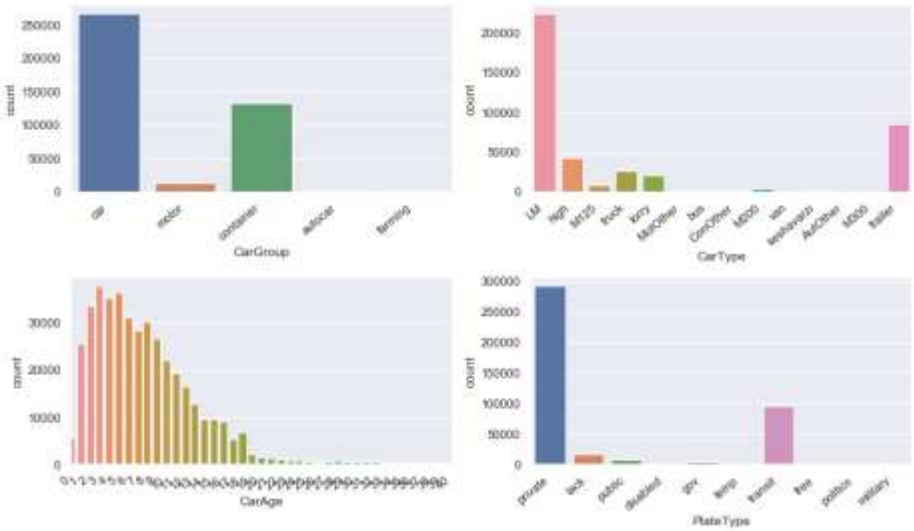
گروه خودرو	تعداد	درصد فراوانی	درصد خسارت‌دیده مالی
موتور سیکلت	13476	3	0/86
اتوکار	1307	---	0/27
بارکش	131998	32	18/25
سواری	267415	65	80/44
کشاورزی	1491	---	0/18

مأخذ: یافته‌های پژوهش

جدول 5: نوع خودرو

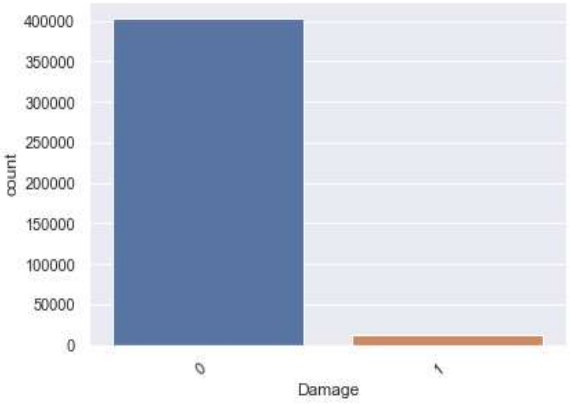
	نوع خودرو	تعداد	درصد فراوانی	درصد خسارت‌دیده مالی
موتورسیکلت‌ها	کمتر از 125 سی‌سی	8320	62	0/53
	بین 125 تا 200 سی‌سی	3440	26	0/25
	بین 200 تا 300 سی‌سی	236	2	0/02
	سایر	1479	10	0/06
اتوکار	اتوبوس	948	73	0/10
	ون	230	18	0/17
	سایر	113	9	---
باربری	وانت	26001	20	9/05
	کامیون	19910	15	6/82
	تریلر	85070	64	2/23
	سایر	987	1	0/13
سواری	ژاپنی، کره‌ای، آمریکایی و اروپایی	43184	16	11/18
	خودروهای ایرانی و چینی	224278	٪84	69/28
کشاورزی	کشاورزی	1491	100	0/18

در ادامه شکل (1) توزیع فراوانی متغیرهای مستقل را نمایش می‌دهد. شایان ذکر است که متغیر «سن خودرو»، تنها متغیر عددی داده‌های مورد بررسی می‌باشد. در شکل (1) سن خودرو، بر بازه صفر تا چهل سال (با افزایش سن خودرو به سمت چهل سال و بالای چهل سال فراوانی بسیار کم می باشد) نمایش داده شده است.



شکل ۱: توزیع فراوانی متغیرهای مستقل

همان‌طور که در شکل (1) مشاهده می‌شود در بین متغیرهای گروه خودرو، نوع خودرو و نوع پلاک، طبقه‌های سواری (Car)، خودروهای ایرانی و چینی (LM) و پلاک شخصی (Private) از بیشترین مقدار فراوانی برخوردار هستند. در مورد متغیر سن خودرو نیز هر چقدر به سمت خودروهای قدیمی‌تر حرکت می‌کنیم؛ از فراوانی آن‌ها کاسته می‌شود و خودروهای بالای ۳۵ سال، بیشتر شامل گروه خودروهای بارکش مانند کامیون و تریلر می‌باشند. متغیر خسارت شامل دو کلاس خسارت دیده (Damaged) و خسارت ندیده (Not Damaged) می‌باشد که به ترتیب با کدهای 1 و 0 نامگذاری شده‌اند. شکل (2) توزیع داده‌ها در دو کلاس فوق را نمایش می‌دهد.



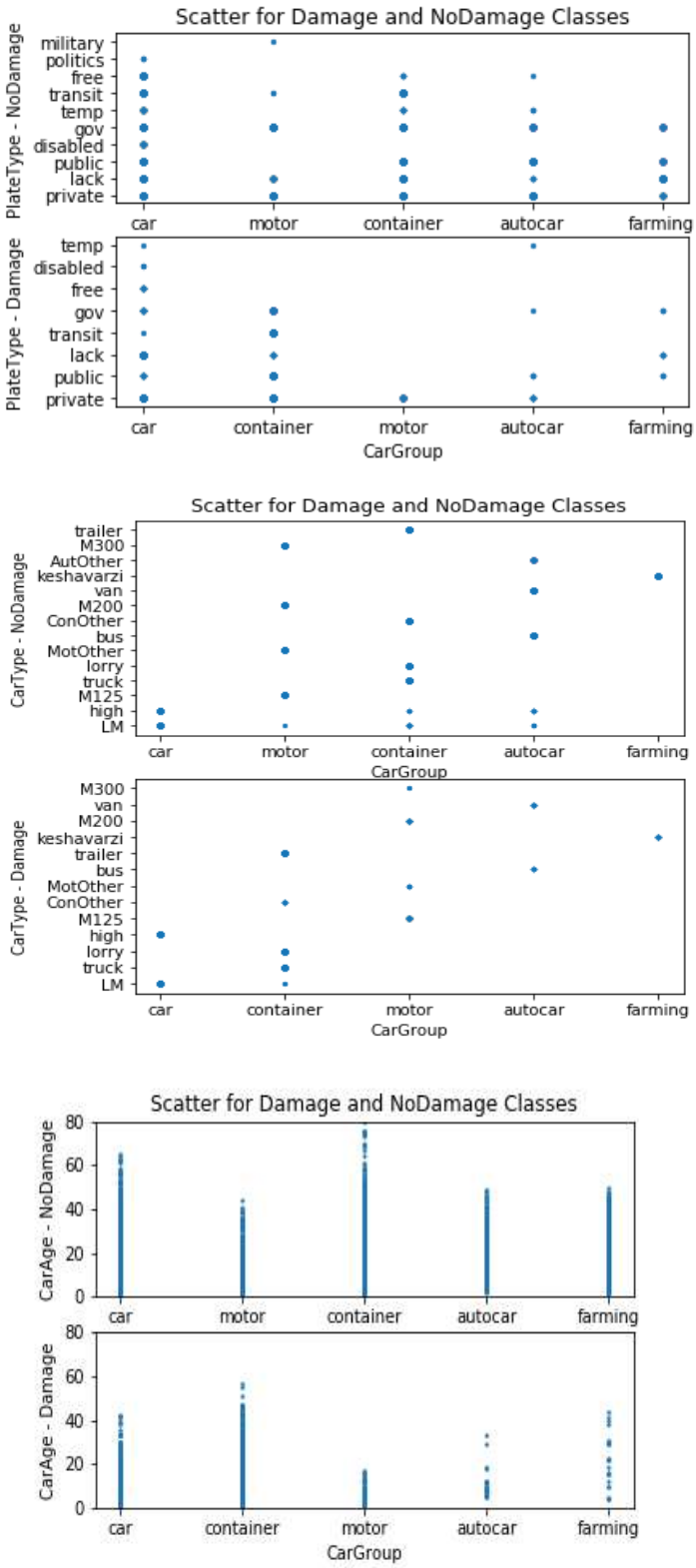
شکل 2: توزیع فراوانی متغیر وابسته

از مجموع 415687 نمونه، کلاس خسارت‌دیده‌ها (D=1) نمونه مثبت (Positive)، شامل 11641 نمونه و کلاس خسارت‌ندیده‌ها (D=0)که نمونه منفی (Negative) هستند، شامل 404046 نمونه است. به عبارتی نمونه‌های خسارت‌دیده 2/8 درصد و نمونه‌های خسارت ندیده 97/2 درصد از داده‌ها را شامل می‌شود. با توجه به اینکه نسبت توزیع دو کلاس 35:1 است، این مسئله یک مسئله طبقه‌بندی باینری نامتعادل[1] محسوب می‌شود.

این عدم تعادل در داده‌ها در رشته بیمه شخص ثالث، امری متداول و معمول می‌باشد. از طرفی تعداد خودروهایی که دچار خسارت نمی‌شوند به مراتب بیشتر از خودروهای خسارت دیده می‌باشند و از طرف دیگر بیمه‌گذاران برای خسارت‌های جزئی به شرکت مراجعه نمی‌کنند تا بتوانند از تخفیف عدم خسارت در سال‌های آتی استفاده نمایند.

شکل شماره (۳) نمودار پراکندگی[2] نمونه‌ها براساس گروه خودرو و دیگر ویژگی‌ها را به تفکیک دو کلاس مثبت و منفی نمایش می‌دهد. این نمودارها مشخص می‌کند در کدام گروه خودرو، نوع خودرو یا نوع پلاک یا سن خودرو؛ نمونه‌ای وجود دارد یا نمونه‌ای وجود ندارد. این نمودارها اطلاعاتی در مورد پراکندگی داده‌ها و عدم وجود نمونه در برخی گروه‌ها را مشخص می‌کند.

[1]. Imbalanced Binary Classification Problem
[2]. Scatter plot



شکل ۳: نمودارهای پراکندگی متغیرها

در انتهای این بخش، پراکندگی داده‌ها در فضای ویژگی مورد توجه قرار گرفته است. بررسی پراکندگی نمونه‌ها در فضای ویژگی چهاربعدی نشان می‌دهد که از مجموع 415687 نمونه داده، تنها ۱۵۱۱ نقطه یکتا[1] وجود دارد. منظور از نقاط یکتا سطرهایی از مجموعه داده می‌باشند که حداقل به وسیله ارزش مقداری یک متغیر مستقل از یکدیگر متمایز می‌شوند. دلیل اصلی این مشکل، تعداد کم ویژگی‌ها و نوع کدگذاری آن‌ها است. در واقع می‌توان نتیجه گرفت، برای آنکه به سطرهای یکتای بیشتری دست پیدا کنیم می‌بایست متغیرها و ویژگی‌های متنوع‌تری جمع‌آوری گردد. همچنین با توجه به متغیرهایی که در پایگاه داده‌ها در دسترس می‌باشد، این نتیجه حاصل می‌شود که مشخصات فردی راننده در تعیین سطح ریسک مشتریان و تمایز سطرهای داده بسیار تأثیرگذار می‌باشد که امید است با توجه به رویکرد قانون جدید بیمه شخص ثالث مبنی بر صدور بیمه‌نامه شخص ثالث براساس ویژگی‌های راننده (بهادر، استاد رمضان و خانی‌زاده، 1396)، در تحقیقات آتی برطرف گردد. در بخش بعد به چگونگی برخورد با این شرایط و مدل‌های استفاده‌شده در این وضعیت خواهیم پرداخت.

روش کار و مدل‌ها

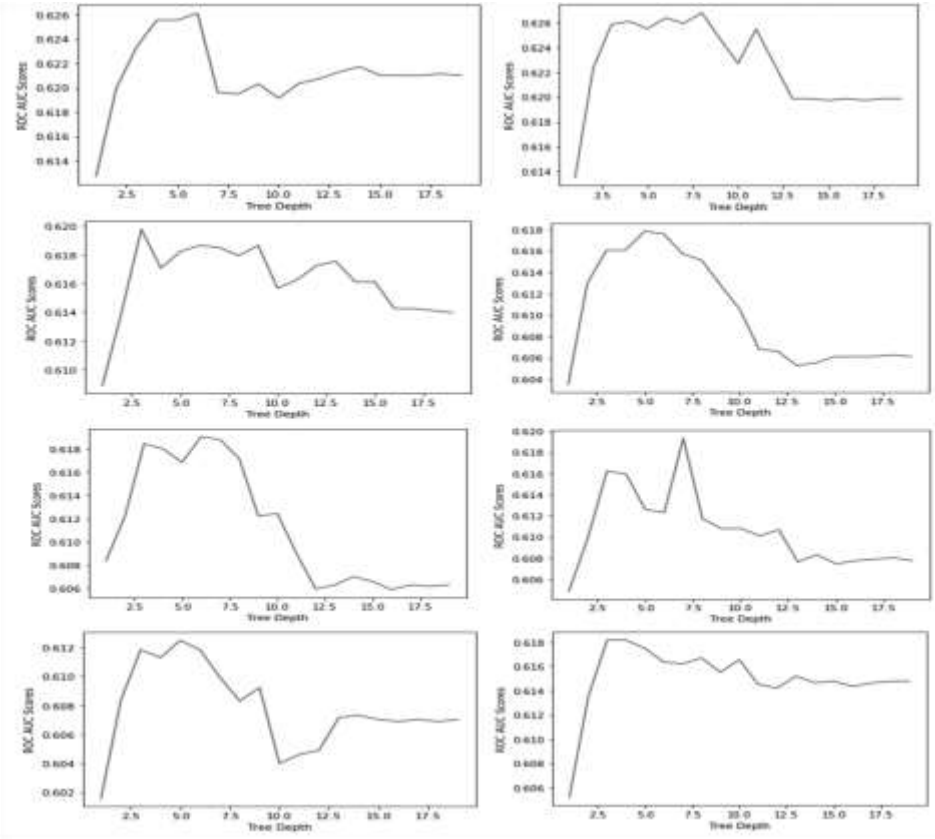
همان‌طور که در بخش قبل اشاره شد، در مجموعه داده‌های بیمه شخص ثالث، نسبت تعداد بیمه‌نامه‌هایی که منجر به پرداخت خسارت شده به موارد بدون خسارت در حدود 1 به 35 است. این نسبت آماری در صنعت بیمه طبیعی می‌باشد و به معنی توزیع نامتعادل داده‌ها[2] در دو کلاس مثبت (خسارت‌دیده مالی) و منفی (خسارت‌ندیده مالی) است. این پدیده در تحلیل داده سبب ایجاد خطای اریبی طبقه‌بندی می‌شود. به عبارتی، با توجه به اینکه تعداد نمونه مثبت، 2/8 درصد و تعداد نمونه منفی، 97/2 درصد داده‌ها است، چنانچه خروجی طبقه‌بندی به ازای تمام نمونه‌ها ثابت و منفی باشد، دقتی معادل 97/2 درصد خواهیم داشت که علی‌رغم دقت قابل توجه فاقد هرگونه ارزش عملیاتی است. در واقع در این حالت مدل، یادگیری خود را تنها براساس خروجی‌های مربوط به داده‌های خسارت‌ندیده انجام می‌دهد و داده‌های خسارت، نادیده گرفته می‌شوند.

یک روش شناخته‌شده برای تحلیل داده‌ها با توزیع نامتعادل، ارزیابی طبقه‌بندی داده‌ها در زیر مجموعه‌های متعادل[3] است که به صورت تصادفی از مجموعه اصلی نمونه‌برداری[4] شده است (Doucette and Heywood, 2008). در این روش با توجه عدم توازن بین نمونه‌های موجود در بین دو کلاس برچسب‌گذاری شده، از بین کلاسی که تعداد مشاهدات آن به میزان چشمگیری از کلاس دیگر بیشتر می‌باشد با روش نمونه‌گیری تصادفی، نمونه‌ای به تعداد رکوردهای موجود در کلاس دیگر به‌دست خواهیم آورد که با توجه به برابری مشاهدات در هر دو مجموعه نتایج تحلیل‌ها قابل اتکاء می‌باشند. در ادامه برای ساخت مجموعه داده متعادل، نمونه‌های مثبت به صورت ثابت و نمونه‌های منفی با تعداد مساوی به صورت تصادفی از مجموعه اصلی نمونه‌برداری می‌شود. این کار 50 بار تکرار شده و نتایج طبقه‌بندی به ازای هر تکرار ثبت می‌شود. نتایج، شامل ماتریس آشفتگی، دقت[5] (ACC)، نرخ تشخیص صحیح نمونه‌های مثبت[6] (TPR)، نرخ تشخیص صحیح نمونه‌های منفی[7] (TNR) و معیار اف وان (F1) است.

در تحقیق حاضر، از الگوریتم‌های طبقه‌بندی یادگیری ماشین و داده‌کاوی بهره گرفته‌ایم. برای طبقه‌بندی از درخت تصمیم، نایو بیز، شبکه عصبی و ماشین بردار پشتیبان استفاده شده و پارامترهای هریک به صورت جداگانه محاسبه و نتایج مقایسه شده است. در هر فرایند آموزش، مجموعه داده به صورت تصادفی به دو بخش داده‌های آموزشی[8] و داده‌های آزمون[9] با نسبت 70 به 30 تقسیم گردید. در حالت نامتعادل تعداد نمونه‌های آزمون 124707 نمونه و در حالت متعادل تعداد نمونه‌های آزمون 6985 نمونه می‌باشد.

---

1. Unique points
2. Imbalanced Data
3. Balanced subsamples
4. Subsampling
5. Accuracy
6. True Positive Rate (TPR)
7. True Negative Rate (TNR)
8. Training Set
9. Test Set



شکل 4: نمودار ROC-AUC مربوط به 8 نمونه‌گیری مختلف

لازم به ذکر است در استفاده از مدل درخت تصمیم و جهت تعیین عمق بهینه درخت تصمیم برای هر بار نمونه‌گیری نمودار ROC-AUC (Gajowniczek, et al., 2014), (Bowers and Zhou, 2019) نسبت به عمق درخت محاسبه و ترسیم می‌شود و میانگین مقادیر بهینه در ۵۰ نمونه‌گیری تصادفی محاسبه شده و به عنوان نتیجه نهایی در نظر گرفته می‌شود که طبق محاسبات انجام گرفته عدد ۶ به‌دست آمد. در شکل 4 نمودارهای ROC-AUC مربوط به ۸ نمونه‌گیری مختلف ترسیم شده است. در واقع منحنی ROC معیار ارزیابی برای سنجش کارایی مسائل طبقه‌بندی دودویی می‌باشد. این نمودار یک نمودار احتمال می‌باشد که از تقاطع TPR و FPR در مقادیر مختلف و برای یک آستانه مشخص از مدل طبقه‌بندی به‌دست می‌آید. معیار AUC خلاصه نتایج منحنی ROC را ارائه داده و برای اندازه‌گیری قابلیت مدل، جهت تشخیص بین دو کلاس استفاده می‌شود. برای مثال مقدار AUC=1 بیانگر این موضوع می‌باشد که مدل قادر است تمام نقاط قرار گرفته در کلاس مثبت و منفی را به طور دقیق و صحیح طبقه‌بندی کند. در ادامه به ارزیابی و بررسی مدل‌های استفاده‌شده می‌پردازیم و در نهایت نتایج به‌دست آمده از این تحقیق ارائه می‌گردد.

ارزیابی مدل‌ها

در ابتدا دو مدل درخت تصمیم و شبکه عصبی را بر روی حالتی که داده‌ها نامتعادل هستند بررسی می‌کنیم. برای شبکه عصبی از ۱۰ نرون برای لایه پنهان استفاده شده است. جدول 6 و 7 به ترتیب نتایج مربوط به درخت تصمیم و شبکه عصبی را نمایش می‌دهند. همانطور که در بخش قبل اشاره شد؛ ۷۰ درصد از داده‌ها به عنوان مجموعه آموزشی در نظر گرفته می‌شود. همان‌طور که قابل پیش‌بینی نیز بود؛ در حالتی که محاسبات بر روی داده‌های نامتعادل انجام گیرد، مدل از دقت بسیار بالایی (۹۷٪) برخوردار می‌باشد. لیکن در این شرایط معیار دقت (ACC) ابزار مناسبی برای تشخیص کارایی مدل محسوب نمی‌شود. در واقع همانطور که در جداول 6 و 7 مشاهده می‌شود؛ نرخ تشخیص صحیح نمونه‌های منفی ۱۰۰٪ می‌باشد ولی در مقابل، نرخ تشخیص صحیح نمونه‌های

مثبت صفر درصد است. نتیجه فوق بیانگر این موضوع است که مدل استفاده شده از نمونه‌های مثبت صرف‌نظر کرده و یادگیری را تنها براساس نمونه های منفی انجام داده است. معیار مناسبی که برای ارزیابی مدل برروی داده‌های نامتعادل و یا مدل‌هایی که FP[1] و FN [2] آن‌ها هم‌ارزش نیستند؛ قابل استناد می‌باشد، معیار F1 است که در جداول 6 و 7، صفر می‌باشد و بیانگر ضعف و قابلیت پایین مدل در پیش‌بینی داده‌های جدید می‌باشد. در واقع معیار F1 تابعی از صحت[3] و فراخوانی[4] می‌باشد. پس نیاز است برای درک بهتر معیار F1 ابتدا روابط مربوط به صحت و فراخوانی را مشاهده کنیم:

$$\text{Precision} = \frac{\text{True Positive (TP)}}{\text{True Positive+False Positive(FP)}} \qquad (1)$$

$$\text{Recall} = \frac{\text{True Positive (TP)}}{\text{True Positive} + \text{False Negative(FN)}} \qquad (2)$$

اگر به مخرج کسرها توجه کنیم، برای رابطه صحت تمام مقادیری که مثبت پیش‌بینی شده‌اند را خواهیم داشت و در مورد رابطه بازخوانی، تمام مقادیر مثبت واقعی به‌دست می‌آید. بنابراین روابط بالا می‌توان به صورت ذیل بازنویسی کرد:

$$\text{Precision} = \frac{\text{True Positive (TP)}}{\text{Total Predicted Positive}} \qquad (3)$$

$$\text{Recall} = \frac{\text{True Positive (TP)}}{\text{Total Actual Positive}} \qquad (4)$$

همان‌طور که از روابط بالا می‌توان برداشت کرد، معیار صحت زمانی استفاده می‌شود که برای محقق اهمیت دارد مشخص شود که چه تعداد از موارد مثبت پیش‌بینی شده، واقعا مثبت هستند و به درستی پیش‌بینی شده‌اند. کاربرد اصلی این معیار زمانی است که هزینه نتایج مثبت‌های کاذب (FP) بالا باشد. از طرفی دیگر معیار بازخوانی بیانگر تعداد نمونه‌هایی است که از بین کل نمونه‌های واقعی به درستی مثبت پیش‌بینی شده‌اند. به طور مشابه این معیار زمانی استفاده می‌شود که هزینه نتایج منفی‌های کاذب بالا باشد. در نهایت معیار F1 ابزاری است که بین معیارهای صحت و بازخوانی نوعی تعادل برقرار می‌کند و از رابطه زیر به‌دست می‌آید:

$$F1 = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \qquad (5)$$

جدول 6: طبقه‌بندی با درخت تصمیم (داده‌های نامتعادل)

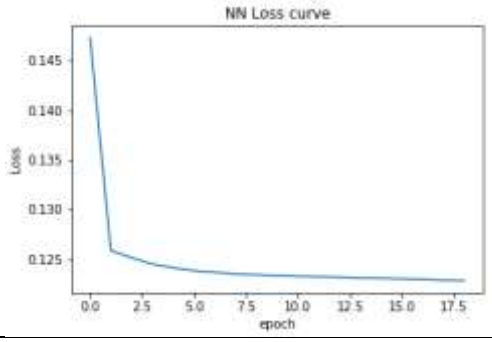
تعداد داده‌های آموزشی و آزمون	TestSet=12470 TrainSet=290980
ماتریس آشفتگی	$CM = \begin{bmatrix} TP = 0 & FP = 2 \\ FN = 3526 & TN = 121179 \end{bmatrix}$
معیار دقت	$ACC = \frac{TP + TN}{TP + FP + TN + FN} = 0.97$
نرخ تشخیص صحیح نمونه‌های مثبت	$TPR = \frac{TP}{TP + FN} = 0.00$
نرخ تشخیص صحیح نمونه‌های منفی	$TNR = \frac{TN}{FP + TN} = 1.00$
معیار F1	$F_1 = 0.00$

مأخذ: یافته‌های پژوهش

جدول 7: طبقه‌بندی با شبکه عصبی (داده‌های نامتعادل)

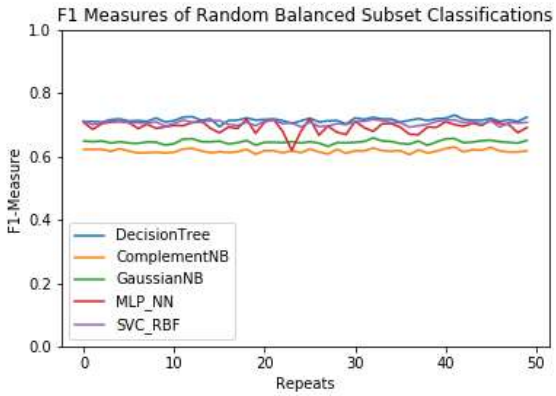
[1]. False Positive
[2]. False Negative
[3]. Precision
[4]. Recall

تعداد داده‌های آموزشی و آزمون	12470TestSet= 290980TrainSet=
ماتریس آشفتگی	$CM = \begin{bmatrix} TP = 0 & FP = 0 \\ FN = 3526 & TN = 121181 \end{bmatrix}$
معیار دقت	$ACC = 0.97$
نرخ تشخیص صحیح نمونه‌های مثبت	$TPR = 0.00$
نرخ تشخیص صحیح نمونه‌های منفی	$TNR = 1.00$
معیار F1	$F_1 = 0.00$



مأخذ: یافته‌های پژوهش

در ادامه برای رفع مشکل ناشی از نامتعادل بودن داده‌ها و بهبود نتایج، روش نمونه‌گیری از مجموعه داده‌ها استفاده شده است. لذا ابتدا با نمونه‌گیری تصادفی، تعداد ۵۰ مجموعه داده متعادل تولید کرده و مدل‌ها را روی هر مجموعه جداگانه محاسبه کرده و برآیند نتایج (منظور میانگین‌گیری روی اندازه‌ها برای 50 بار تکرار انجام دادید؟) را به عنوان خروجی نهایی در نظر می‌گیریم. در شکل (5) معیارF1 برای مدل‌های درخت تصمیم، نایو بیز، ماشین بردار پشتیبان و شبکه عصبی بازاء پنجاه تکرار قابل مشاهده می‌باشد.



شکل 5: معیارF1 برای 50 تکرار مدل‌ها روی داده‌های متعادل (نمونه‌گیری تصادفی)

چنانچه از نتایج شکل (5) مشخص می‌شود، با کاربرد داده‌های متعادل عملکرد مدل‌ها بهبود چشمگیری داشته و تقریبا تمام مدل‌ها نتایج نزدیک به نرخ 70 درصد دارند. همان‌طور که در شکل (5) مشاهده می‌شود درخت تصمیم نسبت به سایر مدل‌های طبقه‌بندی، کارایی بهتری را از خود نشان می‌دهد. در جداول (8) تا (11) نتایج مربوط به هر مدل (به ترتیب شبکه عصبی، ماشین بردار پشتیبان، درخت تصمیم، نایو بیز) به‌طور جداگانه ارائه گردیده است.

جدول 8: ارزیابی داده‌های متعادل‌شده (50 تکرار با نمونه‌های تصادفی)– شبکه عصبی

تعداد داده‌های آموزشی و آزمون	TestSet=6985 TrainSet=16297
ماتریس آشفتگی	$CM = \begin{bmatrix} TP = 3200 & FP = 2382 \\ FN = 311 & TN = 1092 \end{bmatrix}$
معیار دقت	$ACC = 0.61 \pm 0.01$
نرخ تشخیص صحیح نمونه‌های مثبت	$TPR = 0.91 \pm 0.03$
نرخ تشخیص صحیح نمونه‌های منفی	$TNR = 0.32 \pm 0.03$
معیار F1	$F_1 = 0.70 \pm 0.01$

مأخذ: یافته‌های پژوهش

جدول 9: ارزیابی داده‌های متعادل‌شده (50 تکرار با نمونه‌های تصادفی)–ماشین بردار پشتیبان

تعداد داده‌های آموزشی و آزمون	TestSet=6985 TrainSet=16297
ماتریس آشفتگی	$CM = \begin{bmatrix} TP = 3351 & FP = 2543 \\ FN = 119 & TN = 972 \end{bmatrix}$
معیار دقت	$ACC = 0.62 \pm 0.01$
نرخ تشخیص صحیح نمونه‌های مثبت	$TPR = 0.95 \pm 0.02$
نرخ تشخیص صحیح نمونه‌های منفی	$TNR = 0.29 \pm 0.02$
معیار F1	$F_1 = 0.71 \pm 0.01$

مأخذ: یافته‌های پژوهش

جدول 10: ارزیابی داده‌های متعادل‌شده (50 تکرار با نمونه‌های تصادفی)– درخت تصمیم

تعداد داده‌های آموزشی و آزمون	TestSet=6985 TrainSet=16297
ماتریس آشفتگی	$CM = \begin{bmatrix} TP = 3335 & FP = 2551 \\ FN = 143 & TN = 956 \end{bmatrix}$
معیار دقت	$ACC = 0.62 \pm 0.01$
نرخ تشخیص صحیح نمونه‌های مثبت	$TPR = 0.95 \pm 0.01$
نرخ تشخیص صحیح نمونه‌های منفی	$TNR = 0.29 \pm 0.02$
معیار F1	$F_1 = 0.72 \pm 0.01$

مأخذ: یافته‌های پژوهش

جدول 11: ارزیابی داده‌های متعادل‌شده (50 تکرار با نمونه‌های تصادفی)– نایو بیز گوسی

تعداد داده‌های آموزشی و آزمون	TestSet=6985 TrainSet=16297
ماتریس آشفتگی	$CM = \begin{bmatrix} TP = 2689 & FP = 2055 \\ FN = 830 & TN = 1411 \end{bmatrix}$

تعداد داده‌های آموزشی و آزمون	TestSet=6985 TrainSet=16297
معیار دقت	$ACC = 0.58 \pm 0.00$
نرخ تشخیص صحیح نمونه‌های مثبت	$TPR = 0.76 \pm 0.01$
نرخ تشخیص صحیح نمونه‌های منفی	$TNR = 0.41 \pm 0.01$
معیار F1	$F_1 = 0.65 \pm 0.00$

مأخذ: یافته‌های پژوهش

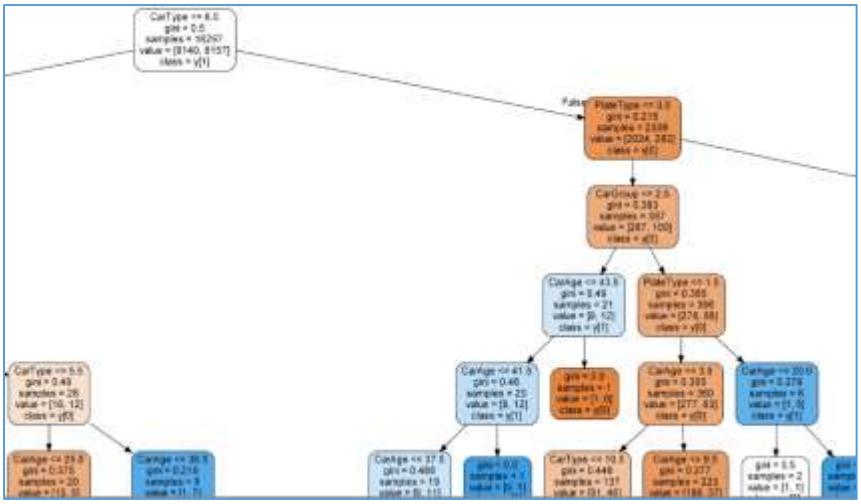
با توجه به جداول فوق ملاحظه می‌گردد که معیار دقت در هر چهار مدل کاهش پیدا کرده؛ لیکن نرخ تشخیص صحیح نمونه‌های مثبت در حال بهبود و افزایش است که نشانه بهبود مدل نیز می‌باشد. به طور کلی در مسائل طبقه‌بندی، پیش‌بینی درست نمونه‌های مثبت از ارزش زیادی برخوردار می‌باشد. همچنین همان‌طور که در بخش‌های قبل اشاره شد، معیار F1 یکی از ابزارهای مناسب جهت ارزیابی مدل‌های مورد استفاده می‌باشد. همانطور که در جداول (8) تا (11) دیده می‌شود در تمامی مدل‌ها، معیار F1 افزایش یافته و باعث بهبود مدل شده است. در این میان مدل درخت تصمیم از مقدار F1 بالاتری نسبت به سایر مدل‌ها برخوردار است. جدول (12) عملکرد چهار مدل را براساس معیار F1 نمایش می‌دهد.

جدول 12: مقایسه دقت مدل‌ها

مدل	معیار F1
نایو بیز	۰/۶۵ ± ۰/۰۰
شبکه عصبی	۰/۷۰ ± ۰/۰۱
ماشین بردار پشتیبان	۰/۷۱ ± ۰/۰۱
درخت تصمیم	۰/۷۲ ± ۰/۰۱

مأخذ: یافته‌های پژوهش

در انتها برای مطالعه تاثیر هر یک از ویژگی‌ها اعم از نوع خودرو، گروه خودرو، نوع پلاک و سن خودرو روی نتایج طبقه‌بندی از مدل درخت تصمیم استفاده شده است. در شکل (6) بخشی از نمودار درخت تصمیم نمایش داده شده است. مسیر درخت تصمیم از گره اول (ریشه) تا یکی از گره‌های برگ که نمایانگر یکی از قوانین موجود در درخت تصمیم می‌باشد مورد بررسی قرار می‌گیرد.



شکل 6: بخشی از ساختار درخت تصمیم این تحقیق

با توجه به نمودار شکل (6) اگر نوع خودرو غیر از خودروهای سواری، اتوبوس، ون، کامیون، وانت و سایر اتوکارها باشد؛ آنگاه برای ارزیابی دقیق‌تر ریسک می‌بایست نوع پلاک خودرو نیز مورد بررسی قرار گیرد. اگر نوع پلاک شامل یکی از دسته‌های «خصوصی»، «فاقد پلاک»، «عمومی» و یا «منطقه آزاد» باشد؛ آنگاه به سراغ متغیر گروه خودرو خواهیم رفت. در این حالت اگر گروه خودرو «سواری»، «اتوکار» و یا «بارکش» باشد، سن خودرو را مورد بررسی قرار می‌دهیم و در صورتی که سن خودرو کمتر از ۴۳ سال باشد، پیش‌بینی می‌شود که خودرو خسارت نمی‌بیند و بنابراین بیمه‌گذار و خودرو مورد نظر به طبقه کم‌ریسک تعلق داشته و باید حق بیمه کمتری از وی دریافت گردد. شایان ذکر است برای تعیین دقیق حق بیمه نیاز است تحقیق مشابهی بر روی خسارت‌های جانی شرکت بیمه انجام پذیرد و برآیند نتایج آن با خروجی مقاله فعلی می‌تواند نرخ کاربردی‌تری برای شرکت‌های بیمه ارائه دهد. جمع‌بندی میزان تاثیرگذاری متغیرها در وقوع خسارت مالی به ترتیب اولویت عبارتند از: نوع خودرو، نوع پلاک، سن خودرو و گروه خودرو.

جمع‌بندی و پیشنهادها

پژوهش فوق در صدد آن بود که با توجه به اهمیت زیاد رشته بیمه شخص ثالث در صنعت بیمه ایران، نتایجی در ارتباط با تأثیرگذاری متغیرهای موجود در احتمال وقوع خسارت تنها در نوع مالی و نه جانی ارائه دهد. در همین راستا از الگوریتم‌های یادگیری ماشین و داده‌کاوی جهت تحلیل داده‌ها بهره گرفته شد. مجموعه داده‌های موجود شامل ۴۱۵۶۸۷ نمونه به همراه پنج متغیر بود که متغیرهای مستقل عبارتند از نوع پلاک، گروه خودرو، نوع خودرو و سن خودرو و متغیر وابسته شامل یک متغیر دو ارزشی (باینری) با مقدار صفر برای خودروهای خسارت ندیده (نمونه منفی) و مقدار یک برای خودروهای خسارت دیده (نمونه مثبت) می‌باشد. در میان متغیرهای مستقل به غیر از سن خودرو که یک متغیر عددی می‌باشد، از نوع اسمی هستند.

در این مقاله سعی شد از اکثر الگوریتم‌های طبقه‌بندی شناخته‌شده یادگیری ماشین استفاده شود و از این لحاظ، پژوهش حاضر دارای جامعیت قابل قبولی می‌باشد. نتایج، یافته‌ها و پیشنهادات پژوهش را می‌توان به شرح زیر خلاصه نمود:

با بررسی داده‌ها مشخص گردید که از میان ۴۱۵۶۸۷ نمونه تنها ۱۵۱۱ نمونه یکتا وجود دارد. براساس این نتیجه، می‌توان پیش‌بینی کرد که متغیرهای بیشتری در محاسبه احتمال وقوع خسارت تأثیرگذارند. همچنین با توجه به نوع داده‌های در دسترس که تنها شامل ویژگی‌های خودرو می‌باشد؛ این نتیجه حاصل می‌شود که ویژگی‌های شخصی راننده می‌تواند در افزایش نقاط و نمونه‌های یکتا بسیار تعیین‌کننده باشد. شایان ذکر است از آنجاکه در سال‌های گذشته بیمه‌نامه‌های شخص ثالث تنها براساس ویژگی‌های خودرو صادر می‌گردید؛ لذا دسترسی به اطلاعات فردی رانندگان بسیار دشوار و غیر قابل اتکا می‌باشد. با این حال تیم نویسنده این مقاله، در حال جمع‌آوری این اطلاعات از طریق

تجمیع داده‌های برخی شرکت‌های بیمه می‌باشد و امیدوار است در آینده‌ای نزدیک تحقیقی در همین راستا و بر روی متغیرهایی شامل اطلاعات راننده انجام شود که مکمل مقاله فوق نیز خواهد بود.

با مرور نتایج ارزیابی مدل‌های استفاده شده در تحقیق، مشاهده می‌شود که مدل درخت تصمیم از کارایی بهتر و قدرت پیش‌بینی بالاتری برخوردار می‌باشد ($F1=0.72 \pm 0.01$). این امر می‌تواند نتیجه‌ای مثبت برای فعالان صنعت بیمه باشد؛ چرا که درخت تصمیم هم از لحاظ ارائه، قابلیت ارائه ساده و تصویری را داشته و هم قادر است بین ویژگی‌های موجود، اولویت‌بندی مناسبی را براساس میزان تأثیرگذاری ویژگی‌ها انجام دهد (و قدرت تفسیرپذیری بالاتری هم دارد). در انتها نتیجه به‌دست آمده در رابطه با اولویت‌بندی متغیرهای استفاده‌شده در تحقیق اشاره می‌گردد.

جهت تعیین اولویت متغیرهای مستقل موجود در مجموعه داده‌ها برای هر یک از ۵۰ نمونه تصادفی، درخت‌های تصمیم رسم گردید و بر اساس میزان فراوانی متغیرها در گره‌ها و سطوح مختلف درخت، میزان اهمیت و اولویت متغیرهای مستقل به ترتیب اولویت عبارتند از:

نوع خودرو،

نوع پلاک،

سن خودرو،

گروه خودرو.

در انتها مجدداً شایان ذکر است که برای دستیابی به یک سیستم ارزیابی ریسک دقیق لازم است که شرکت‌های بیمه در صحت گردآوری اطلاعات شخصی رانندگان نیز کوشا باشند.

منابع و ماخذ

اصغری‌اسکوئی، محمدرضا (1394). کاربرد روش پنجره لغزان برای انتخاب ساختار شبکه عصبی با تأخیر زمانی در پیش‌بینی سری‌های زمانی مالی. فصلنامه پژوهشنامه اقتصادی، سال پانزدهم، شماره 57، صص 75-108.

اصغری‌اسکوئی، محمدرضا قاسم‌زاده، محمد (1395). کاربرد قواعد کشفی و الگوریتم ژنتیک در ساخت مدل ARMA برای پیش‌بینی سری‌زمانی. فصلنامه مدیریت فناوری اطلاعات، دانشگاه تهران، دوره 8، شماره 1، صص 1-26.

ایزدپرست، محمود (1390). دسته‌بندی مشتریان بیمه با استفاده از داده‌کاوی. تازه‌های جهان بیمه، شماره 161.

بهادر، آزاده، استادرمضان، آذین خانی‌زاده، فربد (1396). بررسی امکان صدور بیمه‌نامه شخص ثالث بر اساس ویژگی‌های راننده (تبصره 1 ماده 18 قانون جدید بیمه شخص ثالث) و ارائه آیین‌نامه پیشنهادی. پژوهشکده بیمه.

ترکستانی، محمد صالح، ده‌پناه، آرمان، تقوی‌فرد، محمدتقی شفیعی، شهرام (1395). ارائه چارچوبی برای اصلاح نرخ حق بیمه در رشته بدنه اتومبیل با استفاده از مدل شبکه‌های عصبی (مطالعه موردی: شرکت بیمه آسیا)، مدیریت فناوری اطلاعات، دوره 8، شماره 4.

حاجی‌حیدری، نسترن، خالهء، سامرند فراهی، احمد (1390). طبقه‌بندی میزان ریسک بیمه‌گذاران بیمه بدنه خودرو با استفاده از الگوریتم‌های داده‌کاوی (مورد مطالعه: یک شرکت بیمه). پژوهشنامه بیمه، سال بیست‌وششم، شماره 4.

حنفی‌زاده، پیام رستخیز پایدار، ندا (1390). مدلی جهت دسته‌بندی ریسکی گروه‌های مشتریان بیمه بدنه اتومبیل بر اساس ریسک با استفاده از تکنیک داده‌کاوی (مورد مطالعه: بیمه بدنه اتومبیل در یک شرکت بیمه‌ای). پژوهشنامه بیمه، سال بیست و ششم، شماره 2.

عمرانی نوش‌آبادی، مصطفی (1390). ارائه مدل اقتصادسنجی، جهت تعیین حق بیمه بیمه‌گذار در بیمه شخص ثالث خودرو براساس متغیرهای تأثیرگذار بر آن. پایان‌نامه کارشناسی‌ارشد، پژوهشکده بیمه.

فتح‌نژاد، فرامرز ایزدپرست، محمود (1390)، ارائه چهارچوب برای پیش‌بینی سطح خسارت مشتریان بیمه بدنه اتومبیل با استفاده از راهکار داده‌کاوی. تازه‌های جهان بیمه، شماره 156.

کریم‌زادگان مقدم، داود بهروان، مجید (1394)، ارائه راهکاری برای تعرفه‌گذاری پویا در صنعت بیمه با استفاده از تکنیک داده‌کاوی (مورد مطالعه: بیمه شخص ثالث). پژوهشنامه بیمه، شماره 4.

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

Bowers, A.J.; Zhou, X., (2019). Receiver operating characteristic (ROC) area under the curve (AUC): A diagnostic measure for evaluating the accuracy of predictors of education outcomes. Journal of Education for Students Placed at Risk. 24(1), 20-46.

Chawla, N.V., (2009). Data mining for imbalanced datasets: An overview. In Data mining and knowledge discovery handbook (875-886). Springer, Boston, MA.

David, M., (2015). Auto insurance premium calculation using generalized linear models. Procedia Economics and Finance, 20(15), 147-156.

Doucette, J.; Heywood, M.I., (2008). GP classification under imbalanced data sets: Active sub-sampling and AUC approximation. In European Conference on Genetic Programming (266-277). Springer, Berlin, Heidelberg.

Frempong, N.K.; Nicholas, N.; Boateng, M.A., (2017). Decision tree as a predictive modeling tool for auto insurance claims. Int. J. Statist. Appl., 7(2), 117-120.

Gajowniczek, K.; Ząbkowski, T.; Szupiluk, R., (2014). Estimating the roc curve and its significance for classification models'assessment. Metody Ilościowe w Badaniach Ekonomicznych, 15(2), 382-391.

Guo, H.; Viktor, H.L., (2004). Learning from imbalanced data sets with boosting and data generation: the databoost-im approach. ACM Sigkdd Explorations Newsletter, 6(1), 30-39.

Kašćelan, V.; Kašćelan, L.; Novović Burić, M., (2016). A nonparametric data mining approach for risk prediction in car insurance. Economic Research-Ekonomska Istraživanja. 29(1), 545-558.

Menardi, G.; Torelli, N., (2014). Training and assessing classification rules with imbalanced data. Data Mining and Knowledge Discovery, 28(1), 92-122.

Provost, F., (2000). Machine learning from imbalanced data sets 101. In Proceedings of the AAAI'2000 workshop on imbalanced data sets (Vol. 68, No. 2000, 1-3). AAAI Press.

Salma, D.F.; Murfi, H.; Sarwinda, D., (2019). July. The Performance of One-Dimensional Naïve Bayes Classifier for Feature Selection in Predicting Prospective Car Insurance Buyers. In International Conference on Data Mining and Big Data (124-132). Springer, Singapore.

Thakur, S.S.; Sing, J.K., (2013). Mining Customer's Data for Vehicle Insurance Prediction System using k-Means Clustering-An Application. InternationalJjournal of Computer Applications in Engineering sciences, 3(4), 148.

Wang, S.; Liu, W.; Wu, J.; Cao, L.; Meng, Q.; Kennedy, P.J., (2016). Training deep neural networks on imbalanced data sets. In international joint conference on neural networks (4368-4374). IEEE.

Wuyu, S.; Cerna, P., (2019). Risk Assessment Predictive Modelling in Insurance Industry Using Data Mining. Software Engineering, 6(4), 121.

Yunos, Z.M.; Ali, A.; Shamsuddin, S.M.; Ismail, N., (2016). Predictive Modelling for Motor Insurance Claims Using Artificial Neural Networks. Int. J. Advance Soft Compu. Appl, 8(3).