The Comparative Assessment of Data Mining Methods Effectiveness to Forecasting Return and Risk of Stock in Companies Listed in Tehran Stock Exchange

*** A. Soroushyar**
Department of Accounting, Isfahan (Khorasgan) Branch, Islamic Azad University, Isfahan, Iran
M. Akhlaghi
Department of Accounting, Najaf Abad Branch,Islamic Azad University, Isfahan, Iran

Abstract

Risk and stock returns has always been the most important factors in making financial decisions for investors. Hence, forecasting of these factors is very important for investors and other capital market participants. The purpose of this study is to use data mining techniques to predict stock returns and systematic risk. In this study, with using of four algorithms: linear discriminant analysis algorithm, quadratic discriminant analysis algorithm, K-nearest neighbors algorithm and decision tree with the help of 16 independent variables has been addressing to predict stock returns and systematic risk. Four algorithms are running once with using of whole independent variables and one again with using of 4 independent variables that are known with using of filtering approach as the most effectiveness of variables in predicting the return and risk. Then the accuracy of forecasting 4 algorithms in both cases (in total, 8 predictions for return and 8 predictions for risk) compares and chooses the best algorithm. For this purpose data of 107 companies listed in Tehran stock exchange is used during the period of 2002 to 2014. The results show that in the case of using 16 independent variables, the linear discriminant analysis algorithm provides the best prediction for return and the quadratic discriminant analysis algorithm provides the best prediction for systematic risk. But in the case of using independent variables that are chosen, the quadratic discriminant analysis algorithm offers the prediction for return and linear discriminant analysis algorithm offers the best prediction for systematic risk. In general, using of selected independent variables (instead of using whole independent variables), improves the algorithm's ability in prediction of return and systematic risk.

Keywords: Return, Systematic Risk, Data Mining, Linear Discriminant Analysis, Quadratic Discriminant Analysis, K-Nearest Neighbors, Decision Tree.

* a_soroushyar@yahoo.com

پژوهش‌های حسابداری مالی
سال نهم، شماره اول، پیاپی (۳۱)، بهار ۱۳۹٦
تاریخ وصول: ۱۳۹٤/۱۲/۲٦
تاریخ پذیرش: ۱۳۹٦/٤/۱۱
صص: ۷٦-۵۷

ارزیابی مقایسه‌ای اثربخشی تکنیک‌های داده‌کاوی در پیش‌بینی ریسک و بازده سهام شرکت‌های پذیرفته شده در بورس اوراق بهادار تهران

افسانه سروش یار [*] ، محمد اخلاقی [**]

[*] استادیار حسابداری، دانشگاه آزاد اسلامی، واحد اصفهان (خوراسگان)، اصفهان، ایران
a_soroushyar@yahoo.com

[**] دانشجوی کارشناسی ارشد حسابداری، دانشگاه آزاد اسلامی، واحد نجف آباد، اصفهان، ایران
mohamad.a007@yahoo.com

چکیده

ریسک و بازده سهام همواره از مهم‌ترین عوامل در اتخاذ تصمیمات مالی سرمایه‌گذاران بوده است. از این رو پیش‌بینی آنها برای سرمایه‌گذاران و سایر فعالان بازار سرمایه حائز اهمیت بسیار است. هدف پژوهش حاضر به کارگیری تکنیک‌های داده‌کاوی در پیش‌بینی بازده و ریسک سیستماتیک سهام در شرکت‌های پذیرفته شده در بورس اوراق بهادار تهران می‌باشد. در این پژوهش با استفاده از چهار الگوریتم تحلیل جداساز خطی، الگوریتم تحلیل جداساز غیرخطی، الگوریتم نزدیکترین K همسایگی و درخت تصمیم و به کمک ۱٦ متغیر مستقل به پیش‌بینی بازده و ریسک سیستماتیک سهام پرداخته می‌شود. چهار الگوریتم مذکور یک بار با استفاده از کل متغیرهای مستقل و بار دیگر با استفاده از ٤ متغیر مستقل که با استفاده از رویکرد فیلترینگ به عنوان مؤثرترین متغیرها در پیش‌بینی بازده و ریسک شناخته شده‌اند، اجرا می شود. سپس صحت پیش‌بینی چهار الگوریتم در دو حالت (مجموعاً ۸ پیش‌بینی برای بازده و ۸ پیش‌بینی برای ریسک) مقایسه و بهترین الگوریتم انتخاب می‌گردد. بدین منظور داده‌های ۱۰۷ شرکت پذیرفته شده در بورس اوراق بهادار تهران طی سال‌های ۱۳۸۰ تا ۱۳۹۲ مورد استفاده قرار گرفته است. نتایج حاصل شده حاکی از این است که در حالت به کارگیری

۱ـ نشانی مکاتباتی نویسنده مسؤول: اصفهان، خوراسگان، دانشگاه آزاد اسلامی واحد اصفهان (خوراسگان)، دانشکده علوم انسانی و حقوق، گروه حسابداری

۱۶ متغیر مستقل الگوریتم تحلیل جداساز خطی بهترین پیش‌بینی بازده و الگوریتم تحلیل جداساز غیرخطی بهترین پیش‌بینی ریسک سیستماتیک را به دست می‌دهد. لیکن در حالت استفاده از متغیرهای مستقل منتخب الگوریتم تحلیل جداساز غیرخطی بهترین پیش‌بینی بازده و الگوریتم تحلیل جداساز خطی بهترین پیش‌بینی ریسک سیستماتیک را ارائه می‌دهد. به طور کلی استفاده از متغیرهای مستقل منتخب (به جای استفاده از کل متغیرهای مستقل) توان الگوریتم‌ها در پیش‌بینی بازده و ریسک سیستماتیک را بهبود می‌بخشد.

واژه‌های کلیدی: بازده، ریسک سیستماتیک، داده‌کاوی، تحلیل جداساز خطی، تحلیل جداساز غیرخطی، نزدیکترین K همسایگی، درخت تصمیم طبقه‌بندی کننده.

مقدمه

یکی از مهم‌ترین مسائلی که در اقتصاد هر کشوری بسیار حایز اهمیت است، بازار سرمایهٔ آن کشور است. بازار سرمایه، بازاری است که هر سرمایه‌گذار نهادی و یا فردی برای اتخاذ تصمیمات اقتصادی و در نهایت سرمایه‌گذاری در پرتفوی مناسب خود، با آن روبه‌روست. ریسک و بازده، دو مؤلفهٔ اساسی است که در تصمیم‌گیری‌های مالی سرمایه‌گذاران تأثیر بسزایی دارد؛ چراکه سرمایه‌گذاران همواره درصدد کسب بازدهی بیشتر و تحمل ریسک کمتر هستند. بازده سهام شامل سود نقدی و تغییرات قیمت سهام و ریسک شاخصی برای اندازه‌گیری بی‌اطمینانی در حصول بازده مورد انتظار است. یکی از مباحث مهم در بازار سرمایه، آگاهی از میزان ریسک شرکت‌ها، به‌ویژه ریسک سیستماتیک است که نقش بسزایی در تصمیم‌گیری‌ها ایفا می‌کند؛ زیرا اعتقاد بر این است که بازده مورد انتظار سهام شرکت‌ها تابعی از ریسک سیستماتیک است و ریسک سیستماتیک، تغییرات نرخ بازده یک سهم نسبت به نرخ بازده کل بازار سهام را بیان می‌کند. به‌نظر می‌رسد اطلاعات حسابداری در قیمت بازار سهام و ریسک بازار شرکت‌ها مؤثرند. از دیرباز تأثیر اطلاعات حسابداری بر ریسک و بازده، توجه پژوهشگران را به‌دنبال داشته است. برای مثال، بیور و همکاران [۱۲] و الگرز و ماری [۱۸] رابطه بین اطلاعات حسابداری و ریسک سیستماتیک را بررسی و وجود این رابطه را تأیید کرده‌اند. همچنین، رابطهٔ سود با بازده سهام که نخستین‌بار بال و براون [۹] مطرح کردند، به‌دلیل گستردگی و سهولت استفاده از اطلاعات حسابداری، ملاکی برای ارزیابی عملکرد مالی حال و آیندهٔ شرکت استفاده شده است.

پژوهش حاضر، روشی نظام‌مند را برای جست‌وجوی ویژگی‌های بالقوهٔ اثرگذار بر پیش‌بینی ریسک و بازده سهام در بورس از بین نسبت‌های نقدینگی، نسبت‌های فعالیت، نسبت‌های اهرمی، نسبت‌های سودآوری و نسبت‌های بازار به‌کار می‌گیرد. همچنین، در این پژوهش از چهار الگوریتم تحلیل جداساز خطی[1] (LDA)، الگوریتم تحلیل جداساز غیرخطی[2] (QDA)، الگوریتم نزدیک‌ترین همسایگی[3] (KNN) و الگوریتم درخت تصمیم طبقه‌بندی‌کننده[4] (CDT) برای پیش‌بینی ریسک و بازده سهام استفاده شده است و دقت پیش‌بینی این روش‌ها با یکدیگر مقایسه می‌شود. به‌کمک الگوریتم‌های گزینش[5] متغیرهای مستقل و تحلیل خوشه‌ای[6]، مربوط‌ترین متغیرها از بین نسبت‌های

[1] Linear Discriminant Analysis
[2] Quadratic Discriminant Analysis
[3] K-Nearest Neighbors
[4] Classification Decision Tree
[5] Filter
[6] Function Based Clustering

مورد بررسی انتخاب می‌شود و دوباره به مقایسهٔ دقت پیش‌بینی چهار روش فوق با استفاده از نسبت‌های منتخب پرداخته می‌شود. بر این اساس، هدف پژوهش حاضر تعیین بهترین الگوریتم برای پیش‌بینی ریسک و بازده سهام شرکت‌هاست. استفاده از چهار الگوریتم متفاوت و نیز استفاده از الگوریتم‌های گزینش متغیرهای مستقل و تحلیل خوشه‌ای برای تعیین تأثیرگذارترین عوامل بر ریسک و بازده سهام از نوآوری‌های این پژوهش است. مقایسهٔ دقت پیش‌بینی چهار الگوریتم فوق و انتخاب الگوریتم بهینه، سرمایه‌گذاران و تحلیل‌گران مالی را در پیش‌بینی ریسک و بازده شرکت یاری خواهد کرد.

مبانی نظری و پیشینهٔ پژوهش

از مهم‌ترین دغدغه‌های متخصصان بازار، اطلاعاتی است که شرکت‌ها ارائه می‌دهند. پیش‌بینی قابل اتکای وضعیت شرکت، فرصتی را در اختیار سرمایه‌گذار قرار می‌دهد تا سرمایه‌گذاری مطمئن‌تری انجام دهند و بازده بیشتری را عاید شوند [۲۰]. برخلاف بازده که اغلب مورد توجه فعالان بازار قرار دارد، به پیش‌بینی ریسک کمتر توجه شده است. این در حالی است که آنان معمولاً باید بازده خود را با سطح متناسبی از ریسک تنظیم کنند، زیرا بدون ارزیابی ریسک، نتایج و یافته‌های کارآمد در زمینهٔ مجموعه اوراق بهادار معنا و مفهومی ندارد. همچنین، پیش‌بینی تغییرات قیمت سهام به‌صورت نادرست، دیدگاه دقیقی از آیندهٔ سهام و جذب سرمایه‌گذاران در اختیار قرار نمی‌دهد. از این‌رو، ریسک و بازده هر دو از مهم‌ترین عوامل برای اتخاذ تصمیمات مالی محسوب می‌شوند [۱۰].

به‌منظور پیش‌بینی دقیق ریسک و بازده، تعیین عوامل اثربخش حایز اهمیت است. اگرچه عوامل متعددی در تعیین ریسک و بازده سهام مؤثر است، اما نادیده‌گرفتن اقلام صورت‌های مالی در این پیش‌بینی دور از هدف اصلی گزارشگری مالی است. صورت‌های مالی اطلاعاتی دربارهٔ عملیات گذشتهٔ شرکت ارائه می‌کنند و سرمایه‌گذاران از این اطلاعات برای پیش‌بینی بازدهی آتی شرکت در تصمیمات تخصیص منابع بهره می‌گیرند [۱۵]. در صورتی‌که ارقام حسابداری به‌میزان کافی قابل اتکا باشد، منعکس‌کنندهٔ ارزش بازار حقوق صاحبان سهام است و اطلاعاتی را برای ارزشیابی شرکت به بازار منتقل می‌کند. بینش زیربنایی در تبیین این دیدگاه چنین است که توابع حسابداری، اطلاعاتی فراهم می‌آورد که منعکس‌کنندهٔ عملکرد شرکت است و متعاقباً در قیمت سهام شرکت منعکس می‌شود [۱۳].

به اعتقاد بارث و همکاران [۱۱] حذف صورت سود و زیان در تحلیلگری، احتمالاً منجر به تبیین اشتباه مدل می‌شود و تفسیر نتایج به‌دست‌آمده دشوار خواهد شد. مفیدبودن صورت سود و زیان و ترازنامه به‌میزان زیادی به توانایی آنها در پیش‌بینی جریان‌های نقد آتی بستگی دارد [۱۵]. چنین انتظار می‌رود که انعطاف‌پذیری مالی و ارزش نقدشوندگی دارایی‌ها از طریق صورت ترازنامه و جریان‌های نقدی مورد انتظار آتی از طریق صورت سود و زیان منعکس شود. افزون بر این، هرچه همبستگی متغیرهای حسابداری و ریسک سیستماتیک بیشتر باشد، قیمت اوراق بهادار و در نتیجه ریسک آن در بازار نسبت به اطلاعات جدید سریع‌تر واکنش نشان می‌دهد [۷]. واتز و زیمرمن [۲۷] معتقدند ارقام حسابداری، اطلاعاتی را دربارهٔ جریان‌های نقد مورد انتظار و نرخ

تنزیل به بازار مخابره می‌کند؛ زیرا نرخ بازده مورد انتظار به ریسک دارایی و ریسک دارایی نیز به اعداد حسابداری وابسته است. از آنجا که سود حسابداری با جریان‌های نقدی جاری و آتی شرکت مرتبط است، سود حسابداری حاوی اطلاعاتی در رابطه با ریسک شرکت است. بر این اساس، چنین استنباط می‌شود که نسبت‌های حسابداری به‌عنوان شاخص ریسک استفاده می‌شوند [۲].

محاسبهٔ نسبت‌های مالی که بیانگر ساختارهای زیربنایی همچون سودآوری، نقدینگی، کارایی و اهرمی است، به‌منظور درک رابطهٔ صورت‌های مالی و ریسک ضروری است. افزون بر این، استفاده‌کننده به هنگام تصمیم‌گیری باید چگونگی وزن‌دهی به این اطلاعات را دریابد. در این راستا، پژوهش حاضر درصدد به‌کارگیری الگوریتم‌های داده‌کاوی برای انجام وزن‌دهی به متغیرهای حسابداری مختلف برای پیش‌بینی دقیق‌تر بازده و ریسک سهام شرکت است. داده‌کاوی به فرآیند جست‌وجو و کشف مدل‌های گوناگون، مختصرسازی و اخذ مقادیر از مجموعه‌ای از داده‌های معلوم گفته می‌شود. داده‌کاوی سودمندترین سناریوی تحلیلی اکتشافی است که در آن تصور و برداشت از پیش تعیین‌شده‌ای دربارهٔ نتیجه‌ای که به‌دست می‌آید، وجود ندارد. در حقیقت، داده‌کاوی جست‌وجوی لازم برای یافتن اطلاعات کلی جدید، ارزشمند و غیربدیهی از میان حجم زیاد داده‌هاست [٦]. در ادامه برخی از پژوهش‌های انجام‌شده در خصوص پیش‌بینی بازده و ریسک سهام ارائه می‌شود.

ردر و همکاران [۲۵] با انجام پژوهشی به بررسی توان پیش‌بینی بازده سهام در مدل غیرخطی شبکهٔ عصبی بازگشت‌کننده و دو مدل خطی شامل مدل میانگین متحرک خودکاهنده و مدل هموارسازی تصاعدی پرداختند. نتایج حاصل‌شده، درستی پیش‌بینی عملکرد شبکهٔ عصبی بازگشت‌کننده را تأیید کرد. همچنین عملکرد مدل ترکیبی پیشنهادی در پیش‌بینی بازده سهام به‌طور قابل‌توجهی بهبود یافت.

ژانگ و همکاران [۲۸] در پژوهشی با استفاده از داده‌های ۱۳ساله از بازار سهام شانگهای، توان الگوریتم انتخاب ویژگی علّی (CFS) و سه الگوریتم انتخاب ویژگی شناخته‌شده، یعنی تجزیه و تحلیل محتوای اصلی (PCA)، درخت تصمیم (CART) و حداقل انقباض خالص و عملگر انتخاب (LASSO) را در پیش‌بینی بازده سهام مقایسه کردند. نتایج نشان داد CFS در شرایطی که با هریک از هفت مدل خطی پایه و شناسایی ۱۸ ویژگی سازگار مهم ترکیب شود، بهترین عملکرد پیش‌بینی را از نظر صحت و دقت خواهد داشت.

چنگ‌لی و همکاران [١٤] طی انجام پژوهشی به پیش‌بینی ریسک و بازده سرمایه‌گذاری در سهام از طریق شبیه‌ساز عددی، یعنی زمان تأخیر و تابع چگالی احتمال بازده سهام در مدل اصلاح‌شدهٔ هستون[1] با تأخیر زمانی پرداختند. آنها تأخیر زمانی و موقعیت اولیهٔ ریسک و بازده سرمایه‌گذاری را تجزیه و تحلیل کردند و دریافتند که یک تأخیر زمانی بهینه مطابق با حداقل ریسک سرمایه‌گذاری سهام، حداکثر متوسط بازده قیمت سهام و ثبات قوی از بازده سهام برای کشش قوی تقاضای سهام (EDS) وجود دارد.

جایاواردنا و همکاران [۲۲] با انجام پژوهشی به پیش‌بینی نوسانات سهام با استفاده از اطلاعات یک ساعت بعد، با روش مربع بازگشت شبانه پرداختند. آنها سودمندی استفاده از نوسانات قبل از بازشدن

[1] Heston

بازار و نوسانات شناسایی‌شده از دارایی‌های مرتبط از بازارهای دیگر را زمانی‌که بورس ایتالیا بسته است، تأیید کردند. آنها دریافتند قدرت پیش‌بینی اطلاعات شبانه در دورهٔ زمانی بازشدن بازار بالاتر است و در نهایت این مدل ابزار مهمی را برای سرمایه‌گذار فراهم می‌کند.

اوزتکین و همکاران [٢٤] با استفاده از سه روش تطبیق سیستم استنتاج فازی- عصبی، شبکه‌های عصبی و پشتیبانی ماشین‌بردار به پیش‌بینی بازده روزانهٔ سهام پرداختند. آنها دریافتند روش ماشین‌بردار پیش‌بینی‌های دقیق‌تری را نسبت به دو روش دیگر به دست می‌دهد.

تسای و هسیائو [٢٦] در پژوهشی از سه روش تجزیه و تحلیل محتوای بنیادی (PCA)، الگوریتم‌های ژنتیک (GA) و درخت تصمیم (CART) با استفاده از روش فیلترکردن متغیرهای نماینده بر مبنای راهبردهای واحد، متقاطع و چندتقاطعی برای پیش‌بینی بازده سهام استفاده کردند. نتیجهٔ حاصل از به‌کارگیری این روش‌ها با استفاده از دو شیوهٔ متقاطع و چندتقاطعی به‌ترتیب به انتخاب ١٤ و ١٧ شاخص مهم برای پیش‌بینی بازده سهام منتهی شد که می‌توانند برای تصمیم‌گیری سرمایه‌گذاران در آینده استفاده شوند.

رهنمای رودپشتی و همکاران [٥] کارایی بهینه‌سازی پرتفوی سهام را براساس مدل پایدار با بهینه‌سازی کلاسیک، برای پیش‌بینی ریسک و بازده پرتفوی مقایسه کردند. این پژوهش تلاشی است به‌منظور بهینه‌سازی پرتفوی با استفاده از بهینه‌سازی پایدار و تخمین بازده و ریسک پرتفوی و مقایسهٔ بازده و ریسک پیش‌بینی‌شدهٔ مدل کلاسیک با ریسک و بازده پیش‌بینی‌شدهٔ این مدل. در این پژوهش مشخص شد بازده پیش‌بینی‌شده پرتفوی در مدل پایدار با بازده پیش‌بینی‌شده در مدل کلاسیک تفاوت معناداری دارد و ریسک پیش‌بینی‌شده در مدل پایدار تفاوت معناداری با ریسک پیش‌بینی‌شده در مدل کلاسیک ندارد.

نیکو اقبال و همکاران [٨] به ارزیابی دقت عملکرد مدل‌های شبکهٔ عصبی ایستا و پویا در پیش‌بینی بازدهی شاخص قیمت و بازده نقدی بورس تهران پرداختند تا بتوانند بهترین مدل را برای پیش‌بینی بازدهی شاخص قیمت انتخاب کنند. در این پژوهش از مدل‌های شبکهٔ عصبی اتورگرسیون پویا، ایستای فازی و ایستای چندلایه پیش‌خور استفاده شده است که طبق نتایج به‌دست‌آمده مدل شبکه عصبی فازی عملکرد بهتری در پیش‌بینی متغیرهای مورد بررسی داشته است.

ایزدی‌نیا و کربلایی‌کریم [٣] به بررسی نقش برخی متغیرهای حسابداری از جمله جریان نقدی آزاد، بازده نقدی سرمایه‌گذاری ارزش افزوده اقتصادی و سود هر سهم در پیش‌بینی بازده سهام پرداختند. آنان دریافتند از بین متغیرهای یادشده تنها سود هر سهم ارتباط معناداری با بازده سهام دارد.

احمدپور و غلامی‌جمکرانی [١] به بررسی برخی از نسبت‌های مالی از جمله نسبت دارایی جاری به بدهی جاری، سود خالص به حقوق صاحبان سهام، فروش به حقوق صاحبان سهام، بدهی به حقوق صاحبان سهام و جمع دارایی‌ها با ریسک سیستماتیک پرداختند. در این پژوهش شواهدی دالّ بر رابطهٔ معنادار بین اطلاعات حسابداری با ریسک سیستماتیک یافت نشد.

نمازی و خواجوی [٧] به بررسی نقش متغیرهای حسابداری در پیش‌بینی ریسک سیستماتیک

شرکت‌های پذیرفته‌شده در بورس اوراق بهادار تهران پرداختند. آنان از ۱۷ متغیر حسابداری در پنج‌دسته نسبت‌های نقدینگی، اهرمی، سودآوری، اهرم عملیاتی، اهرم مالی به‌عنوان متغیر مستقل استفاده کردند. در این پژوهش از رگرسیون چندمتغیره و از روش گزینش دنباله‌ای متغیرها با عنوان حذف پس‌رو به‌منظور انتخاب مدل بهینه استفاده شده است. نتایج به‌دست‌آمده از این پژوهش نشان داد از بین متغیرهای مستقل ۱۲ متغیر با ریسک سیستماتیک رابطه‌ای معنادار دارند.

خواجوی و همکاران [٤] به آزمون مدل بازده و مدل قیمت با استفاده از الگوی پانل با داده‌های متوازن پرداختند و به این نتیجه رسیدند که چون محتوای اطلاعاتی سود هر سهم نسبت به ارزش دفتری آن بیشتر است و با توجه به نتایج مدل که نشان می‌دهد محتوای اطلاعاتی نسبت تغییرات سود هر سهم به قیمت، بیشتر و در مقابل آن محتوای اطلاعاتی نسبت سود هر سهم به قیمت کمتر است، پس اطلاعات حسابداری در تعیین قیمت سهام و نرخ بازده و همچنین در تصمیم‌گیری‌های سرمایه‌گذاران بسیار با اهمیت است.

پرسش‌های پژوهش

از آنجا که هدف پژوهش حاضر مقایسهٔ توان چهار تکنیک داده‌کاوی در پیش‌بینی بازده و ریسک سیستماتیک است، پرسش‌های پژوهش را به این صورت ارائه می‌شود:

– کدام‌یک از چهار الگوریتم تحلیل جداساز خطی (LDA)، الگوریتم تحلیل جداساز غیرخطی (QDA)، الگوریتم نزدیکترین K همسایگی (KNN) و الگوریتم درخت تصمیم طبقه‌بندی‌کننده قادر است بازده سهام شرکت‌های پذیرفته‌شده در بورس اوراق بهادار تهران را با دقت بیشتری پیش‌بینی کند؟

– کدام‌یک از چهار الگوریتم تحلیل جداساز خطی (LDA)، الگوریتم تحلیل جداساز غیرخطی (QDA)، الگوریتم نزدیک‌ترین K همسایگی (KNN) و الگوریتم درخت تصمیم طبقه‌بندی‌کننده قادر است ریسک سیستماتیک شرکت‌های پذیرفته‌شده در بورس اوراق بهادار تهران را با دقت بیشتری پیش‌بینی کند؟

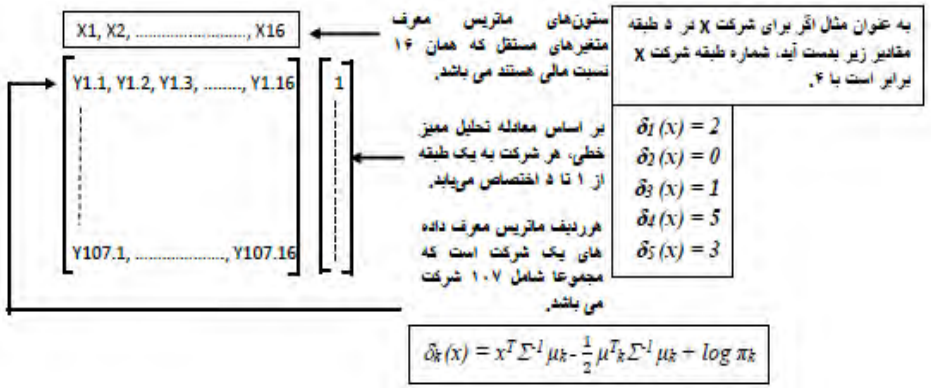
روش پژوهش

پژوهش حاضر از این‌رو که در پی یافتن بهترین تکنیک برای پیش‌بینی ریسک و بازده است، در زمره پژوهش‌های همبستگی و پیش‌بینی قرار دارد که تحلیل‌های آن مبتنی بر روش‌های اکتشافی است. از سویی دیگر، این پژوهش از نوع کاربردی است. در این پژوهش به‌منظور پیش‌بینی ریسک سیستماتیک و بازده سهام شرکت به کمک اطلاعات حسابداری و نسبت‌های مالی از روش داده‌کاوی استفاده شده است. استفاده از این تکنیک مستلزم اجرای سه مرحله است. در اولین مرحله، فهرست کاملی از نسبت‌های مالی و متغیرهای حسابداری تهیه می‌شود که قرار است به‌کمک آنها به پیش‌بینی ریسک و بازده پرداخته شود. این متغیرها شامل نسبت‌های نقدینگی، نسبت-های فعالیت، نسبت‌های اهرمی، نسبت‌های سودآوری و نسبت‌های بازار هستند. همچنین دو متغیر پاسخ (وابسته) شامل ریسک سیستماتیک و بازده سهام هر شرکت محاسبه می‌شود. بازده سهام شامل دو بخش سود نقدی و تغییرات قیمت بازار سهام است و ریسک سیستماتیک نیز از تقسیم کورایانس بازده سهام و بازده بازار بر واریانس بازده بازار محاسبه

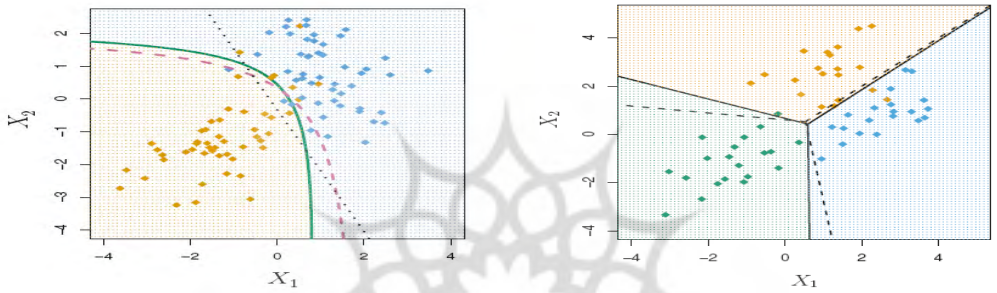
می‌شود. سپس داده‌های مربوط به ریسک و بازده پس از حذف داده‌های پرت به‌طور مجزا در ۵ طبقه در قالب طبقات خیلی کم، کم، متوسط، زیاد و خیلی زیاد دسته‌بندی شده است. طبقه‌بندی بازده به این شرح است: خیلی کم (۰/۲۷۶- تا ۰/۹۳۱-)، کم (۰/۰۴۶- تا ۰/۲۷۴-)، متوسط (۰/۱۵۸ تا ۰/۰۴۴-)، زیاد (۰/۶۰۵ تا ۰/۱۵۸) و خیلی زیاد (۸/۱۰۳ تا ۰/۶۰۶). همچنین طبقه‌بندی ریسک سیتماتیک نیز به این صورت است: خیلی کم (۰/۱۹۰- تا ۴/۵۹-)، کم (۰/۱۲۰- تا ۰/۱۸۰-)، متوسط (۰/۶۱ تا ۰/۱۲۲)، زیاد (۱/۲۷۶ تا ۰/۶۱۷) و خیلی زیاد (۱۹/۲۲ تا ۱/۲۸). پس از انجام طبقه‌بندی به‌کمک چهار الگوریتم مذکور به پیش‌بینی ریسک و بازده پرداخته می‌شود. در پژوهش حاضر این فرآیند دو بار تکرار می‌شود. یک‌بار به‌کمک ۱۶ نسبت مالی (به شرح نگارهٔ ۱) به پیش‌بینی ریسک سیستماتیک و بازده با استفاده از چهار الگوریتم پرداخته می‌شود. بار دیگر، ابتدا با استفاده از تکنیک‌های خوشه‌بندی و فیلترینگ تأثیرگذارترین نسبت‌های مالی بر ریسک سیستماتیک و بازده تعیین و دوباره با استفاده از متغیرهای گزینش‌شده ریسک و بازده به‌کمک این الگوریتم‌ها پیش‌بینی می‌شود. در پایان نیز به مقایسه بین الگوریتم‌های مختلف پرداخته و بهترین الگوریتم بر اساس دقت پیش‌بینی ریسک و بازده انتخاب خواهد شد [۱۰].

از جمع ۱۱۹۳ ردیف داده جمع‌آوری‌شده، ابتدا ۷۵ درصد از داده‌ها (تقریباً تعداد ۹۰۰ ردیف داده) با به‌کارگیری الگوریتم‌های مربوطه آموزش داده می‌شوند و سپس با استفاده از ۲۵ درصد از داده‌های باقیمانده (که تقریباً تعداد ۲۹۳ ردیف داده است) و با عنوان داده‌های آزمون شناخته می‌شوند، میزان دقت پیش‌بینی الگوریتم‌ها آزمون می‌شود. به‌عبارت‌دیگر، ۰/۷۵ از داده‌ها موسوم به داده‌های آموزش برای تبیین ارتباط متغیرهای مستقل و وابسته و ۰/۲۵ مابقی موسوم به داده‌های آزمون برای آزمایش رابطه ایجادشده به‌کار می‌رود. برای اجرای یک پیش‌بینی معتبر و مقاوم (معتبرسازی) از مدل k-fold cross-validation استفاده شده است [۲۱]؛ بنابراین در مرحلهٔ اول هر الگوریتم با استفاده از ۹۰۰ داده، نوع رابطه بین ۱۶ متغیر مستقل و بازده (ریسک) را به نرم‌افزار آموزش می‌دهد، سپس الگویی را که آموخته است، برای پیش‌بینی بازده (ریسک) ۲۹۳ ردیف داده باقیمانده به‌کار می‌گیرد. بازده (ریسک) پیش‌بینی‌شده در قالب یکی از طبقات خیلی کم، کم، متوسط، زیاد و خیلی زیاد به‌کمک نرم‌افزار ارائه می‌شود. سپس طبقهٔ پیش‌بینی‌شده با طبقه‌ای مقایسه می‌شود که بازده (ریسک) واقعی در آن قرار دارد. در صورتی‌که طبقهٔ پیش‌بینی‌شده با طبقهٔ واقعی بازده (ریسک) مشابه باشد، این الگوریتم پیش‌بینی را به‌درستی انجام داده است. این فرآیند برای هر چهار الگوریتم خبره در پیش‌بینی طبقات استفاده شده است.

در الگوریتم تحلیل جداساز خطی براساس معادلهٔ تحلیل ممیز خطی، برای هر ردیف شرکت در هر طبقه از یک تا پنج، یک بردار ورودی $\delta_k(x)$ محاسبه می‌شود که در نهایت شمارهٔ طبقهٔ بیشترین مقدار بردار ورودی از بین پنج بردار محاسبه‌شده، طبقهٔ آن شرکت در نظر گرفته می‌شود. الگوریتم تحلیل جداساز غیرخطی کاملاً مطابق با روش قبل است با این تفاوت که در این روش، ماتریس واریانس کواریانس برای هر طبقه به‌صورت جداگانه محاسبه می‌شود؛ یعنی به‌جای یک ماتریس واریانس کواریانس کلی، پنج ماتریس واریانس کواریانس وجود دارد.



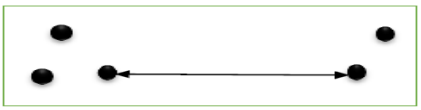
شکل ۱- تحلیل جداساز خطی و ماتریس واریانس کواریانس



تحلیل جداساز خطی (برای یک مسأله سه‌کلاسه) **تحلیل جداساز غیرخطی (برای یک مسأله دوکلاسه)**

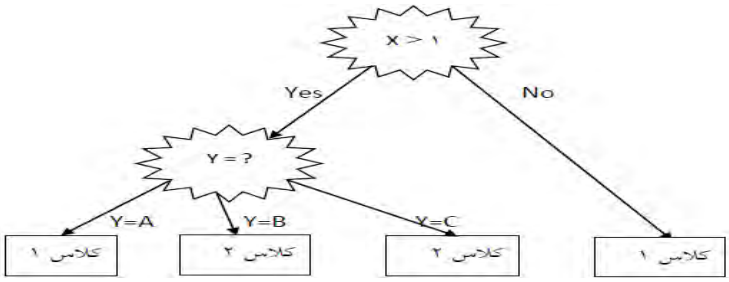
شکل ۲- نمودار تحلیل جداساز خطی و غیرخطی

در الگوریتم نزدیکترین K همسایگی، تعداد K نقطه شناسایی می‌شود که از لحاظ معیار شباهت در متغیرهای مستقل، به نقطهٔ مورد ارزیابی، نزدیک‌ترین نقاط هستند. پس از آن احتمال اینکه نقطهٔ مورد نظر در هر طبقه پیش‌بینی شود، بر اساس نسبت تعداد نزدیک‌ترین نقاط در هر طبقه به تعداد کل نقاط همسایگی محاسبه می‌شود و در نتیجه، نقطهٔ مورد نظر در طبقه‌ای دسته‌بندی می‌شود که دارای بیشترین مقدار عددی احتمال باشد.



شکل ۳- فاصلهٔ نزدیک‌ترین همسایگی

درخت تصمیم، شامل گره‌هایی است که بر روی آنها آزمایشاتی صورت گرفته است. شاخه‌های بیرونی یک گره از نتایج آزمون‌های انجام‌گرفته در هر گره منتج شده است. یک درخت تصمیم برای طبقه-بندی نمونه‌ها با دو مشخصهٔ ورودی X و Y در شکل ٤ نمایش داده شده است. نمونه‌های با مقادیر ویژگی X >۱ و Y = B در طبقهٔ دوم جای می‌گیرند. در حالی‌که نمونه‌هایی با مقادیر X < ۱ در طبقهٔ اول دسته‌بندی می‌شوند (با هر مقدار Y)[6] . لازم به ذکر است که در این پژوهش از شاخص جینی و معیار واریانس برای انشعاب در درخت تصمیم استفاده شده است و عمق درخت تصمیم ده انشعاب بوده است.



شکل ٤- درخت تصمیم همراه با آزمون بر روی صفات X و Y

نگارهٔ ۱- متغیرهای مستقل و وابسته

متغیرهای پاسخ (وابسته)	
= بازده سهام $\frac{\text{سود تقسیمی} + \text{قیمت بازار سهام در ابتدای سال} - \text{قیمت بازار سهام در پایان سال}}{\text{بدهی جاری}}$	
= ریسک سیستماتیک $\frac{cov\,(\text{بازده بازار، بازده سهام})}{var\,(\text{بازده بازار})}$	
متغیرهای مستقل	
نسبت‌های سودآوری:	نسبت‌های نقدینگی:
= بازده فروش $\frac{\text{سود خالص}}{\text{فروش خالص}}$	= نسبت جاری $\frac{\text{دارایی جاری}}{\text{بدهی جاری}}$
= بازده دارایی‌ها $\frac{\text{سود خالص}}{\text{جمع کل دارایی‌ها}}$	= نسبت آنی $\frac{\text{موجودی مواد و کالا} - \text{دارایی جاری}}{\text{بدهی جاری}}$
= بازده حقوق صاحبان سهام $\frac{\text{سود خالص}}{\text{حقوق صاحبان سهام}}$	بدهی‌های جاری ° دارایی‌های جاری = خالص سرمایه در گردش
نسبت‌های اهرمی:	نسبت‌های فعالیت:
= نسبت بدهی به دارایی $\frac{\text{جمع کل بدهی‌ها}}{\text{جمع کل دارایی‌ها}}$	= نسبت گردش موجودی کالا $\frac{\text{بهای تمام‌شده کالای فروش‌رفته}}{\text{متوسط موجودی کالا}}$
= نسبت بدهی به حقوق صاحبان سهام $\frac{\text{جمع کل بدهی‌ها}}{\text{حقوق صاحبان سهام}}$	= نسبت گردش دارایی‌ها $\frac{\text{درآمد خالص فروش}}{\text{جمع کل دارایی‌ها}}$
نسبت‌های بازار:	
= نسبت سود هر سهم $\frac{\text{سود متعلق به سهامداران عادی}}{\text{میانگین موزون تعداد سهام عادی}}$	= نسبت قیمت به درآمد هر سهم $\frac{\text{قیمت بازار هر سهم}}{\text{سود هر سهم}}$
(سود هر سهم) Δ = تغییرات سود هر سهم	= نسبت قیمت به فروش $\frac{\text{قیمت بازار هر سهم}}{\text{فروش هر سهم}}$
توبین Q = نسبت $\frac{\text{ارزش بازار}}{\text{ارزش جایگزینی یا ارزش دفتری دارایی‌های شرکت}}$	= نسبت پرداخت سود سهام $\frac{\text{سود تقسیمی هر سهم}}{\text{سود هر سهم}}$

جامعه و نمونهٔ آماری

جامعهٔ آماری این پژوهش شامل کلیهٔ شرکت‌های پذیرفته‌شده در بورس اوراق بهادار تهران است. در

این مطالعه برای اینکه نمونهٔ پژوهش نمایندهٔ مناسبی از جامعهٔ آماری مورد نظر باشد، برای انتخاب نمونه از روش حذف سیستماتیک استفاده شده است. برای این‌منظور، این معیارها در نظر گرفته شده است و در صورتی‌که شرکتی همهٔ معیارها را احراز کرده باشد، یکی از شرکت‌های نمونهٔ انتخاب شده است:

۱) قبل از سال ۱۳۸۱ در بورس پذیرفته شده باشد، ۲) سال مالی آنها منتهی به پایان اسفندماه باشد، ۳) در قلمرو زمانی پژوهش تغییر سال مالی نداشته باشد، ٤) جزء شرکت‌های سرمایه‌گذاری و واسطه‌گری مالی نباشد، ٥) اطلاعات مورد نیاز شرکت در دورهٔ مورد بررسی در دسترس باشد.

نمونهٔ آماری پژوهش حاضر شامل ۱۰۷ شرکت است. از این تعداد، داده‌های ۱۱۹۳ سال- شرکت به‌طور کامل در اختیار بود که به‌عنوان نمونه استفاده شده است.

پیش‌بینی متغیرهای پاسخ با استفاده از همهٔ متغیرهای مستقل

در این بخش نتایج مربوط به پیش‌بینی بازده و ریسک با استفاده از ١٦ متغیر مستقل مورد اشاره، به کمک چهار الگوریتم به تفکیک ارائه می‌شود.

یافته‌های الگوریتم تحلیل جداساز خطی

نتایج حاصل از به‌کارگیری الگوریتم تحلیل جداساز خطی در پیش‌بینی متغیر بازده و ریسک به‌شرح نگاره‌های (۲) و (۳) است. عناوین در هر ستون معرف طبقهٔ پیش‌بینی و عناوین در هر ردیف معرف طبقهٔ واقعی است. مقادیر واقع‌شده بر روی قطر نگاره نشان‌دهنده تعداد پیش‌بینی‌هایی است که با بازده (ریسک) واقعی مطابقت دارد. این نتایج بدین معناست که برای تعداد ۳۱ سال- شرکت بازده واقعی در طبقه خیلی کم بوده است که این الگوریتم آن را به‌درستی در همین طبقه پیش‌بینی کرده است، اما ۳۱ بازده مابقی (۱۷+٥+٥+٤) را اشتباهی در طبقات دیگر پیش‌بینی کرده است. همچنین، به‌ترتیب بازده تعداد ۱٥، ۱۸، ۲۰ و ۱۷ سال- شرکت به‌درستی در طبقات کم، متوسط، زیاد و خیلی زیاد پیش‌بینی شده است. به‌طورکلی، از ۲۹۳ سال- شرکت، بازده ۱۰۱ سال- شرکت (۳۱+۱٥+۱۸+۲۰+۱۷) به‌طور صحیح پیش-بینی شده است.

نگارهٔ ۲- نتایج طبقه‌بندی پیش‌بینی بازده

طبقه	خیلی کم	کم	متوسط	زیاد	خیلی زیاد
خیلی کم	۳۱	۱۷	٥	٥	٤
کم	۹	۱٥	۱۰	۱۲	۷
متوسط	۱۸	۱٦	۱۸	۱۳	۱٥
زیاد	٥	۱۲	۱۲	۲۰	۱۳
خیلی زیاد	٦	۲	٥	٦	۱۷

نگارهٔ ۳- نتایج طبقه‌بندی پیش‌بینی ریسک

طبقه	خیلی کم	کم	متوسط	زیاد	خیلی زیاد
خیلی کم	١٦	۱۸	۱۷	۸	۱۲

کم	٢١	١٥	١٥	٨
متوسط	٦	١٠	١٠	١١
زیاد	١٠	٣	١٣	١٦
خیلی زیاد	٨	١٤	٧	٢٣

در پیش‌بینی ریسک نیز همانند نگارهٔ قبل، ارقامی که روی قطر نگاره قرار گرفته‌اند، نشان‌دهندهٔ تعداد سال- شرکت‌هایی هستند که ریسک سیستماتیک پیش‌بینی‌شدهٔ آنها در طبقهٔ واقعی ریسک قرار گرفته است. برای مثال، برای ١٦ سال- شرکت ریسک واقعی در طبقه خیلی کم بوده است و این الگوریتم ریسک را در همین طبقه پیش‌بینی کرده است؛ بنابراین پیش‌بینی ریسک برای ٥٥ سال- شرکت (١٨+٨+١٧+١٢) اشتباه انجام شده است و مطابق با طبقه واقعی ریسک نیست. به‌گونه‌ای مشابه ریسک سیستماتیک تعداد ١٥، ١٠، ١٣ و ٢٣ سال- شرکت به‌درستی در طبقات کم، متوسط، زیاد و خیلی زیاد پیش‌بینی شده است.

به‌طورکلی و براساس نتایج نگاره‌های (٢) و (٣)، درستی پیش‌بینی بازده به‌وسیلهٔ الگوریتم تحلیل جداساز خطی ٣٤/٤٧٪ (٢٩٣÷١٠١) و برای ریسک ٢٦/٢٨٪ (٢٩٣÷٧٧) است.

یافته‌های الگوریتم تحلیل جداساز غیرخطی

نتایج حاصل از به‌کارگیری الگوریتم تحلیل جداساز غیرخطی در پیش‌بینی متغیر بازده و ریسک با داده‌های آزمون (٢٩٣ سال- شرکت)، به‌ترتیب در نگاره‌های (٤) و (٥) نمایش داده شده است. همانند الگوریتم تحلیل جداساز خطی، پیش‌بینی‌های انجام‌شده با این الگوریتم، با نتایج واقعی بازده برای ٢٩٣ سال- شرکت مقایسه شد. همان‌طورکه در نگارهٔ ذیل مشاهده می‌شود، تعداد ٩٦ سال- شرکت (١٣+٢+٣٢+٦+٤٣) به‌طور صحیح و مطابق با نتایج واقعی پیش‌بینی شده است.

در پیش‌بینی ریسک سیستماتیک نیز همانند بازده، اعداد روی قطر نگاره که ٨١ سال- شرکت (١٠+٨+٣٦+٢٣+٤) است، تعداد پیش‌بینی‌شده‌های صحیح از میان ٢٩٣ سال- شرکت آزمون‌شده هستند؛ یعنی تعداد ٨١ سال- شرکت دقیقاً در همان طبقه‌ای پیش‌بینی شده‌اند که ریسک واقعی آنها قرار گرفته است. به‌طورکلی، با توجه به نتایج نگاره‌های (٤) و (٥)، مقدار صحت پیش‌بینی با الگوریتم تحلیل جداساز غیرخطی برای بازده ٣٢/٧٦٪ (٢٩٣÷ ٩٦) و برای ریسک ٢٧/٦٤٪ (٢٩٣÷ ٨١) است.

نگارهٔ ٤- نتایج طبقه‌بندی پیش‌بینی بازده

طبقه	خیلی کم	کم	متوسط	زیاد	خیلی زیاد
خیلی کم	٤٣	٢٠	١٧	٨	٢١
کم	٢	٦	١	٣	٥
متوسط	٢٢	٢٩	٣٢	٢٣	١٧
زیاد	٣	٤	٣	٢	١
خیلی زیاد	٢	٢	٧	٧	١٣

نگارهٔ ٥- نتایج طبقه‌بندی پیش‌بینی ریسک

طبقه	خیلی کم	کم	متوسط	زیاد	خیلی زیاد
خیلی کم	٤	١	١	٣	٤
کم	١٦	٢٣	١٨	١٦	٩
متوسط	١٨	٢١	٣٦	٣٢	٤٣
زیاد	٧	١	٨	٨	٣
خیلی زیاد	٣	٣	١	٤	١٠

یافته‌های الگوریتم نزدیک‌ترین K همسایگی

الگوریتم سومی که برای پیش‌بینی ریسک و بازده به‌کار برده شده است، الگوریتم نزدیک‌ترین K همسایگی است. نتایج حاصل از به‌کارگیری این الگوریتم در پیش‌بینی متغیر بازده و ریسک با داده‌های آزمون، در نگاره‌های (٦) و (٧) منعکس شده است.

نگارهٔ ٦- نتایج طبقه‌بندی پیش‌بینی بازده

طبقه	خیلی کم	کم	متوسط	زیاد	خیلی زیاد
خیلی کم	١٢	١٠	١١	١٠	١٠
کم	١١	١٨	٧	١٩	١٠
متوسط	٧	١٣	١٠	٥	١٤
زیاد	١٤	١٤	١٣	١٣	١٣
خیلی زیاد	٨	٩	١١	١٧	١٤

نگارهٔ ٧- نتایج طبقه‌بندی پیش‌بینی ریسک

طبقه	خیلی کم	کم	متوسط	زیاد	خیلی زیاد
خیلی کم	١٦	١٤	١١	٩	١٣
کم	٢٠	١٦	١٣	١٤	١٢
متوسط	٧	١٢	١٤	٧	٩
زیاد	٤	٨	١١	١٣	١١
خیلی زیاد	١٦	٨	٧	١٣	١٥

بنابر نتایج به‌دست‌آمده در نگاره‌های (٦) و (٧)، مقدار صحت پیش‌بینی با به‌کارگیری الگوریتم نزدیکترین K همسایگی (k=١) برای بازده ٢٢/٨٦٪ (٢٩٣÷ ٦٧) و برای ریسک ٢٥/٢٦٪ (٢٩٣÷ ٧٤) است.

تعداد اعداد روی قطر نگاره‌ها که همان پیش‌بینی‌های انجام‌گرفته منطبق با نتایج واقعی است، برای نگاره بازده برابر با ٦٧ سال- شرکت (١٢+١٨+١٠+١٣+١٤) و برای نگارهٔ ریسک تعداد ٧٤ سال- شرکت (١٦+١٦+١٤+١٣+١٥) است.

یافته‌های الگوریتم درخت تصمیم طبقه بندی‌کننده

آخرین الگوریتم به‌کاررفته و به‌عبارتی چهارمین الگوریتم خبره استفاده‌شده برای داده‌های آزمون، الگوریتم درخت تصمیم طبقه‌بندی‌کننده است. همانند

روش‌های گفته‌شده در بالا، نتایج این الگوریتم نیز در نگاره‌های (۸) و (۹) به‌ترتیب برای بازده و ریسک که شامل طبقات پیش‌بینی در هر ستون و طبقات واقعی در هر ردیف هستند، نمایش داده شده است. همان‌طور که مشاهده می‌شود، اعداد روی قطر نگاره برای بازده که عبارت است از مجموع ۹۷ سال- شرکت (۲۵+۱٤+۱۱+۲٦+۲۱)، معرف تعداد پیش‌بینی‌های صحیح انجام‌شده به‌وسیلهٔ این الگوریتم است. برای نگارهٔ ریسک نیز مانند بازده اعداد روی قطر نگاره معرف تعداد پیش‌بینی‌های صحیح انجام‌شده به‌وسیلهٔ این الگوریتم است که عبارت‌اند از: ٥٦ سال- شرکت (۷+۱۷+۳+۱٥+۱٤).

به‌طورکلی، با توجه به نتایج به‌دست‌آمده از نگاره‌های (۸) و (۹)، مقدار درصد صحت پیش‌بینی با به‌کارگیری الگوریتم درخت تصمیم طبقه‌بندی‌کننده برای بازده ۳۳/۱۰٪ (۲۹۳÷ ۹۷) و برای ریسک ۱۹/۱۱٪ (۲۹۳÷ ٥٦) است.

نگارهٔ ۸- نتایج طبقه‌بندی پیش‌بینی بازده

طبقه	خیلی کم	کم	متوسط	زیاد	خیلی زیاد
خیلی کم	۲٥	۱۱	۸	٥	۲
کم	۱۷	۱٤	۱٥	۱۳	٥
متوسط	٤	۹	۱۱	٤	۸
زیاد	۸	۱٤	۱۷	۲٦	۲٤
خیلی زیاد	٦	۷	۱۰	۹	۲۱

نگارهٔ ۹- نتایج طبقه‌بندی پیش‌بینی ریسک

طبقه	خیلی کم	کم	متوسط	زیاد	خیلی زیاد
خیلی کم	۷	۹	۱٦	۱۱	۸
کم	۱٥	۱۷	۱۳	۱٥	٤
متوسط	۱۱	٥	۳	۱۱	٦
زیاد	۱٦	۱٤	۲۲	۱٥	۲۲
خیلی زیاد	۱٤	۸	۷	۱۰	۱٤

پیش‌بینی متغیرهای پاسخ با وجود متغیرهای مستقل منتخب فیلترینگ

در این بخش اهمیت وزنی ۱٦ متغیر مستقل در پیش‌بینی بازده (ریسک) با استفاده از الگوریتم فیلترینگ و تحلیل خوشه‌ای تعیین و از بین آنها ٤ متغیر، مؤثرترین متغیرها انتخاب شدند. سپس با استفاده از چهار الگوریتم پیش‌گفته دوباره به پیش‌بینی ریسک و بازده پرداخته می‌شود. بدین‌ترتیب، تأثیر رویکرد فیلترینگ بر افزایش یا کاهش صحت پیش‌بینی الگوریتم‌ها بررسی می‌شود. الگوریتم فیلترینگ با به‌کارگیری شاخص‌های مربع کای [۲۳]، ضریب پیرسون، ضریب اسپیرمن، آنتروپی Infogain، آنتروپی Gainratio [۱٦]، شاخص عدم‌قطعیت متقارن [۱۷]، الگوریتم OneR [۱۹] و الگوریتم Relief-f [۲۳] به این گزینش می‌پردازد. با توجه به نتایج تحلیل خوشه‌ای بر روی داده‌های مربوط به متغیر بازدهی، ٤ متغیر سود هر سهم، تغییرات سود هر سهم، بازده حقوق صاحبان سهام و

گردش دارایی، مؤثرترین متغیرها در پیش‌بینی بازده و ٤ متغیر سود هر سهم، تغییرات سود هر سهم، بازده حقوق صاحبان سهام و خالص سرمایه در گردش، مؤثرترین متغیرها در پیش‌بینی ریسک انتخاب شد.

یافته‌های الگوریتم تحلیل جداساز خطی

همانند حالت قبل، نتایج پیش‌بینی‌های ریسک سیستماتیک و بازده در دو نگاره به‌طور مجزا نمایش داده شده است. نتایج حاصل از به‌کارگیری این الگوریتم در پیش‌بینی متغیر بازده و ریسک به شرح نگاره‌های (۱۰) و (۱۱) است. همان‌طورکه ملاحظه می‌شود، اعداد روی قطر نگاره نمایانگر پیش‌بینی‌های بازده صحیح مطابق با بازده واقعی هستند که درجمع ۱۰۲ سال- شرکت (۳۱+٤+۲٦+۱٤+۲۷) از ۲۹۳ سال- شرکت در طبقات کم، متوسط، زیاد و خیلی زیاد پیش‌بینی شده است و مابقی ۲۹۳ سال- شرکت یعنی تعداد ۱۹۱ سال- شرکت (۱۰۲-۲۹۳) به صورت اشتباه در طبقات دیگر پیش‌بینی شده‌اند. برای ریسک سیستماتیک نیز تعداد ۸۹ سال- شرکت (۲۳+۲۳+۳+۱۳+۲۷) که روی قطر نگاره با رنگ متفاوت نمایش داده شده است، پیش‌بینی‌های صحیح را نشان می‌دهد.

نگارهٔ ۱۰- نتایج طبقه‌بندی پیش‌بینی بازده

طبقه	خیلی کم	کم	متوسط	زیاد	خیلی زیاد
خیلی کم	۳۱	۲۱	۱۰	۱۰	۲
کم	۲	٤	۰	۲	۱
متوسط	۱۵	۲٤	۲٦	۲۱	۳۰
زیاد	۲	٦	۹	۱٤	٦
خیلی زیاد	۷	۵	۷	۱۱	۲۷

نگارهٔ ۱۱- نتایج طبقه‌بندی پیش‌بینی ریسک

طبقه	خیلی کم	کم	متوسط	زیاد	خیلی زیاد
خیلی کم	۲۳	۱۸	۱۷	۱۰	۲۰
کم	۱۷	۲۳	۲۲	۱٤	۸
متوسط	٤	۲	۳	۲	۱
زیاد	٤	۳	۳	۱۳	۹
خیلی زیاد	۱٤	۹	۱۲	۱۵	۲۷

به‌طورکلی و با توجه به این نگاره‌ها می‌توان نتیجه گرفت مقدار صحت پیش‌بینی الگوریتم تحلیل جداساز خطی برای بازده ۳٤/۸۱٪ (۲۹۳÷ ۱۰۲) و برای ریسک ۳۰/۳۷٪ (۲۹۳÷ ۸۹) است.

یافته‌های الگوریتم تحلیل جداساز غیرخطی

نتایج حاصل از به‌کارگیری الگوریتم تحلیل جداساز غیرخطی در پیش‌بینی متغیر بازده و ریسک به شرح نگاره‌های (۱۲) و (۱۳) است. این نتایج بدین معناست که از ۲۹۳ سال- شرکت، بازده ۱۰۳ سال- شرکت (۵۰+۹+۱۵+۱٤+۱۵) به‌طور صحیح پیش‌بینی شده، اما ۱۹۰ بازده مابقی را اشتباهی در طبقات

دیگری پیش‌بینی کرده است. در خصوص ریسک نیز برای ۸۰ سال- شرکت (۰+٤+٥٤+١٦+٦) پیش‌بینی به‌درستی در طبقات خیلی کم، متوسط، زیاد و خیلی زیاد انجام شده است، اما برای ۲۱۳ سال- شرکت این پیش‌بینی نادرست است. به‌طورکلی و براساس نتایج نگاره‌های (۱۲) و (۱۳) صحت پیش‌بینی بازده به‌وسیلهٔ الگوریتم تحلیل جداساز غیرخطی برای بازده ۳۵/۱۵٪ (۲۹۳÷ ۱۰۳) و برای ریسک ۲۷/۳۰٪ (۲۹۳÷ ۸۰) است.

نگارهٔ ۱۲- نتایج طبقه‌بندی پیش‌بینی بازده

طبقه	خیلی کم	کم	متوسط	زیاد	خیلی زیاد
خیلی کم	٥٠	٣٤	٣٢	٣٤	٢٢
کم	٣	٩	١	٠	٤
متوسط	٥	٧	١٥	١١	٥
زیاد	١	٥	٣	١٤	١
خیلی زیاد	٤	٥	٥	٨	١٥

نگارهٔ ۱۳- نتایج طبقه‌بندی پیش‌بینی ریسک

طبقه	خیلی کم	کم	متوسط	زیاد	خیلی زیاد
خیلی کم	٠	٣	٠	٠	٢
کم	٢	٤	٦	٢	٢
متوسط	٥٥	٤٠	٥٤	٣٢	٣٩
زیاد	٤	٣	٨	١٦	١١
خیلی زیاد	٠	١	١	٢	٦

یافته‌های الگوریتم نزدیک‌ترین K همسایگی

نتایج حاصل از به‌کارگیری این الگوریتم در پیش‌بینی متغیر بازده و ریسک با داده‌های آزمون، در نگاره‌های (١٤) و (١٥) نشان داده شده است. تعداد پیش‌بینی صحیح بازده برابر با ۸۰ سال- شرکت (١٩+١٥+١٥+١٦+١٥) است. در پیش‌بینی ریسک سیستماتیک نیز اعداد روی قطر نگاره که ٧٤ سال- شرکت (١٤+١٥+١٥+٢٠+١٠) است، تعداد پیش‌بینی‌شده‌های صحیح از میان ۲۹۳ سال- شرکت آزمون‌شده هستند. بدین‌ترتیب می‌توان نتیجه گرفت مقدار صحت پیش‌بینی با به‌کارگیری الگوریتم نزدیک‌ترین K همسایگی (k=١) برای بازده ۲۷/۳۰٪ (۲۹۳÷۸۰) و برای ریسک ۲۵/۲٦٪ (۲۹۳÷٧٤) است.

نگارهٔ ١٤- نتایج طبقه‌بندی پیش‌بینی بازده

طبقه	خیلی کم	کم	متوسط	زیاد	خیلی زیاد
خیلی کم	١٩	١٢	٨	٤	٦
کم	١٦	١٥	١٣	٥	٨

متوسط	۱۳	۱۲	۱۵	۱۰	۱۱
زیاد	۱۱	۱۵	۱۵	۱۶	۱۷
خیلی زیاد	۳	۹	۱٤	۱۱	۱۵

نگارهٔ ۱۵- نتایج طبقه‌بندی پیش‌بینی ریسک

طبقه	خیلی کم	کم	متوسط	زیاد	خیلی زیاد
خیلی کم	۱٤	۱۰	۸	۹	۱۵
کم	۱۸	۱۵	۱۹	۱۵	٤
متوسط	۱۷	۱۰	۱۵	۱۳	۱۲
زیاد	٦	۱۰	٦	۲۰	۹
خیلی زیاد	۷	۵	۱۰	۱٦	۱۰

یافته‌های الگوریتم درخت تصمیم طبقه‌بندی‌کننده

در آخرین الگوریتم به‌کاررفته، نتایج الگوریتم درخت تصمیم نیز در نگاره‌های (۱٦) و (۱۷) به‌ترتیب برای بازده و ریسک که شامل طبقات پیش‌بینی در هر ستون و طبقات واقعی در هر ردیف هستند، نمایش داده شده است. همان‌طور که مشاهده می‌شود، اعداد روی قطر نگاره برای بازده که برای تفهیم مطلب با رنگ متفاوت نشان داده شده است، عبارت است از مجموع ۹۵ سال- شرکت (۲۱+۱۷+۱۷+۲۷+۱۳)، معرف تعداد پیش‌بینی‌های صحیح انجام‌شده به‌وسیلهٔ این الگوریتم است. برای نگارهٔ ریسک نیز مانند بازده اعداد روی قطر نگاره معرف تعداد پیش‌بینی‌های صحیح انجام‌شده با این الگوریتم است که عبارت‌اند از: ۵۹ سال- شرکت (٦+۱٤+۱۳+۱٦+۱۰) و تعداد پیش‌بینی‌های اشتباه انجام‌شده با الگوریتم درخت تصمیم در این مرحله برابر است با ۲۳٤ سال- شرکت (۵۹-۲۹۳).

به‌عبارت دیگر، با توجه به نتایج نگاره‌های (۱٦) و (۱۷)، مقدار صحت پیش‌بینی با به‌کارگیری الگوریتم درخت تصمیم برای بازده ۳۲/٤۲٪ (۹۵÷۲۹۳) و برای ریسک ۲۰/۱٤٪ (۵۹÷۲۹۳) است که با این حال نسبت به مرحلهٔ قبلی بهبود یافته است.

نگارهٔ ۱٦- نتایج طبقه‌بندی پیش‌بینی بازده

طبقه	خیلی کم	کم	متوسط	زیاد	خیلی زیاد
خیلی کم	۲۱	۷	۵	۲	۱
کم	۲۰	۱۷	۱۵	۱۱	۷
متوسط	۱۰	۱۵	۱۷	۱٤	۱۳
زیاد	۷	۱٤	۲۱	۲۷	۲٦
خیلی زیاد	۲	۲	۳	۳	۱۳

نگارهٔ ۱۷- نتایج طبقه‌بندی پیش‌بینی ریسک

طبقه	خیلی کم	کم	متوسط	زیاد	خیلی زیاد

خیلی کم	٦	٦	١٢	٦	٦
کم	١٦	١٤	١٢	٢١	٧
متوسط	٢١	١٤	١٣	١٤	٧
زیاد	١٥	١٨	٢٢	١٦	٢٤
خیلی زیاد	٥	١	٢	٥	١٠

مقایسهٔ نهایی الگوریتم‌ها و پاسخ به پرسش‌های پژوهش

در نگارهٔ (١٨) نتایج پیش‌بینی ریسک و بازده با حضور همهٔ متغیرهای مستقل و با استفاده از متغیرهای منتخب ارائه شده است. مقادیر انحراف صحت پیش‌بینی بازده و ریسک نشان می‌دهد رویکرد فیلترینگ و گزینش متغیرهای مهم در افزایش صحت پیش‌بینی تأثیرگذار است. به‌طوری‌که انحراف‌های مثبت، میزان بهبود در صحت پیش‌بینی را نشان می‌دهند و انحراف‌های منفی، میزان کاهش صحت پیش‌بینی را نمایش می‌دهند.

همان‌طور که در نگارهٔ زیر نشان داده شده است، با اجرای روش‌های فیلترینگ و انتخاب متغیرهای مهم در پیش‌بینی بازده، صحت پیش‌بینی دو الگوریتم تحلیل جداساز غیرخطی و نزدیک‌ترین K همسایگی به‌ترتیب به‌میزان ٢/٣٩٪ و ٤/٤٤٪ افزایش یافته است، درحالی‌که تأثیر بسزایی در صحت پیش‌بینی الگوریتم تحلیل جداساز خطی دیده نمی‌شود (٠/٣٤٪) و صحت پیش‌بینی درخت تصمیم هم تاحدودی کاهش یافته است (٠/٦٨٪-). همچنین با گزینش متغیرهای مهم در پیش‌بینی ریسک، بهبودی در صحت پیش‌بینی دو الگوریتم تحلیل جداساز خطی (٤/٠٩٪) و درخت تصمیم (١/٠٣٪) ملاحظه می‌شود؛ ولی تأثیری در نزدیک‌ترین K همسایگی ندارد (٠٪) و صحت پیش‌بینی تحلیل جداساز غیرخطی تاحدودی کاهش یافته است (٠/٣٤٪-).

به‌طورکلی نتایج حاصل‌شده نشان می‌دهد اجرای تحلیل خوشه‌ای و روش‌های فیلترینگ برای گزینش متغیرهای پیش‌بینی‌کننده بااهمیت نقش مؤثری در بهبود صحت پیش‌بینی الگوریتم‌های مورد استفاده در این پژوهش دارد و می‌تواند به‌عنوان یک روش مناسب به‌منظور افزایش صحت پیش‌بینی طبقات استفاده شود.

نگارهٔ ١٨- نتایج تأثیر گزینش متغیرهای مستقل با اهمیت در تغییرات صحت پیش‌بینی بازده و ریسک با به‌کارگیری چهار الگوریتم

ردیف	الگوریتم	با استفاده از ١٦ متغیر	با استفاده از متغیرهای منتخب	انحراف صحت بازده	با استفاده از ١٦ متغیر	با استفاده از متغیرهای منتخب	انحراف صحت ریسک
١	LDA	٪٣٤/٤٧	٪٣٤/٨١	٪٠/٣٤	٪٢٦/٢٨	٪٣٠/٣٧	٪٤/٠٩
٢	QDA	٪٣٢/٧٦	٪٣٥/١٥	٪٢/٣٩	٪٢٧/٦٤	٪٢٧/٣٠	٪٠/٣٤-
٣	KNN	٪٢٢/٨٦	٪٢٧/٣٠	٪٤/٤٤	٪٢٥/٢٦	٪٢٥/٢٦	٪٠/٠٠
٤	CDT	٪٣٣/١٠	٪٣٢/٤٢	٪٠/٦٨-	٪١٩/١١	٪٢٠/١٤	٪١/٠٣

نتیجه‌گیری

هدف پژوهش حاضر، پیش‌بینی ریسک سیستماتیک و بازده سهام شرکت‌ها با استفاده از نسبت‌های مالی و به‌کمک ٤ الگوریتم کاربردی تحلیل جداساز خطی (LDA)، تحلیل جداساز غیرخطی (QDA)، نزدیک‌ترین همسایگی (KNN) و درخت تصمیم و در نهایت مقایسهٔ صحت پیش‌بینی این الگوریتم‌هاست. بدین‌منظور، ابتدا با استفاده از ١٦ متغیر مستقل ریسک سیستماتیک و بازده پیش‌بینی شد و سپس با استفاده از رویکرد فیلترینگ و خوشه-بندی ٤ متغیر مستقل مؤثرتر انتخاب و دوباره به پیش‌بینی پرداخته شد. بر مبنای نتایج به‌دست‌آمده در حالت استفاده از همهٔ متغیرهای مستقل، الگوریتم تحلیل جداساز خطی و درخت تصمیم با بیشترین صحت پیش‌بینی برای پیش‌بینی بازده و الگوریتم تحلیل جداساز خطی و الگوریتم تحلیل جداساز غیر-خطی برای پیش‌بینی ریسک عملکرد مناسب‌تری داشته‌اند. همین‌طور با به‌کارگیری روش فیلترینگ و تحلیل خوشه‌ای و استفاده از ٤ متغیر برتر انتخاب‌شده برای هرکدام از ریسک و بازده، دو الگوریتم تحلیل جداساز خطی و تحلیل جداساز غیرخطی برای پیش-بینی بازده و دو الگوریتم جداساز خطی و الگوریتم جداساز غیرخطی برای پیش‌بینی ریسک نتایج بهتری را کسب کرده‌اند. به‌طور خلاصه، رویکرد فیلترینگ در انتخاب متغیرهای مستقل به‌طور نسبی در بهبود پیش‌بینی ریسک سیستماتیک و به‌ویژه بازده مؤثر واقع می‌شود. در مجموع چه در حالت به‌کارگیری کل متغیرهای مستقل و چه در حالت استفاده از متغیرهای منتخب، الگوریتم تحلیل جداساز خطی و الگوریتم تحلیل جداساز غیرخطی پیش‌بینی‌های نسبتاً مناسبی از ریسک و بازده ارائه می‌دهد. با توجه به موارد یادشده، کاربرد روش فیلترینگ، عملیات پیش‌بینی ریسک و بازده سهام به‌وسیلهٔ سرمایه‌گذاران و بیشتر فعالان بازار سرمایه را بهبود می‌بخشد و می‌توانند به‌جای استفاده از همهٔ متغیرهای مستقل و طولانی‌شدن زمان تحلیل‌ها، با مهم‌ترین متغیرهای مستقل نتایج قابل اتکایی را در کوتاه‌ترین زمان ممکن به‌دست آورند. اگرچه پژوهشی با شیوهٔ مشابه برای پیش‌بینی ریسک و بازده سهام انجام نشده است، اما روند پیش‌بینی ریسک و بازده سهام شرکت‌ها برای انتخاب پرتفوی بهینه با استفاده از دیگر مدل‌ها در گذشته انجام شده است که از این منظر با پژوهش حاضر مشابهت دارد و از آن جمله می‌توان به پژوهش‌های ردر و همکاران [٢٥]، ژانگ و همکاران [٢٨]، چنگ‌لی و همکاران [١٤]، تسای و هسیائو (٢٠١٠) [26]، رهنمای رودپشتی و همکاران [٥]، نیکواقبال و همکاران [٨]، احمدپور و غلامی‌جمکرانی [١] و نمازی و خواجوی [٧] اشاره کرد.

منابع

١- احمدپور، احمد و رضا غلامی. (١٣٨٤). بررسی رابطهٔ اطلاعات حسابداری و ریسک بازار (شرکت‌های پذیرفته‌شده در بورس اوراق بهادار تهران). مجلهٔ علوم اجتماعی و انسانی دانشگاه شیراز (ویژه‌نامهٔ حسابداری)، دورهٔ ٢٢، شمارهٔ ٢، صص ٣٠-١٨.

٢- ایزدی‌نیا، ناصر؛ طیبی، کمیل و علی‌اکبر کاشف. (١٣٩١). تعیین توان سود عملیاتی و تغییرات آن در تبیین و پیش‌بینی بازده سهام: مورد بازار بورس اوراق بهادار تهران. مجلهٔ دانش حسابداری، سال سوم، شمارهٔ ٩، صص ٣٢-٧.

۳- ایزدی‌نیا، ناصر و امیرحسین کربلایی‌کریم. (۱۳۹۰). شناسایی تأثیر متغیرهای منتخب مالی بر بازده سهام در بورس اوراق بهادار تهران. مجلهٔ پژوهش‌های حسابداری مالی، سال ٤، شمارهٔ ۱، شمارهٔ پیاپی ۱۱، صص ۳۰-۱۷.

٤- خواجوی، شکرالله؛ الله‌یاری، حمید و میثم قاسمی. (۱۳۹۰). آزمون مدل بازده و مدل قیمت در شرکت‌های پذیرفته‌شده در بورس اوراق بهادار تهران با استفاده از الگوی پانل با داده‌های متوازن. مجلهٔ پژوهش‌های حسابداری مالی، سال ۳، شمارهٔ ٤، شمارهٔ پیاپی ۱۰، صص ۵۵-۷۰.

۵- رهنمای رودپشتی، فریدون؛ نیکومرام، هاشم؛ طلوعی اشلقی، عباس؛ حسین‌زاده لطفی، فرهاد و مرضیه بیات. (۱۳۹٤). بررسی کارایی بهینه‌سازی پرتفوی براساس مدل پایدار با بهینه‌سازی کلاسیک در پیش‌بینی ریسک و بازده پرتفوی. مهندسی مالی و مدیریت اوراق بهادار، دورهٔ ٦، شمارهٔ ۲۲، صص ٦۰-۲۹.

٦- کانتاردزیک، مهمد. (۱۳۸۵). داده‌کاوی. ترجمهٔ امیر علیخانزاده. بابل، علوم رایانه، چاپ سوم.

۷- نمازی، محمد و شکرالله خواجوی. (۱۳۸۳). سودمندی متغیرهای حسابداری در پیش‌بینی ریسک سیستماتیک شرکت‌های پذیرفته‌شده در بورس اوراق بهادار تهران. بررسی‌های حسابداری و حسابرسی، سال یازدهم، شمارهٔ ۳۷، صص ۹۳-۱۱۹.

۸- نیکواقبال، علی‌اکبر؛ گندلی‌علیخانی، نادیا و اسماعیل نادری. (۱۳۹۲). ارزیابی مدل‌های شبکه عصبی مصنوعی ایستا و پویا در پیش‌بینی قیمت سهام. فصلنامهٔ علمی پژوهشی دانش مالی تحلیل اوراق بهادار، شمارهٔ ۲۲، صص ۹۱-۷۷.

9- Ball, R. and P. Brown. (1968). An empirical evaluation of accounting income numbers. Journal of Accounting Research, Vol. 6, No. 2, Pp. 159-178.

10- Barak, S. and M. Modarres. (2015). Developing an approach to evaluate stocks by forecasting effective features with data mining methods, Expert Systems with Applications, No. 42, Pp. 1325-1339.

11- Barth, M.E., Beaver, W.H. and W.R. Landsman. (1998). Relative valuation roles of equity book value and net income as a function of financial health. Journal of Accounting and Economics, No. 25, Pp. 1-34.

12- Beaver, W.H., Kettler, P. and M. Scholes. (1970). The Association Between Market Determind and Accounting Determind Risk Measures. Accounting Review, Vol. 45, No. 4, Pp. 654-682.

13- Brimble, M.A. (2003). The Relevance of Accounting Information for Valuation and Risk. Phd Thesis, Pp. 1-304.

14- Cheng Li, J. and D. Cheng Mei. (2013). The risks and returns of stock investment in a financial market. Physics Letters A, No. 377, Pp. 663-670.

15- Chu, L., Mathieu, R., Mbagwu, C. and P. Zhang. (2013). The Usefulness of Accttttigg Iffrrmttinn dddrrrm Oprational Risks. working paper, www.degroote.mcmaster.ca.

16- Duda, R.o., Hart, P.E. and D.G. Strok, (2001). Pattern classification. Wiley.

17- Dumais, S., Platt, J., Heckerman, D. and M. Sahami. (1998). Inductive learning algorithms and representations for text categorization. In Proceedings of the international conference on information knowledge management. Pp. 148˚155.

18- Elgers, P. and D. Murray. (1982). The impact of the choice of market index on the empirical evaluation of accounting risk measures. The Accounting Review, Vol. 57, No. 2, Pp. 358-375.

19- Holte, R.C. (1993). Very simple classification rules perform well on most commonly used datasets. Machine Learning, No. 11, Pp. 63˚90

20- Huang, C.F. (2012). A hybrid stock selection model using genetic algorithms

and support vector regression. Applied Soft Computing, No. 12, Pp. 807˚818.

21- James, G., Witten, D., Hastie, T. and R. Tibshirani. (2013). An introduction to statistical learning with applications in R. Springer science + business media New York.

22- Jayawardena, N., Todorova, N., Li, B. and J. Su. (2016). Forecasting stock volatility using after-hour information: Evidence from the Australian Stock Exchange. Economic modeling, No. 52, Pp. 592-608.

23- Kononenko, I. (1994). Estimating attributes: analysis and extensions of relief. In Proceedings of the seventh European conference on machine learning. Pp. 171˚182.

24- Oztekin, A., Kizilaslan, R., Freund, S. and A. Iseri. (2016). A Data Analytic Approach to Forecasting Daily Stock Returns in an Emerging Market. European Journal of operational research, Vol. 253, No. 3, Pp. 697-710.

25- Rather, A.M., Agarwal, A. and V.N. Sastry. (2015). Recurrent neural network and a hybrid model for prediction of stock returns, Expert Systems with Applications, No. 42, Pp. 3234-3241.

26- Tsai, CH. and Y. Hsiao. (2010), Combining multiple feature selection methods for stock prediction: union, intersection and multi-intersection approaches, Decision support systems, No. 50, Pp.258-269

27- Watts, R. and J. Zimmerman. (1986). Positive Accounting Theory, Prentice Hall International, Inc.

28- Zhang, X., Hu, Y., Xie, K., Wang, Sh., Ngai, E.W.T. and M. Liu. (2014). A causal feature selection algorithm for stock prediction modeling. Neurocomputing, No. 142, Pp. 48-59.