پیش بینی جهت حرکت قیمت سهام شرکت های بیمه بورسی با استفاده از الگوریتم های رده بندی: مقایسه روش های رگرسیون لجستیک، KNN ، درخت تصادفی و جنگل تصادفی.

مصطفی طامندی [a]، محیا عسکری پور [b]
[a]عضو هیئت علمی ، دانشگاه ولیعصر، رفسنجان
[b]کارشناس مسئول تحلیل داده ، بیمه سامان، تهران
نویسنده مسئول: مصطفی طامندی **09155064659** ، (tamandi.m@gmail.com)

چکیده: فعالان بورس درصدد دستیابی و به کارگیری روشهایی هستند تا بتوانند با پیش بینی آتی قیمت سهام، سود سرمایه خود را افزایش دهند. بنابراین، ضروری به نظر میرسد که روشهای مناسب و متکی به اصول علمی در تعیین قیمت آینده سهام در اختیار افراد سرمایه گذار قرار گیرد. تاکنون روشهای مختلفی جهت نیل به این هدف معرفی شده اند که اغلب روشهای آماری و داده کاوی هستند. در پژوهش حاضر با استفاده از رویکردهای رگرسیون لجستیک، KNN، درخت تصمیم و جنگل تصادفی که در زمره روشهای رده بندی یادگیری آماری می باشند به مشاهده روند قیمت سهام شرکت بیمه سامان و مقایسه دقت هر کدام از روشها پرداخته شده است. نتیجه ی پژوهش بر روی این داده ها طی سال های 2012 تا 2022 نشان میدهد که روش رگرسیون لجستیک در برآورد روند قیمت سهام شرکت بیمه سامان نسبت به روشهای ذکر شده دیگر از دقت بالاتری برخوردار است.

کلمات کلیدی: تغییرات قیمتی سهام؛ رگرسیون لجستیک؛ درخت تصمیم؛ جنگل تصادفی؛ مدل KNN.

1. مقدمه

برای مطالعه تغییرات قیمتی یک سهم معمولا از قیمت پایانی معامله در هر روز استفاده می شود. بر اساس این متغیر می توان در مورد وضعیت آینده سهم پیش بینی انجام داد. یکی از روش های پیش بینی آن است که قیمت پایانی امروز یک سهم را با قیمت پایانی روز قبل آن مقایسه کنیم. فرض کنید قیمت پایانی یک سهم را در روز t با c_t نمایش دهیم. در این صورت برای مقایسه فوق می توان از شاخص زیر استفاده کرد:

$$d_t = \begin{cases} 1 & \text{if } c_t \geqslant c_{t-1} \\ 0 & \text{if } c_t < c_{t-1} \end{cases}$$

اگر d_t برای روزهای زیادی در یک سهم مقدار یک را به خود بگیرد، جهت سهم افزایشی است و می توان نسبت به خرید آن تصمیم گیری نمود و اگر صفر باشد، می توان نسبت به فروش سهم اقدام نمود. با توجه به دو مقداری بودن این متغیر، می توان از آن به عنوان متغیر رده بند استفاده کرد. بنابراین اگر اطلاعاتی در مورد سهم داشته باشیم که بتوان از آنها به عنوان متغیرهای مستقل استفاده کرد، می توان با استفاده از روش های رده بندی، جهت حرکت قیمت پایانی یک سهم را پیش بینی کنیم. کارا و همکاران (2011) با در نظر گرفتن مشخصه هایی همچون میانگین متحرک ده روزه، گشتاور، شاخص قدرت نسبی (RSI) و ... جهت حرکت شاخص کل استانبول را با استفاده از روش های یادگیری ماشین ANN و SVM پیش بینی کرد.

رستمخانی و همکاران(1400) به مقایسه دو الگوریتم خفاش و جنگل تصادفی برای انتخاب بهینه سهام پرداختند. آنها اشاره کردند که رفتار سهام در بازار، مانند بسیاری از پدیده های طبیعی، رفتار غیر خطی است و مدلهای خطی از تشخیص صحیح رفتار غیر خطی عاجز هستند و تنها می توانند بخش خطی رفتار را خوب تشخیص دهند. بنابر این، نیاز به الگوها و مدلهای غیر خطی برای شناسایی رفتار سهام تأثیر بسزایی در پیش بینی آتی سهام و اتخاذ تصمیم مناسب دارد. بنابراین، با در نظر داشتن گرایشها و ترجیحات مختلف سرمایه گذاران، یافتن روشی برای انتخاب یک مجموعه مناسب از اوراق بهادار که از طریق آن بتوان بر عدم اطمینان ها و ترجیحات مختلف افراد غلبه کرد، ضروری به نظر می رسد. با توجه به سنجش و این الگوریتم ها در انتخاب سهام، اثبات کارایی آنها می تواند سرمایه گذاران را تشویق کند تا از طریق مدل مذکور به سرمایه گذاری با ریسک کمتر براساس الگویی دقیق و با دقت بالا سوق داده شوند. ماهیت الگوریتم جنگل تصادفی نیاز به آموزش و انتخاب ویژگیها دارد که باعث میشود سرعت الگوریتم پایینتر باشد و زمان همگرایی را بالا میبرد. (رستمخانی و همکاران 1400)

گل ارضی و یعقوبی (98) برای تبیین تغییرات قیمت سهام پذیرفته شده در بورس اوراق بهادار تهران طی سال های 89 تا 96 از روش های رگرسیون چند متغیره استفاده کردند. در صنعت بیمه، اصغری و همکاران (99) با استفاده از الگوریتم های یادگیری ماشین به بررسی تاثیر ویژگیهای خودرو در پیش بینی ریسک خسارت مالی در بیمه شخص ثالث پرداختند. مدل های مورد استفاده آنها الگوریتم درخت تصمیم، نایو بیز و شبکه عصبی بودند. آنها نتیجه گرفتند که میزان تاثیرگذاری متغیرها در وقوع خسارت به ترتیب اولویت عبارتند از: نوع خودرو، نوع پلاک، سن خودرو و گروه خودرو.

در این مطالعه به بررسی روند قیمت سهام شرکت های بیمه بورسی و مقایسه دقت روشهای رگرسیون لجستیک، درخت تصمیم، جنگل تصادفی و مدل KNN که از جمله روش های یادگیری آماری معروف هستند، پرداخته شده است. در بخش دوم این روش ها را معرفی خواهیم کرد. در بخش سوم داده های مورد استفاده در این مقاله را توصیف می کنیم و در بخش چهارم به تجزیه و تحلیل آماری و ارائه نتایج خواهیم پرداخت.

2. معرفی مدل های رده بندی:

در این بخش چهار روش رده بندی را که در یادگیری آماری مورد استفاده قرار می گیرد، معرفی می کنیم. لازم به ذکر است که مدل های دیگری هم برای رده بندی داده ها وجود دارند، اما مدل های حاضر در این مقاله به این دلیل انتخاب شده اند که در ساخت آنها از روش های آماری و احتمالاتی استفاده شده است. سایر روش های رده بندی مثل SVM و شبکه عصبی عمدتا با استفاده از روش های بهینه سازی ریاضی و با یادگیری از طریق خود داده ها ساخته می شوند.

1.2 رگرسیون لجستیک

یک روش کلاسیک برای رده بندی، رگرسیون لجستیک می باشد. مدل های رگرسیونی عموما شامل تعدادی متغیر مستقل (کمی و کیفی) و یک متغیر پاسخ هستند. این متغیر پاسخ مقادیر کمی دارد و هدف از رگرسیون پیش بینی مقدار این متغیر پاسخ بر اساس مقادیر متغیرهای مستقل است. اکنون اگر متغیر پاسخ به صورت کیفی و مقادیر آن به صورت صفر و یک یا بله و خیر باشد و مسئله ما مدل بندی احتمال صفر یا یک بودن متغیر پاسخ باشد، به چنین مدلی رگرسیون لجستیک گفته می شود. در تحقیق حاضر مدل رگرسیون لجستیک به صورت زیر خواهد بود:

$$P(x) = P(Y = 1 \mid \mathbf{X}=\mathbf{x}) = \frac{e^{\beta_0+\sum_{i=1}^{p}\beta_i x_i}}{1 + e^{\beta_0+\sum_{i=1}^{p}\beta_i x_i}}$$

در این معادله، $\mathbf{X}$ متغیرهای مستقل موجود در مسئله و β_i پارامترهای مدل هستند که باید برآورد شوند. در رگرسیون لجستیک پیش بینی مقدار متغیر پاسخ یا احتمال آن موضوعیت ندارد و هدف رده بندی مشاهدات بر اساس تخمینی است که برای احتمال بالا به دست می آید. به طور پیش فرض اگر برای یک مشاهده مثل x ، مشخص شود که $P(x) > 0.5$ آنگاه آن مشاهده را در گروهی رده بندی می کنیم که شامل روزهایی باشد که جهت قیمت افزایشی است.

2.2. k- نزدیک ترین همسایه ها

در روش KNN، برای رده بندی مشاهدات در دو گروه به این صورت عمل می شود. ابتدا یک مشاهده x در نظر گرفته می شود و k مشاهده یا همسایه که نزدیک ترین فاصله با x نسبت به سایر مشاهدات را دارند در گروهی قرار می گیرند که x قرار گرفته است. این کار با همه مشاهدات انجام می گیرد و این فرایند به صورت تکراری و با انتخاب های تصادفی آن قدر تکرار می شود تا نهایتا همه مشاهدات در دو گروه رده بندی شوند.

انتخاب مقدار مناسب k نقش مهمی در دقت مدل به دست آمده دارد. برای انتخاب k، مدل های KNN متفاوت را روی مجموعه آزمایشی برای $k=1,...,30$ اجرا می کنیم. برای هر کدام از مدل ها، خطای رده بندی یا عکس آن یعنی دقت رده بندی را محاسبه می کنیم و مقداری از k را که به ازای آن کمترین خطا یا بیشترین دقت به دست آمده باشد، به عنوان مقدار مناسب انتخاب می کنیم. سپس این مدل را که در داده های آزمایشی صحت خود را نشان داده است، برای مجموعه داده های آموزشی استفاده می کنیم.

در مدل رگرسیون لجستیک فرض می شود که رابطه بین متغیرهای مستقل و پاسخ خطی است. به همین دلیل در این مدل باند جدا کننده دو گروه به شکل خط است. اما در مدل KNN این باند می تواند غیرخطی باشد. بنابراین مدل KNN می تواند نسبت به رگرسیون لجستیک برتری داشته باشد. علاوه بر این مدل KNN یک مدل ناپارامتری است و نیازی به دانستن توزیع متغیرهای مستقل در این حالت نیست.

3.2 درخت تصمیم

یکی دیگر از روش های ناپارامتری برای رده بندی داده ها، درخت تصمیم است. یک درخت تصمیم شامل چند شاخه است که مشخص می کند هر مشاهده از مجموعه داده ها در هر کدام از ناحیه ها یا شاخه ها با چه احتمالی در یکی از گروه های متغیر پاسخ رده بندی می شود. بنابراین یک درخت تصمیم علاوه بر اینکه همانند روش های ذکر شده قبلی مشخص می کند که هر مشاهده در کدام کلاس رده بندی می شود، می تواند نسبت یا سهم هر کلاس در شاخه ای از متغیرها و نقش هر کدام را در این رده بندی تعیین کند. در درخت تصمیم همانند آنچه در روش خوشه بندی داده ها اتفاق می افتد از یک ریشه شروع می کنیم و با روش های تکراری شاخه های درخت تصمیم را مشخص می کنیم. جدا سازی شاخه ها از هم بر اساس نرخ خطای رده بندی انجام می شود. اگر p_{mk} نسبت مشاهداتی از شاخه m باشند که در کلاس k رده بندی می شوند، در این شاخه مشخص نرخ رده بندی نادرست از فرمول زیر محاسبه می شود:

$$E = 1 - p_{mk}$$

بنابراین یک شاخه جدید به درخت تصمیم اضافه می شود اگر مقدار E کوچک باشد. به بیان دیگر اگر ادغام دو شاخه در هم باعث کاهش نرخ خطای رده بندی می شود آنگاه این دو شاخه در هم ادغام می شوند و نهایتا یک درخت خالص به دست می آید. در مسئله درخت تصمیم هم از آنجا که با یک روش ناپارامتری سر و کار داریم باید داده ها را به دو مجموعه آزمایشی و آموزشی تقسیم کرده و با استفاده از داده های آزمایشی به مدل بهینه برسیم.

4.2 جنگل تصادفی

همان طور که در بخش قبل گفتیم روش درخت تصمیم برای رده بندی داده ها مدل مناسبی است. اما این مدل به شدت وابسته به متغیرهای مستقل است. اگر در مجموعه داده ها یک متغیر مستقل وجود داشته باشد که در رده بندی نسبت به بقیه قوی تر باشد روی درختی که در نهایت ساخته می شود تاثیرگذار خواهد بود. برای حل این مشکل می توان درخت هایی را با تعداد متفاوت از متغیرهای مستقل به صورت تصادفی ایجاد کرد و در بین آنها به دنبال حالتی باشیم که کمترین نرخ خطای رده بندی را ایجاد می کند. این درخت های تصادفی یک جنگل به وجود می آورند که به آن جنگل تصادفی می گویند. این جنگل مثل عملیات تکراری عمل می کند که در روش KNN هم به وسیله آن k مناسب را یافتیم. دقت مدل جنگل تصادفی به تعداد درختان و تعداد گره ها یا متغیرهای مستقلی که در ساخت مدل تاثیر گذار هستند، بستگی دارد. بنابراین با روش های تکراری می توان به جنگلی رسید که بیشترین دقت را داشته باشد.

3. داده های پژوهش:

در این تحقیق می خواهیم کاربرد روش های رده بندی را در تعیین روند افزایشی یا کاهشی در قیمت سهام شرکت های بیمه بورسی مورد مطالعه قرار دهیم. برای این منظور قیمت سهام چند شرکت بیمه بورسی را که بیش از 1000 روز از ارائه آنها در بورس گذشته است از طریق سایت شرکت مدیریت فناوری بورس تهران دریافت شده است. نمودار 1، درصد روزهایی که جهت حرکت قیمت سهم برای چند شرکت بیمه مثبت شده است (به این معنا که در معادله 1 داشته باشیم $d_t = 1$) نشان می دهد.

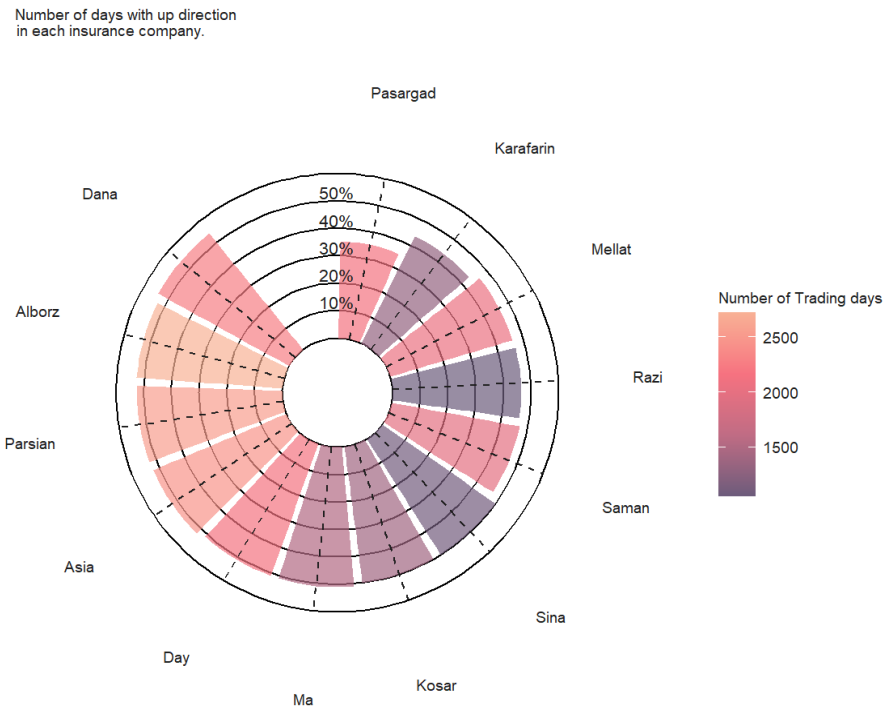
طبق این نمودار شرکت های بیمه دانا و البرز بیشترین روزهای افزایش قیمت سهم نسبت به روز قبل را داشته اند (حدود 52 درصد) و شرکت های بیمه پاسارگاد و کارآفرین کمترین افزایش قیمت سهم (حدود 34 و 41 درصد). رنگ نمودار نمایش دهنده تعداد روزهایی است که سهم هر شرکت در بورس مورد معامله قرار گرفته است. از این نظر شرکت های بیمه البرز و پارسیان بیشترین روز معامله و شرکت های رازی و سینا کمترین روزهای مشارکت در بورس را به خود اختصاص داده اند. به منظور خلاصه شدن مقاله، از بین این شرکت ها شرکت بیمه سامان را بعنوان شرکت مورد مطالعه انتخاب کرده ایم. داده های قیمت سهام شرکت بیمه سامان را از تاریخ 2012/07/14 تا 2022/01/03 از طریق سایت بورس دریافت کردیم. تعداد داده ها که به صورت روزانه ثبت شده است، بعد از حذف داده های گمشده 1951 مشاهده است. متغیرهای پیشگو مورد استفاده در این تحقیق به همراه تعریف هرکدام در جدول 1 آمده است.

جدول 1. متغیرهای پیشگوی مورد استفاده در مدل های رده بندی و تعریف آنها

شاخص	تعریف
OPENt	قیمت ابتدایی سهم در روز t
lag1t	$c_t - c_{t-1}$
StochKt	$\frac{c_t - L_{t-n}}{H_{t-n} - L_{t-n}}$
RSIt	$100 - \frac{100}{1 + (\sum_{i=0}^{n-1} U_{t-i} / \sum_{i=0}^{n-1} D_{t-i})}$
MACDt	$EMA(26)_t - EMA(12)_t$
where	$EMA(n)_t = \frac{2}{n+1} c_t + \frac{n-1}{n+1} EMA(n)_{t-1}$
ADXt	$\frac{(n-1)ADX_{t-1} + DX_t}{n}$
where	$DX_t = \frac{Di^+ - Di^-}{Di^+ + Di^-}$

در جدول 1، Lt-n و Ht-n به ترتیب کمترین و بیشترین قیمت سهم در n روز گذشته است..Dt و Utبه ترتیب میزان تغییر قیمت سهم در لحظه t به سمت پایین و به سمت بالاست. Di- و Di+ به ترتیب شاخص های جهت حرکت سهم به بالا یا پایین هستند که بر اساس متوسط پایین ترین و بالاترین قیمت دو روز متفاوت برای روز های پشت سر هم به دست می آید. (ویلدر، 1978)

در این مقاله قصد داریم مقدار d_t را که در معادله (1) به آن اشاره شده است، با توجه به شاخص هایی که در جدول 1 آمده و با استفاده از روش های رده بندی مورد اشاره در بخش قبل، برای بررسی جهت حرکت قیمت سهام شرکت بیمه سامان پیش بینی کنیم.



شکل 1. نمودار درصد جهت افزایشی در شرکت های بیمه بورسی

جدول 2، نشان دهنده میانگین شاخص های مورد مطالعه در شرکت بیمه سامان است. برای انجام الگوریتم های رده بندی مورد اشاره در فصل قبل، ابتدا باید داده های مربوط به شرکت بیمه سامان را به دو گروه آموزشی (Training data) و آزمایشی (Test data) تقسیم کنیم. معمولا داده های آموزشی 80 درصد از داده ها و در نتیجه داده های آزمایشی 20 درصد از کل داده ها هستند که به طور تصادفی انتخاب می شوند. تقسیم داده ها به این دو گروه برای بررسی دقت مدل انجام می شود. به این صورت که مدل با استفاده از داده های آموزشی ساخته شده و بر روی داده های آزمایشی مورد بررسی قرار می گیرد. در صورتی که مدل ساخته شده دارای دقت مناسبی در داده های آزمایشی باشد، می توان از این مدل برای رده بندی مشاهدات استفاده کرد.

جدول 2. اطلاعات توصیفی شرکت بیمه سامان

تعداد مشاهده	1951
میانگین قیمت پایانی	7062.29
میانگین MACD	0.9284-
میانگین RSI	48.245
میانگین ADX	25.837
میانگین StochK	0.4462

4. تجزیه و تحلیل آماری

4.1 رگرسیون لجستیک:

در این بخش یک مدل رگرسیون لجستیک را با توجه به آنچه در بخش روش شناسی گفتیم به داده ها برازش دادیم. ابتدا 6 متغیر که در بخش های قبل معرفی کردیم را به عنوان متغیرهای مستقل در نظر گرفتیم و یک رگرسیون لجستیک روی آنها انجام دادیم. جدول 3، برآورد پارامترهای مدل به همراه انحراف معیار آنها را برای پارامترهای بیمه سامان نشان می دهد. طبق این جدول، متغیرهای StochKt و MACDt در پیش بینی جهت حرکت قیمت سهام نسبت به سایر متغیرها وزن بیشتری دارند.

جدول 3. برآورد پارامترهای مدل رگرسیون لجستیک به همراه انحراف معیار آنها (در پرانتز) برای شرکت بیمه . علامت "-" به معنای معنادار نبودن آن برآورد در مدل مربوط بوده است.

پارامتر	سامان
عرض از مبدا	-5.9484 (0.5667)
OPENt	0.0011 (0.0002)
lag1t	0.0011- (0.0002)
MACDt	-0.2320 (0.0249)
RSIt	0.0991 (0.0133)
StochKt	1.5611 (0.3470)

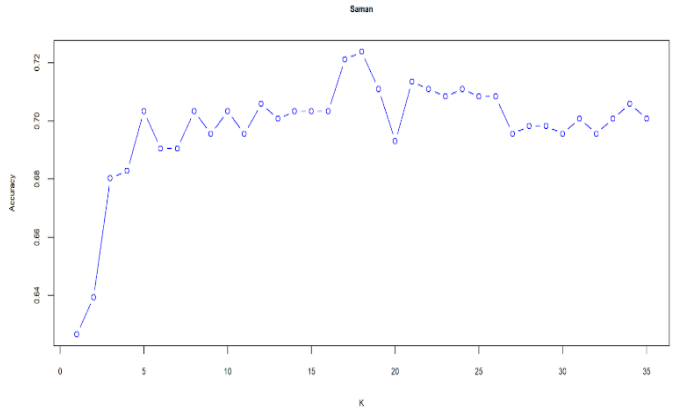
در مدل های رده بندی، دقت مدل در رده بندی درست مشاهدات اهمیت دارد. به این معنا که رده واقعی یک مشاهده با رده ای که توسط مدل پیش بینی می شود یکسان است یا خیر. هرچه در یک مدل تعداد مشاهداتی که به درستی رده بندی شده اند بیشتر باشد، دقت مدل بیشتر است. در جدول 4 تعداد رده بندی درست و نادرست برای داده های آزمایشی شرکت بیمه سامان با توجه به مدل رگرسیون لجستیک، آمده است. طبق این جدول، از 388 مشاهده آزمایشی برای شرکت بیمه سامان، در 174 روز هم در واقعیت جهت کاهشی در قیمت سهم داشتیم و هم مدل لجستیک استفاده شده چنین حالتی را پیش بینی کرده است. در 42 مورد خطا داشته ایم یعنی جهت واقعی به صورت کاهشی بوده است اما مدل جهت افزایشی را پیش بینی کرده است. به همین ترتیب می توان سطر دوم جدول شماره 4 را هم تفسیر کرد. همچنین بر اساس تعداد رده بندی درست در این جدول می توان دقت مدل را اندازه گیری نمود که در جدول شماره 8 به آن اشاره خواهیم کرد.

جدول 4. مقایسه رده بندی بر اساس واقعیت و بر اساس مدل رگرسیون لجستیک برای داده های آزمایشی شرکت های بیمه سامان

نتیجه واقعی	پیش بینی مدل	
	dt=0	dt=1
dt=0	174	42
dt=1	57	118

4.2 تحلیل KNN:

همان طور که قبلا گفتیم در این نوع رده بندی موضوع اصلی انتخاب k مناسب برای خوشه بندی نقاط است. در شکل 2 نمودار دقت مدل KNN در رده بندی برای مقادیر مختلف k رسم شده است.



شکل 2. نمودار دقت مدل های KNN برای شرکت بیمه سامان بر اساس مقادیر مختلف k

جدول شماره 5 نتایج رده بندی را با استفاده از مدل KNN در داده های شرکت بیمه سامان نشان می دهد. این جدول را می توان همچون جدول شماره 4 تفسیر نمود.

جدول 5. مقایسه رده بندی بر اساس واقعیت و بر اساس مدل KNN برای داده های آزمایشی شرکت بیمه سامان

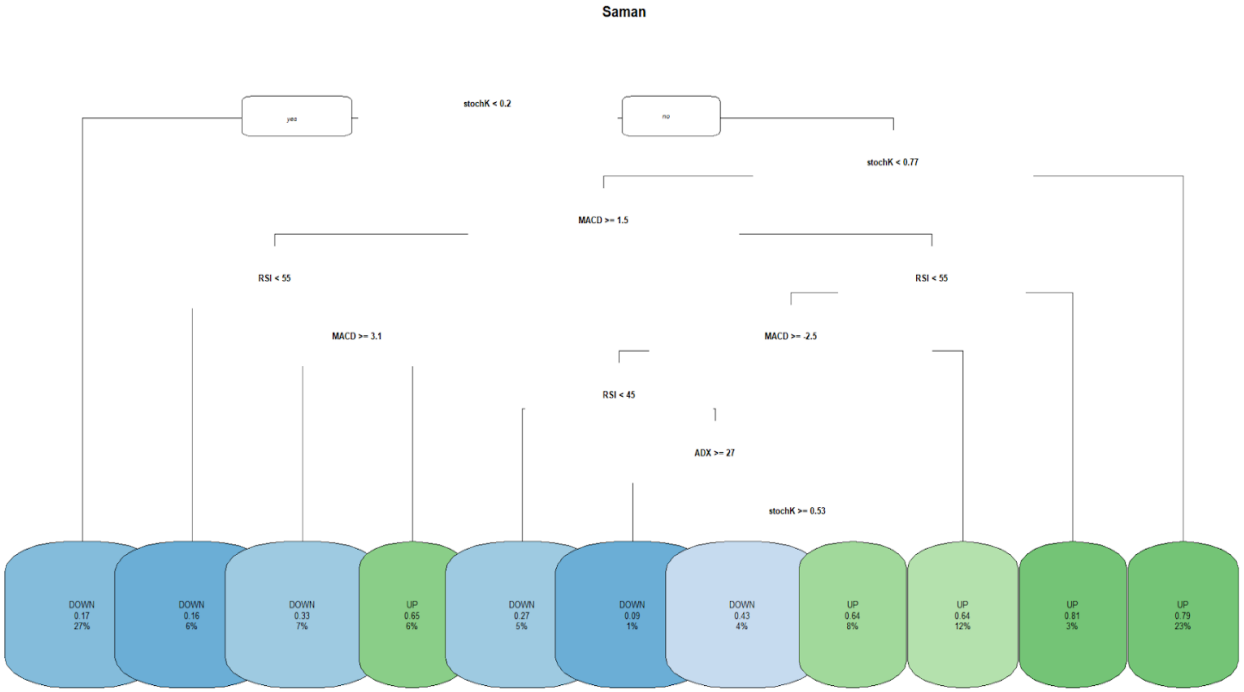
نتیجه واقعی	پیش بینی مدل	
	dt=0	dt=1
dt=0	151	65
dt=1	46	129

4.3 مدل درخت تصمیم:

همچون روش های قبلی در اینجا هم 80 درصد داده ها برای مجموعه آموزشی و 20 درصد باقیمانده برای تست مدل در نظر گرفته شد و با استفاده از پکیج rpart در نرم افزار R درخت تصمیم مربوط به داده ها را رسم کردیم. این نمودار برای شرکت سامان در شکل 3 نمایش داده شده است.

همان طور که در شکل 3 مشاهده می شود، از بین 6 متغیر حاضر در مسئله دو متغیر OPENt و lag1t در ایجاد و رده بندی توسط درخت تصمیم تاثیری ندارد. طبق این درخت تصمیم، در حالتی که شاخص StochK کمتر از 0.2 باشد با احتمال 17 درصد روز بعد شاهد کاهش قیمت خواهیم بود (متغیر پاسخ در حالت Down خواهد بود). داده های حاضر در این شاخه 27 درصد از مجموعه داده های آموزشی را شامل می شوند. از طرفی اگر این شاخص کمتر از 0.77 باشد ، شاخص MACDt هم بیشتر از 1.5 و شاخص RSI<55 باشد، با احتمال 16 درصد روز بعد شاهد کاهش قیمت خواهیم بود که 6 درصد از داده های آموزشی را شامل می شود. به همین ترتیب شاخه های دیگر این درخت را هم می توان تحلیل کرد. بر اساس این شکل ، شاخص های StochK و RSIدر تشکیل و تفسیر درخت تصمیم اهمیت بیشتری نسبت به بقیه شاخص ها دارند.

جدول شماره 6 نتایج رده بندی را با استفاده از درخت تصمیم شکل 3 در داده های شرکت بیمه سامان نشان می دهد. این جدول تعداد رده بندی درست و نادرست را بر اساس درخت تصمیم نشان می دهد.



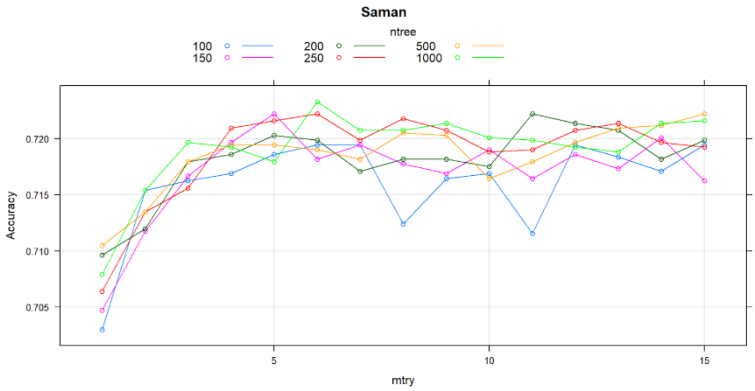
شکل 3. درخت تصمیم برای شرکت بیمه سامان

جدول 6. مقایسه رده بندی بر اساس واقعیت و بر اساس مدل درخت تصمیم برای داده های آزمایشی شرکت بیمه سامان

نتیجه واقعی	پیش بینی مدل	
	dt=0	dt=1
dt=0	154	65
dt=1	50	122

4.4 مدل جنگل تصادفی:

در این بخش با استفاده از پکیج RandomForest در نرم افزار R، الگوریتم را اجرا کردیم. در مدل های جنگل تصادفی همان طور که قبلا اشاره شد، ابتدا تعداد درختان (ntree) و تعداد گره ها (mtry) را با استفاده از داده های آزمایشی تعیین می کنیم. شکل 4 نمودار دقت جنگل تصادفی را برای مقادیر مختلف ntree و mtry نشان می دهد. در داده های بیمه سامان ، بیشترین دقت زمانی است که ntree=1000 و mtry=6 باشد. پس از مشخص شدن تعداد درختان جنگل و تعداد گره ها می توان از این مدل برای رده بندی داده ها آموزشی استفاده کرد.



شکل 4. مقایسه دقت رده بندی با استفاده از جنگل تصادفی برای شرکت بیمه مورد مطالعه برای تعیین تعداد درخت ها و گره ها

جدول شماره 7 نتایج رده بندی را با استفاده از جنگل تصادفی شکل 4 در داده های شرکت بیمه سامان نشان می دهد. این جدول تعداد رده بندی درست و نادرست را نشان می دهد.

جدول 7. مقایسه رده بندی بر اساس واقعیت و بر اساس مدل جنگل تصادفی برای داده های آزمایشی شرکت بیمه سامان

نتیجه واقعی	پیش بینی مدل	
	dt=0	dt=1
dt=0	103	107
dt=1	36	145

در جدول شماره 8، دقت مدل های مورد مطالعه در این مقاله در شرکت بیمه سامان نشان داده شده است. این دقت ها با محاسبه نسبت تعداد رده بندی های درست نسبت به کل مشاهدات آزمایشی با توجه به جدول های 4 تا 7 برای مدل های مختلف به دست آمده اسن. همان طور که ملاحظه می شود، مدل رگرسیون لجستیک با مقدار تقریبی 75 درصد دقت بیشتری در پیش بینی جهت حرکت قیمت سهام نسبت به سایر مدل ها دارد. این بدان معناست که رگرسیون لجستیک در 75 درصد روزها می تواند به درستی تشخیص دهد که روز بعد قیمت سهام روند افزایشی خواهد داشت یا کاهشی خواهد بود. بعد از رگرسیون لجستیک به ترتیب روش های KNN، درخت تصمیم و جنگل تصادفی قرار دارند.

جدول 8. مقایسه دقت رده بندی در مدل های مختلف برای شرکت های بیمه مورد مطالعه

مدل	دقت رده بندی
رگرسیون لجستیک	0.7468
KNN	0.7161
درخت تصمیم	0.7059
جنگل تصادفی	0.6343

5. نتیجه گیری

در این مقاله چند روش رده بندی آماری را معرفی کردیم و با استفاده از آنها به پیش بینی جهت حرکت قیمت سهام شرکت های بیمه بورسی پرداختیم. به طور متمرکز و با هدف کاهش صفحات مقاله صرفا بیمه سامان را مورد تجزیه و تحلیل قرار دادیم و مشخص شد که در بین مدل های مورد مطالعه روش

رگرسیون لجستیک دقت بیشتری نسبت به سایر مدل های رده بندی دارد. این روش ها را برای چند شرکت بیمه دیگر هم مورد استفاده قرار دادیم که نتایج مشابهی در آنها هم به دست آمده است.

6. منابع:

[1] اصغری اسکوئی ، م.، خانی زاده، ف.، جهانشاد، آ. بهادر، آ.(1399). کاربرد داده کاوی با استفاده از الگوریتم یادگیری ماشین برای بررسی تاثیر ویژگی های خودرو در پیش بینی ریسک خسارت مالی در رشته بیمه شخص ثالث. فصلنامه علمی پژوهشی - سال سی و پنجم، بهار 1399، 65-33.

[2] رستمخانی، ح.، خدارحمی، ب.، جهانشاد، آ.(1400). انتخاب بهینه سهام با استفاده از الگوریتم خفاش و جنگل تصادفی. فصلنامه مهندسی مالی و مدیریت اوراق بهادار، ، 461-470.

[3] گل ارضی، غ.، یعقوبی، م. (1398). مقایسه دقت مدل اولسن و مدل پیوتروسکی در تبیین تغییرات قیمت سهام پذیرفته شده در بازار اوراق بهادار تهران. فصلنامه مدلسازی اقتصادسنجی - سال سوم، تابستان 1397، 163-147.

[4] Kara, Y., Boyacioglu, M. A., and 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 systems with Applications*, *38*(5), 5311-5319.

[5] Wilder, J. W. (1978). *New concepts in technical trading systems*. Trend Research.