



ORIGINAL RESEARCH PAPER

Comparison of the accuracy of statistical learning algorithms in predicting of the stock price movement of Saman Insurance Company as a listed insurance company

M. Tamandi^{1,*}, M. Askaripour²

¹ Department of Statistics, Faculty of mathematical sciences, Vali-e-Asr University of Rafsanjan, Rafsanjan, Iran

² Technical Deputy of Non-Life Insurance, Saman Insurance, Tehran, Iran

ARTICLE INFO

Article History:

Received 08 January 2023

Revised 09 April 2023

Accepted 14 May 2023

Keywords:

Decision Tree

Insurance

KNN algorithm

Random Forest

Stock price

ABSTRACT

BACKGROUND AND OBJECTIVES: One of the criteria for deciding to invest in a listed company is the amount or changes in the stock price of the company in the future days and months. Various methods have been studied to predict the stock price and investment risk in a company. In most of these methods, the stock price is predicted as a continuous response variable. For this purpose, time series models are used in which assumptions such as the normality of disturbances or the linearity of the model are important. The purpose of this research is to introduce a two-category response variable based on the direction of share price movement in the next day and to introduce some statistical classification methods to predict it. These models do not have the limitations of the previous models, and for that reason, they are of interest. The main objective of this article is to implement the studied methods and compare their accuracy in predicting the orientation of stock price movement of stock exchange insurance companies.

METHODS: In the current research, we have predicted the direction of stock price movement by using K-nearest neighbors, decision trees, and random forest algorithms, which are among the non-parametric classification methods of statistical learning. The data used in this research includes information on the stock price of one of the insurance companies during the years 2019 to 2020, which has a suitable and high share in the portfolio of the insurance industry. To determine the accuracy of the studied models, the data were randomly divided into two groups, training and testing. Then, the statistical learning models were implemented on training data and their validity was measured using experimental data.

FINDINGS: The research results indicate the high accuracy of all three non-parametric models in predicting the stock price category of the insurance company. Likewise, among the studied models, the K-nearest neighbors algorithm performed better than other algorithms in predicting the direction of stock price movement.

CONCLUSION: Considering the importance of the risk of investing in an insurance company for customers, attainment to a valid model for stock price classification and specifying the variables that increase or decrease the price can help customers and insurance companies make better decisions.

*Corresponding Author:

Email: Tamandi@vru.ac.ir

Phone: +9834 31312281

ORCID: [0000-0003-0056-4642](https://orcid.org/0000-0003-0056-4642)

DOI: [10.22056/ijir.2023.03.02](https://doi.org/10.22056/ijir.2023.03.02)

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





مقاله علمی

مقایسه دقت الگوریتم‌های یادگیری آماری در پیش‌بینی جهت حرکت قیمت سهام شرکت بیمه سامان به عنوان یک شرکت بیمه بورسی

مصطفی طامندی^{۱*}، محیا عسکری پور^۲

^۱ گروه آمار، دانشکده علوم ریاضی، دانشگاه ولی عصر (عج) رفسنجان، رفسنجان، ایران

^۲ معاونت فنی بیمه‌های غیر زندگی بیمه سامان، تهران، ایران

اطلاعات مقاله

تاریخ‌های مقاله:

تاریخ دریافت: ۱۸ دی ۱۴۰۱

تاریخ داوری: ۲۰ فروردین ۱۴۰۲

تاریخ پذیرش: ۲۴ اردیبهشت ۱۴۰۲

کلمات کلیدی:

الگوریتم KNN

بیمه

جنگل تصادفی

درخت تصمیم

قیمت سهام

نویسنده مسئول:

ایمیل: tamandi@vru.ac.ir

تلفن: +۹۸۳۴۳۱۳۱۲۲۸۱

ORCID: 0000-0003-0056-4642

چکیده:

پیشینه و اهداف: یکی از معیارهای تصمیم‌گیری برای سرمایه‌گذاری در یک شرکت بورسی، میزان یا تغییرات قیمت سهام آن شرکت در روزها و ماه‌های آتی است. روش‌های متعددی برای پیش‌بینی قیمت سهام و ریسک سرمایه‌گذاری در یک شرکت، مورد مطالعه قرار گرفته است. در اکثر این روش‌ها، قیمت سهام به عنوان یک متغیر پاسخ پیوسته پیش‌بینی شده است. برای این منظور، از مدل‌های سری زمانی استفاده می‌شود که در آنها پذیره‌هایی از جمله نرمال بودن اغتشاش‌ها و یا خطی بودن مدل اهمیت دارد. هدف از این پژوهش، معرفی یک متغیر پاسخ دو رده‌ای براساس جهت حرکت قیمت سهام در روز آتی و معرفی چند روش رده‌بندی آماری برای پیش‌بینی آن است. این مدل‌ها، محدودیت‌های مدل‌های گذشته را ندارند و به همین دلیل مورد توجه هستند. پیاده‌سازی روش‌های مورد مطالعه و مقایسه دقت آنها در پیش‌بینی جهت حرکت قیمت سهام شرکت‌های بیمه بورسی هدف اصلی این مقاله است.

روش‌شناسی: در پژوهش حاضر با استفاده از الگوریتم‌های K-نزدیک‌ترین همسایه‌ها، درخت تصمیم و جنگل تصادفی که در زمره روش‌های ناپارامتری رده‌بندی یادگیری آماری می‌باشند، به پیش‌بینی جهت حرکت قیمت سهام پرداخته‌ایم. داده‌های مورد استفاده در این تحقیق شامل اطلاعات قیمت سهام یکی از شرکت‌های بیمه در طی سال‌های ۱۳۹۰ تا ۱۴۰۰ است که سهم مناسب و بالایی در پرتفوی صنعت بیمه دارد. برای تعیین دقت مدل‌های مورد مطالعه، داده‌ها به صورت تصادفی به دو دسته آموزشی و آزمایشی تقسیم شدند. سپس مدل‌های یادگیری آماری روی داده‌های آموزشی اجرا و اعتبار آنها با استفاده از داده‌های آزمایشی سنجیده شد.

یافته‌ها: نتایج تحقیق حاکی از دقت بالای هر سه مدل ناپارامتری در پیش‌بینی رده قیمت سهام شرکت بیمه مورد نظر است. همچنین در بین مدل‌های مورد مطالعه، الگوریتم K-نزدیک‌ترین همسایه‌ها نسبت به سایر الگوریتم‌ها در پیش‌بینی جهت حرکت قیمت سهام عملکرد بهتری از خود نشان داد.

نتیجه‌گیری: با توجه به اهمیت ریسک سرمایه‌گذاری در یک شرکت بیمه برای مشتریان، یافتن مدل مناسب برای رده‌بندی قیمت سهام و مشخص نمودن متغیرهای مؤثر در افزایش یا کاهش قیمت، می‌تواند به مشتریان و شرکت‌های بیمه در تصمیم‌گیری بهتر کمک کند.

از روش جایگزین یعنی پیش‌بینی جهت حرکت قیمت به‌جای مقدار قیمت مفیدتر است. در این حالت متغیر پاسخ به‌جای اینکه یک متغیر کمی و پیوسته باشد، یک متغیر کیفی دو‌حالتی است. فرض کنید قیمت پایانی یک سهم را در روز t با C_t نمایش دهیم. در این صورت متغیر پاسخ به‌صورت زیر تعریف می‌شود:

$$d_t = \begin{cases} 1 & \text{if } C_t \geq C_{t-1} \\ 0 & \text{if } C_t < C_{t-1} \end{cases} \quad (1)$$

اگر d_t برای روزهای زیادی در یک سهم مقدار یک را به خود بگیرد، جهت سهم افزایشی است و می‌توان نسبت به خرید آن تصمیم‌گیری نمود و اگر صفر باشد، می‌توان نسبت به فروش سهم اقدام نمود. این متغیر همانند نمودارهای شمعی است که در تحلیل تکنیکال برای تغییرات قیمت یک سهم مورد توجه قرار می‌گیرد و شامل مستطیل‌های سبز و قرمز است. درحقیقت، مستطیل‌های سبز در نمودار شمعی معادل با وقتی است که $d_t = 1$ و مستطیل قرمز زمانی است که $d_t = 0$.

استفاده از متغیر دو مقداری d_t به‌جای قیمت پایانی سهم به‌عنوان متغیر پاسخ این مزیت را دارد که در این حالت می‌توان با استفاده از روش‌های رده‌بندی آماری و بدون نیاز به فرض‌هایی همچون ثبات واریانس یا معادلات خطی پارامتریک، رفتار آینده یک سهم را پیش‌بینی کرد. از طرف دیگر، در نظر گرفتن یک متغیر دو‌مقداری به‌جای مقدار قیمت سهم به‌عنوان متغیر پاسخ، مشکل خودهمبستگی را در داده‌های وابسته به زمان برطرف می‌کند و در نتیجه می‌توان از روش‌های رده‌بندی در این حالت نیز استفاده کرد. (Kim (2003 با در نظر گرفتن یک متغیر دو‌مقداری به شکل بالا و با استفاده از ماشین‌های بردار پشتیبان به پیش‌بینی جهت حرکت قیمت سهم در بورس کره جنوبی پرداخت. (Kara et al. (2011 با در نظر گرفتن چنین متغیری به مقایسه دقت روش‌های یادگیری ماشین در پیش‌بینی جهت حرکت شاخص کل بورس استانبول پرداختند.

روش‌های رده‌بندی آماری بسیار متنوع هستند که از روش‌های پارامتری مثل رگرسیون لوژستیک و تحلیل ممیزی خطی تا روش‌های ناپارامتری همچون Bagging و Boosting را شامل می‌شوند. در روش‌های ناپارامتری برخلاف روش‌های پارامتری نیاز به دانستن شکل ارتباط بین متغیرهای مستقل و پاسخ نیست و توزیع احتمالی داده‌ها هم اهمیتی ندارد. این روش‌ها زمانی که بعد داده‌ها بالا باشد بهتر از روش‌های پارامتری عمل می‌کنند. در این مقاله بر روی سه روش ناپارامتری k -نزدیک‌ترین همسایه‌ها (KNN)، درخت تصمیم (DT) و جنگل تصادفی (RF) که جزو روش‌های یادگیری آماری برای رده‌بندی محسوب می‌شوند، تمرکز می‌کنیم. وجه مشترک این سه روش آن است که در آنها، مشاهدات براساس فاصله اقلیدسی میان متغیرها رده‌بندی می‌شوند. به همین دلیل، محدودیت‌های مدل‌های پارامتری را ندارند و رده‌بندها می‌توانند غیرخطی باشند. این روش‌ها در بخش‌های بعد با جزئیات معرفی خواهند شد. یک مرجع مناسب برای

سرمایه‌گذاری در بورس اوراق بهادار و خرید سهام شرکت‌های بورسی یکی از روش‌های رایج و مبتنی بر پذیرش ریسک برای افزایش سرمایه و کسب ثروت در سرتاسر دنیاست. در این میان، یکی از دغدغه‌های معمول یک سرمایه‌گذار آن است که کدام شرکت را انتخاب کند تا سود بیشتری دریافت کند یا ضرر کمتری ببیند. بنابراین، پیش‌بینی میزان یا تغییرات قیمت سهم یک شرکت در روزها و ماه‌های آتی می‌تواند به سرمایه‌گذاران برای تصمیم‌گیری بهتر کمک کند. به‌طور کلی دو روش برای پیش‌بینی قیمت سهم یک شرکت وجود دارد: ۱- روش‌های تحلیل تکنیکال که در آن با استفاده از تغییرات قبلی در قیمت سهم، ارزش بازار، حجم سهم ارائه شده و ... قیمت سهم پیش‌بینی می‌شود. ۲- روش‌های تحلیل بنیادین که در آنها عوامل بیرونی همچون تصمیمات دولت‌ها، میزان تولید ناخالص ملی، قیمت ارز و طلا و قیمت‌های جهانی برای پیش‌بینی قیمت سهم یک شرکت به‌عنوان متغیرهای تأثیرگذار در نظر گرفته می‌شود (Shah et al., 2019).

از آنجا که روش‌های بنیادی نیاز به دسترسی به اطلاعات گسترده‌ای از خارج از بازار بورس دارد و در کشور ما این اطلاعات کمتر در اختیار است، برای پیش‌بینی قیمت سهم معمولاً روش‌های تکنیکال مورد توجه قرار گرفته است. در این حالت، متغیر پاسخی که می‌خواهیم آن را پیش‌بینی کنیم، به دو صورت تعریف می‌شود. در حالت اول به‌صورت مستقیم قیمت سهم را در روز یا روزهای آتی پیش‌بینی می‌کنیم. در حالت دوم متغیر پاسخ بسته به اینکه قیمت سهم نسبت به روز قبل افزایش خواهد یافت یا کاهش خواهد بود، به‌صورت رسته‌بندی شده و دارای دو مقدار است. حالت اول به فراوانی در مطالعات مختلف مورد توجه قرار گرفته است. در این روش معمولاً از روش‌های آماری کلاسیک همانند سری‌های زمانی ARMA و ARIMA (Singh et al., 2021; Almasarweh and Alwadi, 2018) یا روش‌های جدیدتر مثل مدل‌های ARCH، GARCH و CARR (Roh, 2007) استفاده می‌شود. همچنین روش‌های یادگیری ماشین مثل شبکه‌های عصبی مصنوعی و ماشین‌های بردار پشتیبان (SVM) هم مورد توجه قرار گرفته‌اند (Shah et al., 2019; Başoğlu Kabran and Ünlü, 2021). Mehrabanpour et al. (2022) پیشنهاد کردند که ترکیبی از روش مؤلفه‌های اصلی و نظریه مجموعه‌های راف برای پیش‌بینی مقدار قیمت سهم مورد استفاده قرار گیرد.

روش پیش‌بینی مستقیم مقدار قیمت سهم نقاطضعفی هم دارد. سری‌های زمانی قیمت‌های سهم معمولاً غیرخطی با خطاهای غیرنرمال هستند. از طرفی، در بورس‌های با ثبات اقتصادی کمتر، قیمت سهم تحت تأثیر بحران‌های مختلف دچار نوسان‌های شدید می‌شوند و در نتیجه ثبات واریانس که معمولاً در مدل‌های آماری از جمله فرض‌های اساسی محسوب می‌شود، در این داده‌ها وجود ندارد. بنابراین، نتایج و پیش‌بینی‌های حاصل از این مدل‌ها ممکن است چندان دقیق و مورد اعتماد نباشد. در چنین حالتی استفاده

روش‌های یادگیری و رده‌بندی آماری، کتاب (James et al. (2013 است.

صنعت بیمه و به‌ویژه شرکت‌های خصوصی بیمه که سهامشان در بورس اوراق بهادار ارائه شده است، نقش مهمی در اقتصاد کشور دارند. شرکت‌های بیمه در فاصله بین پرداخت حق بیمه تا ادعای خسارت توسط مشتری این فرصت را دارند که از درآمد حاصل برای سرمایه‌گذاری در بخش‌های مختلف بازار استفاده کنند و در نتیجه باعث افزایش ارزش سهام خود شوند. از طرفی تعداد زیاد شرکت‌های بیمه و رقابت بین آنها برای جذب مشتری، این شرکت‌ها را بر آن داشته است که از روش‌های جدید برای تحلیل بازار و بررسی رفتار مشتریان استفاده کنند. علاوه بر این، شیوع ویروس کرونا و تبعات آن، نگاه بازار را به نحوه سرمایه‌گذاری عوض کرده است و شیوه‌های مرتبط با محاسبات ابری، اینترنت اشیا، هوش مصنوعی و بازاریابی دیجیتال مورد توجه قرار گرفته است. حجم داده‌هایی که باید در چنین مواردی تحلیل شوند آن قدر زیاد است که روش‌های کلاسیک آمار و ریاضیات کاربردی جواب‌گویی آنها نیست. در نتیجه، روش‌های یادگیری ماشین، داده‌کاوی و یادگیری آماری در این زمینه می‌تواند به بهبود عملکرد شرکت‌های بیمه کمک کند و از طرفی مشتریان هم با مقایسه آنها بتوانند شرکت مناسبی را برای سرمایه‌گذاری انتخاب کنند.

در این مقاله قصد داریم به پیش‌بینی جهت حرکت قیمت سهام شرکت‌های بیمه بورسی براساس مطالعه موردی در شرکت بیمه سامان به‌عنوان یکی از شرکت‌های با سابقه در فعالیت بورسی با استفاده از روش‌های KNN، درخت تصمیم و جنگل تصادفی بپردازیم و دقت این روش‌ها را با یکدیگر مقایسه کنیم. در بخش دوم ابتدا مروری بر ادبیات و پیشینه تحقیق در زمینه کاربرد روش‌های یادگیری آماری در داده‌های اقتصادی خواهیم داشت. در بخش سوم به‌طور خلاصه به مبانی نظری روش‌های پیش‌گفته پرداخته می‌شود. متغیرهای مستقل و پاسخ مورد استفاده در این تحقیق و اطلاعاتی در توصیف داده‌های مورد استفاده و نحوه به‌کارگیری روش‌های یادگیری آماری در بخش چهارم ارائه خواهد شد. بخش پنجم به تحلیل آماری داده‌ها و مقایسه دقت روش‌های به‌کار گرفته‌شده، تعلق خواهد داشت. در پایان و در بخش ششم به نتیجه‌گیری و ارائه پیشنهادات برای پژوهش‌های آینده می‌پردازیم.

مبانی نظری پژوهش

در این بخش سه روش رده‌بندی را که در یادگیری آماری مورد استفاده قرار می‌گیرد، معرفی می‌کنیم. لازم به ذکر است که مدل‌های دیگری هم برای رده‌بندی داده‌ها وجود دارند، اما مدل‌های حاضر در این مقاله به این دلیل انتخاب شده‌اند که در ساخت آنها از روش‌های آماری و احتمالاتی استفاده شده است. سایر روش‌های رده‌بندی مثل SVM و شبکه عصبی عمدتاً با استفاده از روش‌های بهینه‌سازی ریاضی ساخته می‌شوند. همچنین روش‌های بیان‌شده در این مقاله در گروهی از مدل‌های آماری رده‌بندی می‌شوند که به آنها

روش‌های مبتنی بر درخت گفته می‌شود.

هدف از روش‌های رده‌بندی آن است که براساس چندین ویژگی یا متغیر مستقل تصمیم بگیریم که یک مشاهده در کدام کلاس یا رده از متغیر پاسخ (دو یا چند سطحی) قرار خواهد گرفت. به‌عنوان مثال، در تحقیق حاضر قرار است براساس چندین متغیر مستقل که در بخش بعد آنها را معرفی خواهیم کرد، پیش‌بینی کنیم آیا قیمت سهم در روز آینده افزایشی خواهد بود یا کاهش. در مدل‌های رده‌بندی، دقت مدل در رده‌بندی درست مشاهدات اهمیت دارد. به این معنا که رده واقعی یک مشاهده با رده‌ای که توسط مدل پیش‌بینی می‌شود یکسان است یا خیر. دقت یک مدل رده‌بندی براساس نسبت حالت‌های تطابق بین آنچه مدل پیش‌بینی می‌کند و آنچه در واقعیت وجود دارد، به‌دست می‌آید. بنابراین، هرچه این عدد به یک نزدیکتر باشد، مدل دقیق‌تر است.

k - نزدیک‌ترین همسایه‌ها (KNN)

در روش KNN، برای رده‌بندی مشاهدات در دو گروه به این صورت عمل می‌شود. ابتدا یک مشاهده X در نظر گرفته می‌شود و k مشاهده یا همسایه که نزدیک‌ترین فاصله با X نسبت به سایر مشاهدات دارند، در گروهی قرار می‌گیرند که X قرار گرفته است. این کار برای بقیه مشاهدات هم انجام می‌گیرد و این فرایند به‌صورت تکراری و با انتخاب‌های تصادفی آن قدر تکرار می‌شود تا نهایتاً همه مشاهدات در دو گروه رده‌بندی شوند.

انتخاب مقدار مناسب k نقش مهمی در دقت مدل به‌دست آمده دارد. در این مقاله برای انتخاب k ، از روش اعتبارسنجی متقابل استفاده می‌کنیم. در این روش، ابتدا داده‌ها به‌صورت تصادفی به دو دسته آموزشی و آزمایشی تقسیم می‌شود. سپس مدل‌های KNN با مقادیر مختلف $k=1, \dots, 30$ به داده‌های آموزشی برازش داده می‌شود. برای هر کدام از مدل‌ها، دقت رده‌بندی براساس داده‌های آزمایشی تعیین می‌شود و مقداری از k را که به‌ازای آن بیشترین دقت رده‌بندی به‌دست آمده باشد، به‌عنوان مقدار مناسب انتخاب می‌کنیم. در مدل KNN، مرز جداکننده دو گروه می‌تواند غیرخطی باشد. بنابراین، این مدل می‌تواند نسبت به برخی دیگر از روش‌های آماری همچون رگرسیون لوژستیک و روش ممیزی خطی برتری داشته باشد. علاوه بر این، مدل KNN یک مدل ناپارامتری است و نیازی به دانستن توزیع متغیرهای مستقل در این حالت نیست.

درخت تصمیم

یکی دیگر از روش‌های ناپارامتری برای رده‌بندی داده‌ها، درخت تصمیم است. یک درخت تصمیم شامل چند شاخه است که مشخص می‌کند هر مشاهده از مجموعه داده‌ها در هر کدام از ناحیه‌ها یا شاخه‌ها با چه احتمالی در یکی از رده‌های متغیر پاسخ رده‌بندی می‌شود. بنابراین، یک درخت تصمیم علاوه بر اینکه همانند روش قبلی مشخص می‌کند که هر مشاهده در کدام کلاس رده‌بندی می‌شود، می‌تواند نسبت یا سهم هر متغیر را به‌صورت مرتب‌شده

متقابل). دقت مدل جنگل تصادفی به تعداد درختان و تعداد گره‌ها یا متغیرهای مستقلی که در ساخت مدل تأثیرگذار هستند، بستگی دارد (Breiman, 2001).

در بخش بعد ابتدا به توصیف داده‌ها و متغیرهای مورد استفاده در این تحقیق خواهیم پرداخت و سپس با توجه به روش‌های معرفی شده در این بخش به رده‌بندی داده‌ها می‌پردازیم.

مروری بر پیشینه پژوهش

نویسندگان متعددی درباره کاربرد روش‌های یادگیری ماشین در تحلیل داده‌های اقتصادی و به‌ویژه بورس اوراق بهادار مطالعه کرده‌اند. (Rostamkhani et al., 2021) با استفاده از جنگل تصادفی به انتخاب سهم بهینه در بازار بورس پرداخت. براساس مطالعه ایشان در سناریوهای مختلف، جنگل تصادفی دقتی حدود ۸۰ تا ۹۰ درصد در پیش‌بینی سهم بهینه را نشان می‌دهد. (Kara et al., 2011) از روش‌های شبکه‌های عصبی مصنوعی و SVM در بازار بورس استانبول استفاده کرده‌اند. طبق نتیجه تحقیق ایشان، مدل شبکه‌های مصنوعی با ۷۵ درصد دقت نسبت به مدل SVM عملکرد بهتری در پیش‌بینی جهت حرکت شاخص کل بورس استانبول دارد.

(Milosevic, 2016) برای تعیین اینکه یک سهم برای سرمایه‌گذاری خوب است یا نه، به مقایسه روش‌های جنگل تصادفی، SVM و بیز ساده پرداخته است و نتیجه گرفته است که روش جنگل تصادفی با دقتی در حدود ۷۵ درصد نسبت به سایر روش‌ها عملکرد بهتری دارد. (Başoğlu Kabran and Ünlü, 2021) با استفاده از یک روش یادگیری ماشین دو مرحله به بررسی حباب موجود در شاخص اقتصادی S&P 500 در بازار آمریکا پرداختند. در تحقیق ایشان روش‌های KNN، رگرسیون لوژستیک، درخت تصمیم، شبکه‌های عصبی مصنوعی و SVM با یکدیگر مورد مقایسه قرار گرفته‌اند. طبق نتیجه این تحقیق، روش‌های SVM و KNN با دقتی در حدود ۹۹ درصد نسبت به سایر روش‌ها پیش‌بینی بهتری درباره وجود حباب در شاخص S&P ارائه می‌دهد. هرچند میزان بالای این دقت ممکن است نشانه‌ای از مفهوم بیش‌برازش در مسئله داشته باشد که نیاز به مذاقه بیشتر دارد. علاوه بر تحقیقات مورد اشاره، مقاله مروری (Shah et al., 2019) یک مرجع مناسب برای علاقه‌مندان به بررسی کاربرد روش‌های یادگیری آماری در تحلیل داده‌های بورسی است. همچنین منابع زیر به کاربرد روش‌های یادگیری ماشین در پیش‌بینی قیمت سهام پرداخته‌اند: (Bing et al., 2012; Huang et al., 2005; Mintarya et al., 2023; Nair et al., 2010; Patel et al., 2015; Rouf et al., 2021; Singh et al., 2019; Zhang and Wu, 2009).

در زمینه کاربرد روش‌های یادگیری ماشین در صنعت بیمه در ایران کارهای بسیار کمی صورت گرفته است. (Izadparast et al., 2012) از درخت تصمیم برای پیش‌بینی سطح خسارت مشتریان بیمه بدنه اتموبیل استفاده کردند و براساس نتایج این تحقیق، دقت درخت تصمیم در رده‌بندی مشتریان بیمه حدود ۶۰ درصد اعلام شده است.

در شاخه‌های درخت تصمیم معین کند. در درخت تصمیم همانند آنچه در روش خوشه‌بندی داده‌ها اتفاق می‌افتد، از یک ریشه شروع می‌کنیم و با روش‌های تکراری شاخه‌های درخت تصمیم را مشخص می‌کنیم. جداسازی شاخه‌ها از هم براساس نرخ خطای رده‌بندی انجام می‌شود. اگر p_{mk} نسبت مشاهداتی از شاخه m باشند که در کلاس k رده‌بندی می‌شوند، در این شاخه نرخ رده‌بندی نادرست از فرمول زیر محاسبه می‌شود:

$$E = 1 - p_{mk} \quad (2)$$

بنابراین، اگر مقدار E کوچک باشد، یک شاخه در درخت تصمیم باقی می‌ماند. به بیان دیگر، اگر ادغام دو شاخه در هم باعث کاهش نرخ رده‌بندی نادرست می‌شود، آنگاه این دو شاخه در هم ادغام می‌شوند و نهایتاً یک درخت خالص به دست می‌آید. در روش درخت تصمیم هم از آنجا که با یک روش ناپارامتری سروکار داریم، باید داده‌ها را به دو مجموعه آزمایشی و آموزشی تقسیم کرده و با استفاده از داده‌های آزمایشی به مدل بهینه برسیم. (Kotsiantis, 2013) یک مطالعه مروری در زمینه روش درخت تصمیم و کاربردهای آن انجام داده است که می‌تواند جزئیات بیشتری از این روش در اختیار علاقه‌مندان قرار دهد.

جنگل تصادفی

همان‌طور که در بخش قبل گفتیم، روش درخت تصمیم برای رده‌بندی داده‌ها مدل مناسبی است. اما این مدل به شدت وابسته به متغیرهای مستقل است. اگر در مجموعه داده‌ها یک متغیر مستقل وجود داشته باشد که در رده‌بندی نسبت به بقیه قوی‌تر باشد، روی درختی که در نهایت ساخته می‌شود، تأثیرگذار خواهد بود. در روش‌های مبتنی بر درخت، حذف شاخه‌های اضافی یا به اصطلاح هرس کردن درخت و رسیدن به یک درخت خالص از اهمیت ویژه‌ای برخوردار است. این کار معمولاً با بهینه‌سازی شاخص نرخ رده‌بندی نادرست که در معادله (۲) به آن اشاره شده است، انجام می‌شود (درختی که کمترین میزان نرخ رده‌بندی نادرست را داشته باشد، خالص‌تر است). در این خالص‌سازی هر کدام از متغیرها می‌توانند به صورت تک‌به‌تک در کاهش این شاخص تأثیرگذار باشند. یک جنگل تصادفی از چندین درخت تصمیم ساخته می‌شود و اگر هر کدام از این درخت‌ها با مجموعه‌ای از متغیرهای مستقل ثابت ساخته شوند، ممکن است یک درخت شامل متغیرهایی با قدرت بالا در کاهش نرخ رده‌بندی نادرست باشد و درخت دیگر شامل متغیرهایی با قدرت کم. برای حل این مشکل می‌توان چندین درخت را با استفاده از مجموعه‌ای از متغیرهای مستقل که به صورت تصادفی انتخاب شده‌اند، ایجاد کرد و در بین آنها به دنبال مجموعه‌ای باشیم که بیشترین دقت رده‌بندی را ایجاد می‌کند. این جنگل تصادفی از درخت‌های تصمیم مثل همان عملیات تکراری عمل می‌کند که در روش KNN هم به وسیله آن k مناسب را یافتیم (روش اعتبارسنجی

نوسانات آن براساس سرعت حرکت قیمت و همچنین جهت حرکت آن می‌باشد. نحوه محاسبه شاخص تصادفی k در جدول ۱ آمده است. در این جدول، L_{t-n} و H_{t-n} به ترتیب کمترین و بیشترین قیمت سهم در n روز گذشته است.

(پ) RSI : یکی از پرطرفدارترین شاخص‌ها در بین تحلیل‌گران بازار بورس، شاخص قدرت نسبی (RSI) است. شاخص قدرت نسبی با بررسی دوره زمانی مشخص که به صورت استاندارد ۱۴ روزه می‌باشد، قدرت خرید و فروش در یک گزینه معاملاتی را مورد بررسی قرار می‌دهد و آن را در یک نمودار با سه بازه مختلف نشان می‌دهد. این شاخص یک نوسانگر است و این نوسانگر بین دو سطح ۰ تا ۱۰۰ نوسان می‌کند. فرمول محاسبه شاخص RSI در جدول ۱ موجود است. در این جدول، D_t و U_t به ترتیب میزان تغییر قیمت سهم در لحظه t به سمت پایین و به سمت بالاست.

(ت) $MACD$: میانگین متحرک همگرا واگرا توسط جرال د اپل فیزیکدان و محقق آمریکایی در اواخر سال ۱۹۷۰ طراحی شد. شاخص $MACD$ در خانواده نوسانگرها قرار می‌گیرد و همان‌طور که از اسمش پیداست، از میانگین متحرک در محاسبات این شاخص استفاده شده است. این شاخص از دو میانگین متحرک ساخته می‌شود که به قیمت‌های نزدیک‌تر به پایان دوره وزن بیشتری می‌دهد. این شاخص هم با توجه به ماهیت آن می‌تواند در پیش‌بینی جهت حرکت قیمت سهم مؤثر باشد (Wilder, 1978). نحوه محاسبه شاخص $MACD$ در جدول ۱ آمده است.

(ث) ADX : این شاخص، میانگین حرکت جهت‌دار است. این شاخص همچون دو شاخص قبلی، از نوع شاخص‌های روندی است و از ۲ بخش اصلی با نام‌های منحنی DI^+ و منحنی DI^- تشکیل شده است. DI^+ و DI^- به ترتیب شاخص‌های جهت حرکت سهم به بالا یا پایین هستند که براساس متوسط پایین‌ترین و بالاترین قیمت دو روز متفاوت برای روزهای پشت‌سرهم به دست می‌آید. زمانی که خط ADX در فرایند تحلیل سهم روند صعودی داشته باشد، یعنی بازار روند دارد. بازار رونددار می‌تواند شامل روند نزولی یا صعودی باشد. حالت افقی و بدون شیب این خط هم به معنای راکد بودن معاملات سهم در بازار بورس است. فرمول شاخص ADX هم در جدول ۱ آمده است.

در این مقاله قصد داریم مقدار d_t را که در معادله (۱) به آن اشاره شده است، با توجه به شاخص‌هایی که در جدول ۱ آمده و با استفاده از روش‌های رده‌بندی مورد اشاره در بخش قبل، برای بررسی جهت حرکت قیمت سهم شرکت بیمه سامان پیش‌بینی کنیم. جدول ۲، نشان‌دهنده میانگین شاخص‌های مورد مطالعه در شرکت بیمه سامان است.

یکی از مشکلاتی که در مدل‌های یادگیری نظارتی چالش برانگیز است، وجود همخطی چندگانه بین متغیرهای پیش‌گو (شاخص‌ها) می‌باشد. شاخص‌های مورد استفاده در این مقاله با توجه به پیشینه تحقیق و مقالات مشابه به کار گرفته شده‌اند و از این مسئله رنج می‌برند. همخطی چندگانه در مدل‌های پارامتری همچون رگرسیون

(Asghari oskoei et al. (2020) به کاربرد روش‌هایی همچون شبکه‌های عصبی، درخت تصمیم و بیز ساده در پیش‌بینی ریسک خسارت مالی بیمه‌های شخص ثالث پرداخته‌اند و به این نتیجه رسیده‌اند که براساس متغیرهای مستقلی همچون نوع و سن خودرو، درخت تصمیم با دقتی معادل ۷۲ درصد بهترین پیش‌بینی‌کننده خسارت مالی در بین روش‌های مورد مطالعه ایشان بوده است. در منابع خارجی که کاربرد روش‌های یادگیری ماشین و رده‌بندی را در صنعت بیمه مطالعه کرده‌اند، می‌توان به منابع زیر اشاره نمود. Frempong et al. (2017) یک درخت تصمیم برای پیش‌بینی نیازهای مربوط به بیمه اتومبیل ارائه کرده‌اند. Yunos et al. (2016) یک مدل پیش‌بینی برای متقاضیان بیمه موتورسیکلت براساس شبکه‌های عصبی مصنوعی پیشنهاد داده‌اند. Khamesian et al. (2022) انواع روش‌های یادگیری ماشین را در پیش‌بینی قیمت بیمه اتومبیل مورد مقایسه قرار دادند.

همان‌طور که اشاره شد، قبلاً برای پیش‌بینی قیمت سهام شرکت‌های بورسی از روش‌های یادگیری ماشین استفاده شده است. اما جهت حرکت قیمت سهم به‌عنوان یک متغیر پاسخ رده‌ای و استفاده از روش‌های یادگیری آماری مبتنی بر درخت برای پیش‌بینی آن کمتر مورد توجه قرار گرفته است. در این تحقیق به معرفی چند روش یادگیری آماری و کاربرد آن در صنعت بیمه می‌پردازیم.

روش‌شناسی پژوهش

در این تحقیق می‌خواهیم کاربرد روش‌های رده‌بندی را در تعیین روند افزایشی یا کاهش‌ی در قیمت سهام شرکت‌های بیمه بورسی مورد مطالعه قرار دهیم. برای این منظور قیمت سهام شرکت بیمه سامان را به‌عنوان یکی از شرکت‌هایی که بیش از ۱۰۰۰ روز از ارائه آنها در بورس گذشته است، از طریق سایت شرکت مدیریت فناوری بورس تهران دریافت کردیم. این داده‌ها شامل اطلاعات قیمت سهم از تاریخ ۱۳۹۱/۰۴/۲۴ تا ۱۴۰۱/۱۰/۱۳ است که به‌صورت روزانه ثبت شده‌اند. تعداد داده‌ها بعد از حذف داده‌های گمشده، ۱۹۵۱ مشاهده است. متغیرهای مستقل مورد استفاده در این تحقیق، شاخص‌هایی هستند که در تحلیل تکنیکال مورد استفاده قرار می‌گیرد. این شاخص‌ها در زیر تعریف شده‌اند:

(الف) EMA : این شاخص میانگین متحرک نمایی است که از رابطه زیر به دست می‌آید:

$$EMA_t = \frac{2}{n+1}c_t + \frac{n-1}{n+1}EMA_{t-1}$$

n تعداد روزهای گذشته را نشان می‌دهد که می‌خواهیم تحت مطالعه قرار گیرد. در این مقاله $n=14$ را در نظر گرفتیم.

(ب) $StochK$: شاخص تصادفی k نوسانگری است که وضعیت قیمت پایانی را نسبت به بالا و پایین‌ترین قیمت نشان می‌دهد. این شاخص از قیمت سهم و حجم معاملات تبعیت نمی‌کند، بلکه

جدول ۱: متغیرهای مستقل مورد استفاده در مدل‌های رده‌بندی و تعریف آنها
Table 1: Independent variables using in classification models and their definition.

تعریف	شاخص
قیمت ابتدایی سهم در روز t	OPEN
	$c_t - c_{t-1}$
	lag1
	$\frac{c_t - L_{t-n}}{H_{t-n} - L_{t-n}}$
	Stochk
	$100 - \frac{100}{1 + \left(\frac{\sum_{i=0}^{n-1} U_{t-i}}{\sum_{i=0}^{n-1} D_{t-i}} \right)}$
	RSI
	$EMA(26)_t - EMA(12)_t$
که در آن	MACD
	$EMA(n)_t = \frac{2}{n+1}c_t + \frac{n-1}{n+1}EMA(n)_{t-1}$
	$\frac{(n-1)ADX_{t-1} + DX_t}{n}$
که در آن	ADX
	$DX_t = \frac{D_i^+ - D_i^-}{D_i^+ + D_i^-}$

جدول ۲: اطلاعات توصیفی قیمت سهام شرکت بیمه سامان
Table 2: Descriptive information on stock price of the Saman insurance company

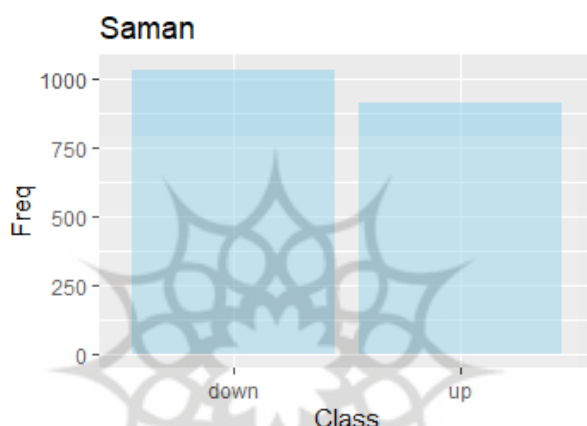
1951	تعداد مشاهده Number of observations
7062/29	میانگین قیمت پایانی Close price mean
-0/9284	میانگین MACD MACD mean
48/245	میانگین RSI RSI mean
25/837	میانگین ADX ADX mean
0/4462	میانگین StochK StochK mean

زمان همگرایی الگوریتم را افزایش می‌دهد، اما باعث کاهش دقت نمی‌شوند، زیرا این مدل‌ها مبتنی بر فاصله‌های اقلیدسی هستند و براساس آن داده‌ها را رده‌بندی می‌کنند نه با استفاده از برآورد پارامترها (Hao and Priestley, 2016; Sharma, 2012).
برای انجام الگوریتم‌های رده‌بندی مورد اشاره در فصل قبل، ابتدا باید داده‌ها را به دو گروه آموزشی و آزمایشی تقسیم کنیم. معمولاً داده‌های آموزشی ۸۰ درصد از داده‌ها و در نتیجه داده‌های آزمایشی ۲۰ درصد از کل داده‌ها هستند که در داده‌های سری زمانی از آخرین

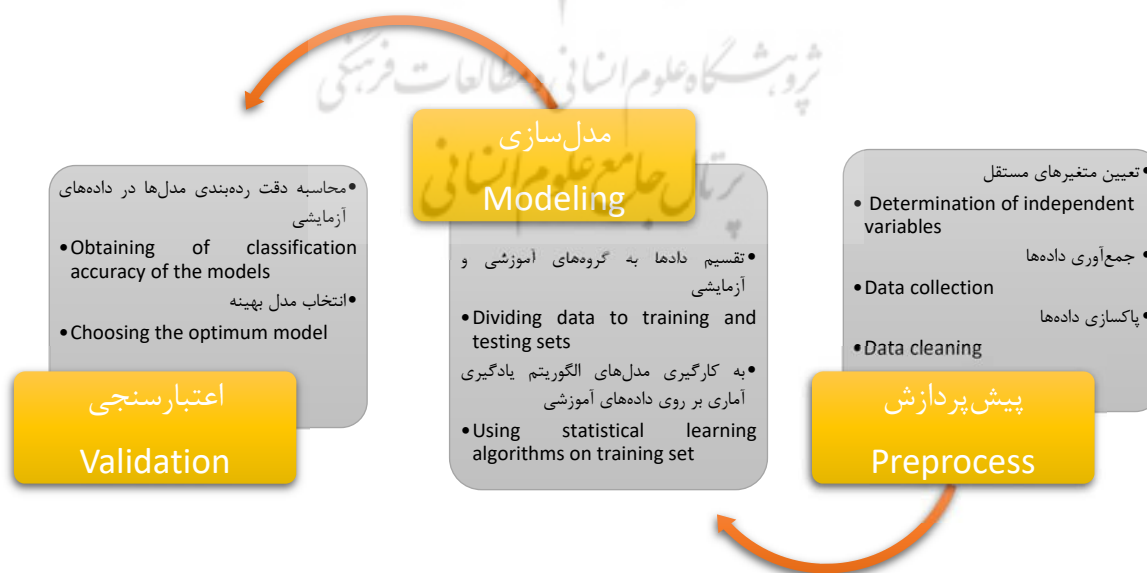
لوژستیک و تحلیل ممیزی فیشر باعث می‌شود ماتریس مشاهدات معکوس پذیر نباشد و در نتیجه امکان برآورد پارامترها وجود ندارد. این مسئله باعث افزایش اریبی برآوردگرها و میانگین مربع خطای مدل می‌شود و در نتیجه این مدل‌ها هرچند ممکن است دقت بالایی (از نظر میزان خطای رده‌بندی نادرست) در رده‌بندی داده‌ها از خود نشان دهند، اما به دلیل وجود همخطی قابل‌اعتماد نیستند و بنابراین در این مقاله استفاده نشده‌اند. در مقابل، همخطی چندگانه در مدل‌های ناپارامتری همچون KNN، درخت تصمیم و جنگل تصادفی، هرچند

مدل‌سازی و اعتبارسنجی مدل‌هاست. در مرحله اول، داده‌های ابتدایی باید پاکسازی شوند و داده‌های گم‌شده و پرت، جایگزین و حذف شوند. در مرحله دوم، مدل‌های یادگیری ماشین باید بر روی داده‌ها اجرا شوند. اما همان‌طور که قبلاً گفته شد، از آنجا که مدل‌های یادگیری از طریق داده‌ها عملکرد خود را بهبود می‌بخشند، ابتدا داده‌ها به دو گروه آموزشی و آزمایشی تقسیم می‌شوند. سپس مدل براساس داده‌های آموزشی ساخته شده و در داده‌های آزمایشی مورد بررسی قرار می‌گیرد. مرحله سوم شامل بررسی دقت مدل‌ها در داده‌های آزمایشی است. هرچه دقت یک مدل بیشتر باشد به این معناست که خطای رده‌بندی آن مدل کمتر است و بنابراین آن مدل قابل‌اعتمادتر می‌باشد.

داده‌ها انتخاب می‌شوند. تقسیم داده‌ها به این دو گروه برای بررسی دقت مدل انجام می‌شود. به این صورت که مدل با استفاده از داده‌های آموزشی ساخته شده و بر روی داده‌های آزمایشی مورد بررسی قرار می‌گیرد. در صورتی که مدل ساخته‌شده دارای دقت مناسبی در داده‌های آزمایشی باشد، می‌توان از این مدل برای رده‌بندی مشاهدات استفاده کرد. **شکل ۱**، تعداد روزهایی را نشان می‌دهد که قیمت سهام شرکت بیمه سامان جهت کاهشی (down) و جهت افزایشی (up) دارد. همان‌طور که ملاحظه می‌شود، تعداد روزهایی که جهت قیمت کاهشی است کمی بیشتر از روزهای افزایشی است. **شکل ۲**، چارچوب انجام مراحل تحلیل داده‌ها و رده‌بندی مشاهدات را نشان می‌دهد، که شامل سه مرحله اصلی پیش‌پردازش،



شکل ۱: نمودار فراوانی جهت افزایشی (up) و جهت کاهشی (down) در قیمت سهام شرکت بیمه سامان
Figure 1: Frequency plot of upward and downward in the stock price of the Saman insurance company



شکل ۲: چارچوب مدل پیشنهادی
Figure 2: The framework of the proposed model

نتایج و بحث

شاخص امتیاز F : میانگین هارمونیک شاخص‌های صحت و پوشش است که به صورت زیر تعریف می‌شود:

$$F. score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

شاخص دقت: همان‌طور که مشخص است این شاخص نسبتی از مشاهدات را که به درستی طبقه‌بندی شده‌اند محاسبه می‌کند و معادله آن به صورت زیر است:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

همه این شاخص‌ها مقادیر بین صفر و یک دارند و هرچه به یک نزدیک‌تر باشند به معنای دقت بالاتر مدل در رده‌بندی مشاهدات است. همچنین با استفاده از شاخص پوشش یا حساسیت و شاخص دیگری به نام اختصاص که نرخ پیش‌بینی درست را در حالت منفی اندازه می‌گیرد، نمودار منحنی مشخصه عملکرد (ROC) رسم می‌شود که مساحت سطح زیر آن (AUC) نشان‌دهنده دقت مدل است. هرچه این منحنی بالاتر باشد یا مقدار مساحت سطح زیر آن به یک نزدیک‌تر باشد، مدل دقیق‌تر است (Ghorbani et al., 2022).

لازم به ذکر است که تمامی محاسبات، با نرم افزار R و با استفاده از بسته‌های $class$ ، $rpart$ و $randomForest$ و توسط کامپیوتری با مشخصات RAM=16GB، Processor=AMD4700، WIN11 و 64GB انجام شده است. کدهای مورد استفاده در آدرس [github](https://github.com) در دسترس می‌باشد.

تحلیل KNN

همان‌طور که قبلاً گفتیم، در این نوع رده‌بندی موضوع اصلی انتخاب k مناسب برای خوشه‌بندی مشاهدات است. در شکل ۳، نمودار دقت مدل KNN در رده‌بندی برای مقادیر مختلف k رسم شده است. براساس این نمودار، $k = 23$ بیشترین دقت را با استفاده از داده‌های آموزشی ایجاد می‌کند. بنابراین، مدل KNN را با این مقدار از k به داده‌های آزمایشی برازش می‌دهیم تا به وسیله آن دقت مدل را بسنجیم. جدول ۴، نتایج رده‌بندی را با استفاده از مدل KNN در

در این بخش، سه مدل ناپارامتری رده‌بندی را که در بخش مبانی نظری به آنها اشاره کردیم، بر روی داده‌های شرکت بیمه سامان برازش می‌دهیم. همان‌طور که قبلاً اشاره شد، داده‌ها به دو گروه آموزشی (۱۵۶۱ روز) و آزمایشی (۳۹۱ روز) تقسیم و استاندارد شدند. هر سه مدل بر روی داده‌های آموزشی ساخته می‌شوند و سپس مدل ساخته‌شده بر روی داده‌های آزمایشی آزموده می‌شود تا دقت مدل بررسی گردد.

همان‌طور که قبلاً گفته شد، دقت مدل‌های رده‌بندی براساس نسبت یا درصد تعداد حالت‌هایی سنجیده می‌شود که پیش‌بینی مدل از رده یا کلاس مشاهده با واقعیت تطابق داشته باشد. زمانی که فقط دو کلاس برای رده‌بندی وجود داشته باشد، معمولاً از ماتریس درهم‌ریختگی برای بررسی دقت مدل استفاده می‌شود. این ماتریس به شکل جدول ۳ است.

در جدول ۳، TP و FP به ترتیب تعداد رده‌بندی درست و نادرست در حالتی است که در واقعیت، رده درست مثبت است (TN, dt=1) و FN و به ترتیب تعداد رده‌بندی درست و نادرست در حالتی است که در واقعیت، رده درست منفی است (dt=0). با توجه به این ماتریس می‌توان شاخص‌های زیر را برای بررسی دقت یک مدل در رده‌بندی مشاهدات معرفی کرد:

شاخص صحت: این معیار نرخ پیش‌بینی درست را نسبت به حالت‌هایی که توسط مدل مثبت ارزیابی شده‌اند، نشان می‌دهد و به صورت زیر تعریف می‌شود:

$$Precision = \frac{TP}{TP + FP}$$

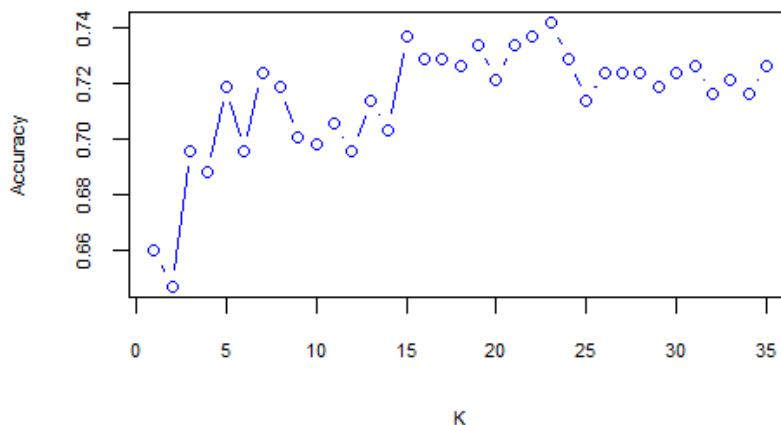
شاخص پوشش یا یادآوری: تعداد پیش‌بینی‌های مثبت صحیح از همه پیش‌بینی‌های مثبت که می‌توانست انجام شود را محاسبه می‌کند و به صورت زیر به دست می‌آید:

$$Recall = \frac{TP}{TP + FN}$$

این شاخص گاهی به نام حساسیت هم نامیده می‌شود.

جدول ۳: نمونه‌ای از یک ماتریس درهم‌ریختگی برای داده‌ها
Table 3: A confusion matrix sample

پیش‌بینی مدل prediction		نتیجه واقعی Observed
dt=1	dt=0	dt=0
FP	TN	dt=1
TP	FN	



شکل ۳: نمودار دقت مدل های KNN برای داده های شرکت بیمه سامان براساس مقادیر مختلف k
Figure 3: The accuracy of the KNN models with different k for Saman insurance company data

جدول ۴: ماتریس درهم‌ریختگی براساس مدل KNN برای داده‌های آزمایشی شرکت بیمه سامان
Table 4: Confusion matrix of the KNN model for testing data of the Saman insurance company

پیش‌بینی مدل prediction		نتیجه واقعی observed
dt=1	dt=0	
42	159	dt=0
125	65	dt=1

آموزشی در این شاخه از درخت تصمیم قرار می‌گیرند. از طرفی اگر $StochK$ کمتر از 0.09 باشد، (شاخه سمت چپ) با احتمال 10% درصد جهت حرکت سهم افزایشی خواهد بود. (18% درصد از مشاهدات). به بیان دیگر، احتمال اینکه در روز آینده قیمت سهم کاهشی باشد 90% درصد است. اما اگر $0.091 < StochK < 0.67$ ، آنگاه بسته به مقادیری که $MACD$ و RSI خواهند داشت، رده‌بندی مطابق با شاخه‌های وسطی درخت تصمیم انجام و مشابه دو حالت قبل تفسیر می‌شود.

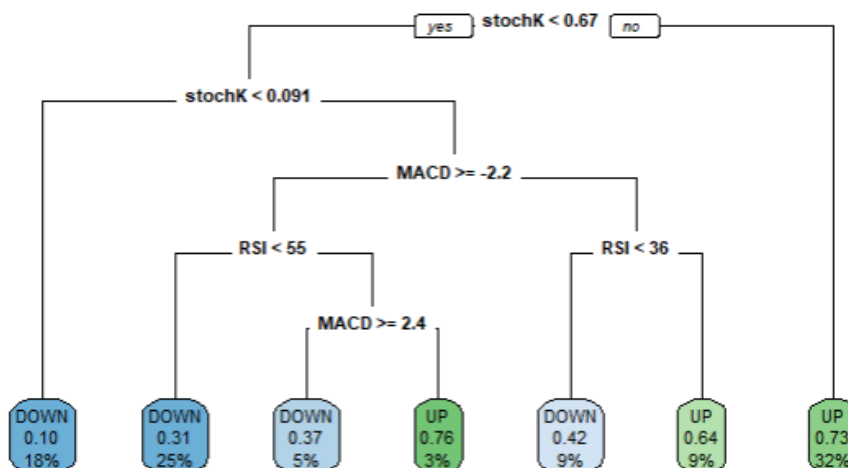
برای مقایسه رده‌بندی این مدل با رده‌بندی واقعی مشاهدات، باید همانند روش قبلی ماتریس درهم‌ریختگی رسم شود. **جدول ۵** این مقایسه را انجام داده است.

استفاده از درخت تصمیم مزیت‌ها و معایبی دارد. تفسیر و توضیح نتایج برای افراد در این روش با توجه به اینکه براساس شکل انجام می‌شود، ساده‌تر از روش‌هایی همچون رگرسیون لوژستیک یا KNN است. اما درمقابل در بحث پیش‌بینی (در مدل‌های رگرسیونی) درخت تصمیم قدرت کمتری نسبت به روش‌های پیش‌گفته دارد. در این روش می‌توان تعیین کرد که در ساخت مدل، اهمیت هر کدام از متغیرها چقدر بوده است. ضریب اهمیت هر متغیر براساس میزان تأثیری که متغیر مورد نظر در کاهش نرخ خطای رده‌بندی نادرست دارد، به‌دست می‌آید. **جدول ۶** میزان اهمیت هر متغیر را در

داده‌های شرکت بیمه سامان نشان می‌دهد. براساس این جدول، در 159 روز جهت حرکت سهم کاهشی بوده و مدل KNN هم آن را به‌صورت کاهشی پیش‌بینی کرده است. به همین ترتیب در 125 روز جهت حرکت سهم رو به بالا بوده است و مدل هم همین را پیش‌بینی کرده است. بنابراین، در $125+159$ مورد، پیش‌بینی به‌صورت درست انجام شده است. پس با توجه به معادله دقت، مدل KNN با $k=23$ ، دقتی به میزان 0.726 را نشان می‌دهد. یعنی این مدل در پیش‌بینی جهت حرکت قیمت سهم بیمه سامان در روز آینده، حدود 73% درصد دقت دارد. سایر شاخص‌های مورد اشاره در بخش قبل در **جدول ۹** آمده‌اند.

درخت تصمیم

در این بخش، با استفاده از بسته $rpart$ در نرم‌افزار R درخت تصمیم مربوط به داده‌های بیمه سامان را رسم کردیم. این نمودار در **شکل ۴** نمایش داده شده است. همان‌طور که در **شکل ۴** مشاهده می‌شود، از بین 6 متغیر حاضر در مسئله تنها سه متغیر $StochK$ ، $MACD$ و RSI به‌عنوان متغیرهای مؤثر در رده‌بندی با این روش در مدل باقی مانده‌اند. طبق این درخت تصمیم، اگر در یک روز $StochK$ بزرگ‌تر از 67% درصد باشد، با احتمال 73% درصد در روز آینده شاهد افزایش قیمت سهم خواهیم بود. 32% درصد از مجموعه داده‌های



شکل ۴: نمودار درخت تصمیم برای داده‌های شرکت بیمه سامان
Figure 4: Decision tree for Saman insurance company data

جدول ۵: ماتریس درهم‌ریختگی براساس مدل درخت تصمیم برای داده‌های آزمایشی شرکت بیمه سامان
Table 5: Confusion matrix of the decision tree for testing data of the Saman insurance company

پیش‌بینی مدل prediction		نتیجه واقعی observed	
dt=1	dt=0	dt=0	dt=1
59	142		
131	59		

جدول ۶: میزان اهمیت هر کدام از متغیرهای مستقل در ساخت درخت تصمیم
Table 6: Importance measures for each independent variables in the decision tree

6	5	4	3	2	1	ترتیب اهمیت Importance order
Lag1	OPEN	ADX	MACD	RSI	StochK	متغیر variable
10.03	11	16.31	53.44	95.78	151.25	ضریب اهمیت Importance coefficient

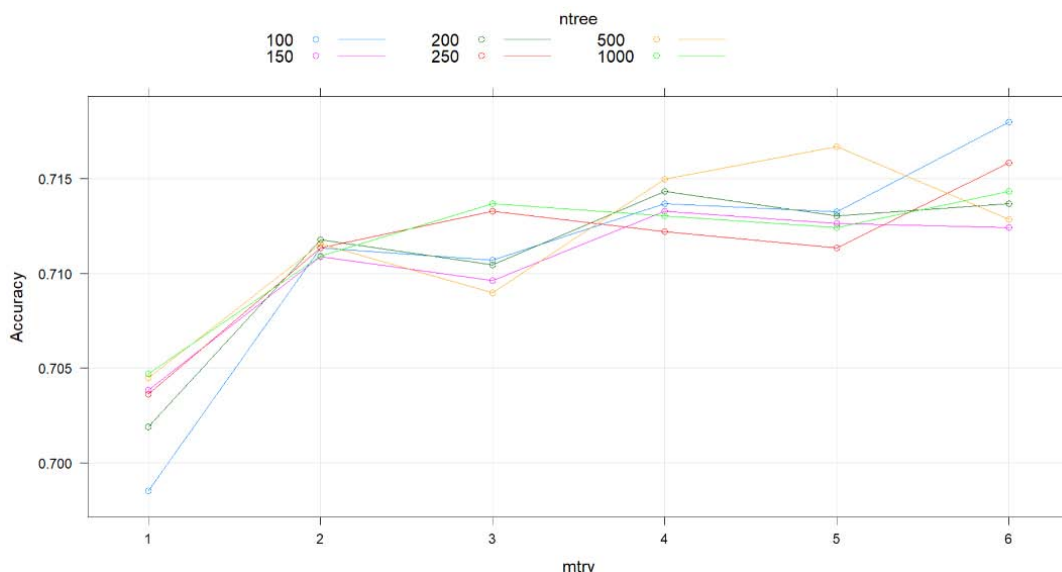
جنگلی از درختان تصمیم می‌گردد که بیشترین دقت را در رده‌بندی ایجاد کند. بنابراین، تعداد درختان اولیه برای رشد جنگل (ntree) و تعداد متغیرهایی که برای شروع ساخت جنگل به صورت تصادفی انتخاب می‌شوند (mtry) اهمیت دارد.

برای یافتن مقدار مناسب برای دو کمیت ntree و mtry از روش اعتبارسنجی متقابل استفاده می‌شود. به این صورت که مقادیر مختلف این دو کمیت در نظر گرفته شده و مدل‌های جنگل تصادفی براساس داده‌های آموزشی در یک الگوریتم تکرار ساخته می‌شوند. هر کدام که بیشترین دقت را ایجاد کردند، به‌عنوان مقادیر قابل‌استفاده در

خالص‌سازی درخت تصمیم شکل ۴ بیان می‌کند. طبق این جدول، متغیرهای Stochk، RSI و MACD اهمیت بیشتری نسبت به سایر متغیرها دارد که در شکل درخت تصمیم هم مشخص است.

جنگل تصادفی

در این بخش با استفاده از بسته RandomForest در نرم‌افزار R، الگوریتم را اجرا کردیم. این مدل در حقیقت تعمیمی از درخت تصمیم است که در آن با شروع از تعداد ثابتی درخت و با در نظر گرفتن مجموعه‌ای از متغیرهای مستقل، الگوریتم به صورت تصادفی به دنبال



شکل ۵: نمودار دقت جنگل های تصادفی با مقادیر مختلف $mtry$ و $ntree$ در داده های شرکت بیمه سامان
Figure 5: The accuracy of the random forests with different values of $ntree$ and $mtry$ for Saman insurance company data

جدول ۷: مقایسه رده بندی براساس واقعیت و براساس مدل جنگل تصادفی برای داده های آزمایشی شرکت بیمه سامان
Table 7: Confusion matrix of the random forest for testing data of the Saman insurance company

پیش بینی مدل prediction		نتیجه واقعی observed	
dt=1	dt=0	dt=0	dt=1
44	161	161	44
115	71	71	115

جدول ۸: میزان اهمیت هر کدام از متغیرهای مستقل در ساخت جنگل تصادفی
Table 8: Importance measures for each independent variables in the random forest

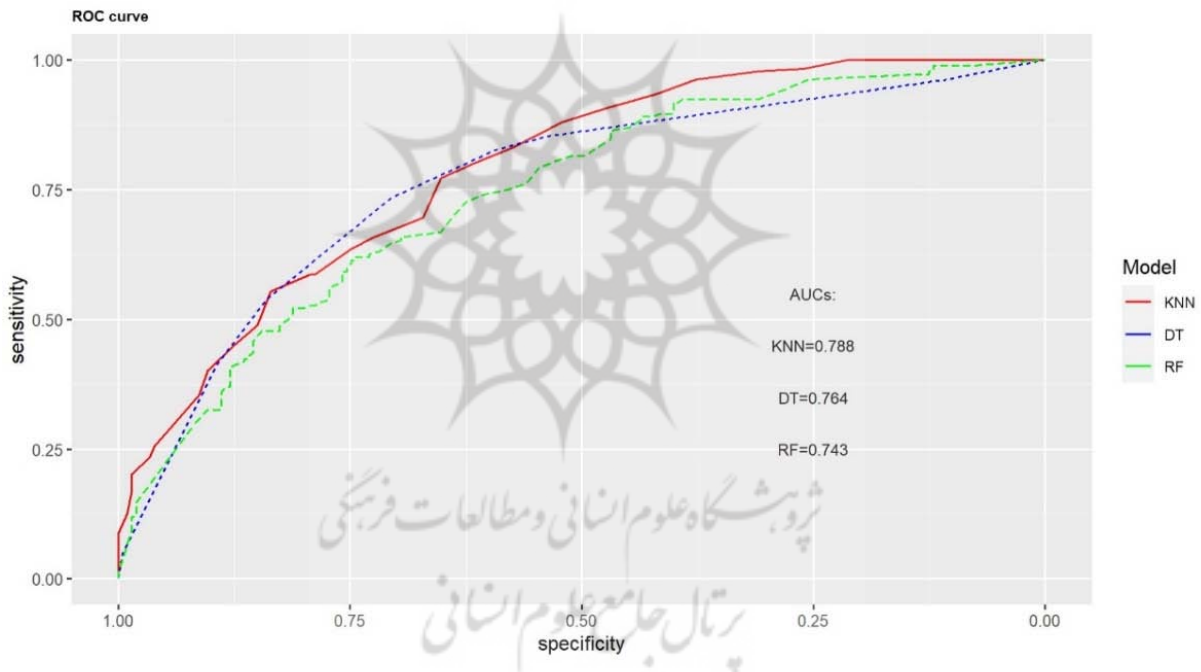
6	5	4	3	2	1	ترتیب اهمیت Importance order
ADX	Lag1	MACD	OPEN	RSI	StochK	متغیر variable
0	0	3.64	5.90	42.16	100	ضریب اهمیت Importance coefficient

نمود که کدام متغیرها نقش مهم تری در ساخت جنگل دارند. جدول ۸، متغیرها را بر حسب اهمیت آنها نشان می دهد. براساس این جدول سه متغیری که از اهمیت بیشتری در ساخت مدل جنگل تصادفی دارند، متغیرهای RSI، StochK و OPEN هستند. در نهایت، در جدول ۹ به مقایسه دقت رده بندی مدل های مورد مطالعه در این مقاله پرداخته ایم. همان طور که ملاحظه می شود، مدل KNN دارای بیشترین دقت رده بندی در بین مدل های مورد مطالعه است. ضمناً همان طور که قبلاً اشاره شد، مهم ترین متغیرها در ساخت مدل های

الگوریتم به کار گرفته می شوند. شکل ۵، نمودار دقت مدل را برای مقادیر مختلف این دو کمیت نشان می دهد. براساس این نمودار، $ntree=100$ و $mtry=6$ در داده های آموزشی دقتی در حدود ۷۲ درصد ایجاد می کند که بیشتر از سایر مقادیر است. جدول ۷، نتایج حاصل از رده بندی با چنین مدلی را در داده های آزمایشی نشان می دهد. همان طور که از جدول مشخص است، دقت رده بندی با استفاده از جنگل تصادفی در حدود ۶۷ درصد می باشد. در این حالت نیز می توان همچون مدل درخت تصمیم مشخص

جدول ۹: میزان دقت رده‌بندی در مدل‌های مختلف رده بندی در داده‌های شرکت بیمه سامان
Table 9: The accuracy of different classification models for the Saman insurance company data

جنگل تصادفی Random forest	درخت تصمیم Decision tree	KNN	مدل model
0.7059	0.6982	0.7263	دقت Accuracy
(0.6580, 0.7506)	(0.6501, 0.7432)	(0.6632, 0.8027)	فاصله اطمینان دقت Confidence interval
0.7232	0.6894	0.7485	صحت Precision
0.6182	0.6894	0.6578	پوشش Recall
0.6666	0.6894	0.7002	معیار F F.score



شکل ۶: نمودار منحنی مشخصه عملکرد برای سه مدل knn، درخت تصمیم و جنگل تصادفی به همراه مقادیر سطح زیر منحنی برای هر کدام از آنها (AUC) در داده‌های بیمه سامان

Figure 6: The ROC curves for the KNN, decision tree and random forest and their AUCs for Saman insurance data

درخت تصمیم و جنگل تصادفی متغیرهای StochK و RSI هستند. شکل ۶، نمودار منحنی مشخصه عملکرد را برای سه مدل KNN، درخت تصمیم و جنگل تصادفی نشان می‌دهد. واضح است که این منحنی برای جنگل تصادفی پایین‌تر از دو منحنی دیگر است. از طرفی این منحنی برای مدل KNN در اغلب موارد نسبت به منحنی

درخت تصمیم بالاتر قرار گرفته است. همچنین مقادیر سطح زیر منحنی نمودار برای هر کدام از مدل‌ها در داخل نمودار مشخص شده است. هرچه این مقدار به یک نزدیک‌تر باشد، به معنای دقیق‌تر بودن مدل است. همان‌طور که مشخص است این مقدار برای مدل KNN بزرگتر از دو مدل دیگر است.

جمع بندی و پیشنهادات

در این مقاله چند روش رده‌بندی آماری را معرفی کردیم و با استفاده از آنها به پیش‌بینی جهت حرکت قیمت سهام شرکت‌های بیمه بورسی پرداختیم. به‌طور متمرکز، صرفاً بیمه سامان را مورد تجزیه و تحلیل قرار دادیم و مشخص شد که در بین مدل‌های مورد مطالعه روش KNN دقت بیشتری نسبت به سایر مدل‌های رده‌بندی دارد. از آنجا که در این مقاله تمرکز نویسندگان بر روش‌های یادگیری آماری مبتنی بر درخت بود، به سایر روش‌ها همچون شبکه‌های عصبی، بیز ساده و بوستینگ اشاره نشده است. این روش‌ها می‌توانند در مقاله‌ای دیگر برای مقایسه با روش‌های بیان‌شده در این تحقیق مورد توجه باشند. همچنین در این تحقیق مدل‌های پارامتری همچون رگرسیون لوژستیک و ممیزی خطی فیشر به دلیل وجود همخطی در متغیرهای مستقل مورد مطالعه قرار نگرفتند. یک راه‌حل برای مشکل همخطی، استفاده از روش‌های کاهش داده همچون تحلیل مؤلفه‌های اصلی است. در این صورت، متغیرهای مستقل به ترکیب‌های خطی ناهمبسته تبدیل می‌شوند. این مسئله در پژوهشی دیگر تحت بررسی می‌باشد و به‌صورت جداگانه ارائه خواهد شد.

در این مطالعه شرکت بیمه سامان به‌عنوان یکی از قدیمی‌ترین شرکت‌های بیمه بورسی، مورد بررسی قرار گرفت. به‌طور مشابه می‌توان این تحقیق را برای سایر شرکت‌های بیمه و بازارهای سرمایه‌ای دیگر هم مورد مطالعه قرار داد. علاوه بر این، متغیر رده‌بندی در این تحقیق برای پیش‌بینی قیمت سهام شرکت‌های بیمه استفاده شده است. این مدل می‌تواند به مشتریان بیمه در تعیین میزان سرمایه‌گذاری در سهام یک شرکت کمک کند. اما در صنعت بیمه متغیرهای دیگری هم وجود دارند که می‌توانند در روش‌های رده‌بندی مورد استفاده قرار گیرند. به‌عنوان مثال، می‌توان مدلی طراحی کرد که بتواند اعتبار یا عدم اعتبار یک مشتری را برای داشتن یک بیمه مشخص همچون بیمه عمر یا بیمه شخص ثالث مورد بررسی قرار دهد. همچنین طراحی یک سامانه برای استفاده از این مدل‌ها در اعتبارسنجی مشتریان بیمه پیشنهاد می‌شود.

مشارکت نویسندگان

جمع‌آوری و پاکسازی داده‌ها و بررسی پیشینه تحقیق بر عهده

نویسنده دوم بوده است. تحلیل داده‌ها و ویرایش مقاله بر عهده نویسنده اول است.

تشکر و قدردانی

نویسندگان مقاله از پیشنهادات مفید سردبیر و داوران محترم نشریه پژوهش‌نامه بیمه که در بهبود نتایج این پژوهش کمک شایانی کرد، تشکر می‌نمایند.

تعارض منافع

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

دسترسی آزاد

کپی‌رایت نویسنده(ها) ©2023 این مقاله تحت مجوز بین‌المللی Creative Commons Attribution 4.0 اجازه استفاده، اشتراک‌گذاری، اقتباس، توزیع و تکثیر را در هر رسانه یا قالبی مشروط به درج نحوه دقیق دسترسی به مجوز CC منوط به ذکر تغییرات احتمالی بر روی مقاله می‌باشد. لذا به استناد مجوز مذکور، درج هرگونه تغییرات در تصاویر، منابع و ارجاعات یا سایر مطالب از اشخاص ثالث در این مقاله باید در این مجوز گنجانده شود، مگر اینکه در راستای اعتبار مقاله به اشکال دیگری مشخص شده باشد. در صورت عدم درج مطالب مذکور و یا استفاده فراتر از مجوز فوق، نویسنده ملزم به دریافت مجوز حق نسخه‌برداری از شخص ثالث می‌باشد.

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


یادداشت ناشر

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

منابع

- Almasarweh, M.; Alwadi, S., (2018). ARIMA model in predicting banking stock market data. *Mod. Appl. Sci.*, 12(11): 309-333 **(25 Pages)**.
- Asghari oskoei, M.R.; Khanizadeh, F.; Bahador, A., (2020). Application of data mining through machine learning algorithms to study effect of car features in predicting financial claim of motor third party liability insurance. *Iran. J. Insur. Res.*, 9(1): 33-66 **(34 Pages)**. [In Persian]
- Başoğlu Kabran, F.; Ünlü, K.D., (2021). A two-step machine learning approach to predict S&P 500 bubbles. *J. Appl. Stat.*, 48(13/15): 2776-2794 **(19 Pages)**.
- Bing, Y.; Hao, J.K.; Zhang, S.C., (2012). Stock market prediction using artificial neural networks. *Adv. Engine. Forum.*, 6(7): 1055-1060 **(6 Pages)**.
- Breiman, L., (2001). Random forests. *Machine Learning.*, 45(1): 5-32 **(28 Pages)**.
- 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 **(4 Pages)**.
- Ghorbani, H.; Ghanbarzadeh, M.; Ofoghi, R., (2022). Investigating the churn of life insurance customers using data mining methods (A case study: One of the Iran's insurance companies). *Iran. J. Insur. Res.*, 11(4): 305-320 **(16 Pages)**. [In Persian]
- Hao, J.; Priestley, J.L., (2016). A comparison of machine learning techniques and logistic regression method for the prediction of Past-Due amount. Phd thesis.
- Huang, W.; Nakamori, Y.; Wang, S.Y., (2005). Forecasting stock market movement direction with support vector machine. *Comput. Oper. Res.*, 32(10): 2513-2522 **(10 Pages)**.
- Izadparast, S.M.; Farahi, A.; Fathnejad, F.; Teimourpour, B., (2012). Using data mining techniques to predict the detriment level of car insurance customers. *Iran. J. Inf. Process. Manag.*, 27(3): 699-722 **(24 Pages)**. [In Persian]
- James, G.; Witten, D.; Hastie, T.; Tibshirani, R., (2013). An introduction to statistical learning. Springer.
- Kara, Y.; Boyacioglu, M.A.; Baykan, Ö.K., (2011). Predicting direction of stock price index movement using artificial neural networks and support vector machines: The sample of the Istanbul stock exchange. *Expert Syst. Appl.*, 38(5): 5311-5319 **(9 Pages)**.
- Khamesian, F.; Esna-Ashari, M.; Dei Ofosu-Hene, E.; Khanizadeh, F., (2022). Risk classification of imbalanced data for car insurance companies: Machine learning approaches. *Int. J. Math. Model. Comp.*, 12(3): 153-162 **(10 Pages)**.
- Kim, K.J., (2003). Financial time series forecasting using support vector machines. *Neurocomputing.*, 55(1/2): 307-319 **(13 Pages)**.
- Kotsiantis, S.B., (2013). Decision trees: a recent overview. *Artif. Intell. Rev.*, 39(1): 261-283 **(23 Pages)**.
- Mehrabanpour, M.; Azar, A.; Shahrami Babakan, M., (2022). Stock price forecasting by presenting a hybrid model using principal component analysis and rough set theory. *Mod. Res. Dec. Mak.*, 7(2): 137-167 **(31 Pages)**.
- Milosevic, N., (2016). Equity forecast: Predicting long term stock price movement using machine learning. *EconPapers*.
- Mintarya, L.N.; Halim, J.N.; Angie, C.; Achmad, S.; Kurniawan, A., (2023). Machine learning approaches in stock market prediction: A systematic literature review. *Procedia Comput. Sci.*, 1(1): 96-102 **(7 Pages)**.
- Nair, B.B.; Mohandas, V.P.; Sakthivel, N.R., (2010). A decision tree-rough set hybrid system for stock market trend prediction. *Int. J. Comput. Appl.*, 6(9): 1-6 **(6 Pages)**.
- Patel, J.; Shah, S.; Thakkar, P.; Kotecha, K., (2015). Predicting stock market index using fusion of machine learning techniques. *Expert. Syst. Appl.*, 42(4): 2162-2172 **(11 Pages)**.
- Roh, T.H., (2007). Forecasting the volatility of stock price index. *Expert. Syst. Appl.*, 33(4): 916-922 **(7 Pages)**.
- Rostamkhani, H.; Khodarahmi, B.; Jahanshad, A., (2021). Optimal stock selection using bat and random forest algorithm. *Financ. Eng. Portf. Manag.*, 12(1): 461-480 **(20 Pages)**. [In Persian]
- Rouf, N.; Malik, M.B.; Arif, T.; Sharma, S.; Singh, S.; Aich, S.; Kim, H.C., (2021). Stock market prediction using machine learning techniques: A decade survey on methodologies, recent developments, and future directions. *Electronics.*, 10(21): 2717-2742 **(26 Pages)**.
- Shah, D.; Isah, H.; Zulkernine, F., (2019). Stock market analysis: A review and taxonomy of prediction techniques. *Int. J. Financ. Stud.*, 7(2): 26-42 **(17 Pages)**.
- Sharma, D., (2012). Improving the art, craft and science of economic credit risk scorecards using random forests: Why credit scorers and economists should use random forests. *J. Acad. Bank. Stud.*, 11(1): 1-33 **(33 Pages)**.
- Singh, S.; Madan, T.K.; Kumar, J.; Singh, A.K., (2019). Stock market forecasting using machine learning: Today and tomorrow. 2nd International Conference on Intelligent Computing, Instrumentation and Control Technologies., 1(1): 738-745 **(8 Pages)**.
- Singh, S.; Parmar, K.S.; Kumar, J., (2021). Soft computing model coupled with statistical models to estimate future of stock market. *Neural. Comput. Appl.*, 33(13): 7629-7647 **(19 Pages)**.
- Wilder, J.W., (1978). New concepts in technical trading systems. Argis.
- 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. Comput. Appl.*, 8(3): 160-172 **(13 Pages)**.
- Zhang, Y.; Wu, L., (2009). Stock market prediction of S&P 500 via combination of improved BCO approach and BP neural network. *Expert. Syst. Appl.*, 36(5): 8849-8854 **(6 Pages)**.

AUTHOR(S) BIOSKETCHES	معرفی نویسندگان
<p>مصطفی طامندی، استادیار گروه آمار، دانشکده علوم ریاضی، دانشگاه ولی عصر (عج) رفسنجان، رفسنجان، ایران</p> <ul style="list-style-type: none">▪ Email: tamandi@vru.ac.ir▪ ORCID: 0000-0003-0056-4642▪ Homepage: https://profile.vru.ac.ir/~tamandi <p>مجیا عسکری پور، کارشناسی ارشد آمار، معاونت فنی بیمه‌های غیر زندگی و کارشناس مسئول تحلیل داده، بیمه سامان، تهران، ایران</p> <ul style="list-style-type: none">▪ Email: m.askari@samaninsurance.ir▪ ORCID: 0009-0008-8689-2922▪ Homepage: https://www.si24.ir/	

HOW TO CITE THIS ARTICLE	
<p><i>Tamandi, M.; Askaripour, M., (2023). Comparison of the accuracy of statistical learning algorithms in predicting of the stock price movement of Saman Insurance Company as a listed insurance company. Iran. J. Insur. Res., 12(2): 181-196.</i></p> <p>DOI: 10.22056/ijir.2023.03.02</p> <p>URL: https://ijir.irc.ac.ir/article_160296.html?lang=en</p>	

