

فصلنامه علمی - پژوهشی مطالعات تجربی حسابداری مالی - سال ۱۵ - شماره ۵۹ - پائیز ۱۳۹۷
صفحات ۱۳۵ تا ۱۵۷

خوشه‌بندی و پیش‌بینی سودآوری شرکت‌های پذیرفته شده در بورس اوراق بهادار تهران با رویکرد درخت تصمیم C5

محمدرضا مهربان‌پور *

ملیحه حبیب‌زاده **

چکیده

امروزه سرمایه‌گذاران با توجه به فضای رقابتی حاکم باید محتاط‌تر از قبل تصمیم بگیرند. به این منظور آنها می‌توانند از بانک‌های اطلاعاتی بورس اوراق بهادار استفاده کنند. اما این اطلاعات به تنهایی مضمّن نیست، بنابراین لازم است با استفاده از فنون داده‌کاوی تجزیه، تحلیل و تفسیر داده‌ها انجام شود تا اطلاعات قابل‌انکاتری در اختیار استفاده‌کنندگان قرارگیرد. هدف این پژوهش خوشه‌بندی و پیش‌بینی سودآوری شرکت‌ها و تعیین عوامل موثر بر سودآوری شرکت‌های عضو بورس اوراق بهادار تهران است. جهت این کار، ۸۸ شرکت - سال در محدوده زمانی ۱۳۸۷-۱۳۹۵ انتخاب شدند. پس از پیش‌پردازش اولیه داده‌ها، با نرم‌افزارهای متلب و Clementine و با استفاده از معیار SSE و روش K-Means شرکت‌ها به ۳ خوشه تبدیل شدند و نتایج این خوشه‌بندی‌ها بوسیله معیار سنجش کیفیت، مورد سنجش قرار گرفت. در ادامه با استفاده از درخت تصمیم C5 خوشه‌ها تحلیل و متغیرهای تاثیرگذار بر سودآوری، شناسایی شد. از ۳۲ متغیر تحلیل شده تنها ۸ متغیر شامل: سودخالص به کل دارایی، فروش به کل دارایی، سودخالص به حقوق صاحبان سهام، سود عملیاتی به فروش خالص، سود و زیان انباشته به حقوق صاحبان سهام، سودخالص به فروش خالص، کل بدهی‌ها به کل دارایی‌ها و دارایی‌های جاری به کل دارایی‌ها بر سودآوری شرکت‌ها تاثیر می‌گذارند. در نهایت با در نظر گرفتن این متغیرها، پیش‌بینی سودآوری شرکت‌ها طبق هر خوشه انجام شد که دقت پیش‌بینی خوشه‌ها به ترتیب ۸۶/۳۴ درصد، ۸۸/۱۵ درصد و ۸۱/۶۸ درصد است.
واژگان کلیدی: پیش‌بینی، سودآوری، خوشه‌بندی، درخت تصمیم C5.

* استادیار حسابداری دانشگاه تهران (نویسنده مسئول) mehrabanpour@ut.ac.ir

** کارشناس ارشد حسابداری، مدرس دانشگاه قم

تاریخ پذیرش: ۱۳۹۷/۰۷/۱۷

تاریخ دریافت: ۱۳۹۷/۰۱/۲۷

مقدمه

پیش‌بینی آینده در زمینه‌های مختلف، همواره برای انسان جالب و جذاب بوده است. بااطمینان می‌توان گفت که پیش‌بینی آینده و روند تغییرات در همه حوزه‌ها، از دغدغه‌های اصلی و همیشگی مدیران سطح بالا و میانی می‌باشد. از آنجایی که شناخت فرصت‌های مناسب سرمایه‌گذاری در بازار اوراق بهادار، نیازمند تحلیل دقیق صورت‌های مالی شرکت‌ها و پیش‌بینی مناسب سودآتی می‌باشد، بنابراین تحلیل و پیش‌بینی متغیرهای موثر بر سودآوری شرکت‌ها همواره مدنظر سرمایه‌گذاران، کارگزاران و سایر افراد ذی‌نفع در بازار اوراق بهادار می‌باشد (Finger, 1994). امروزه فضای رقابتی شدیدی بین شرکت‌های مختلف حاکم شده است؛ بنابراین مدیران باید سریع‌تر و درست‌تر از قبل تصمیم بگیرند. لازمه چنین امری، پیش‌بینی دقیق از سودآوری آتی شرکت‌ها و شناسایی عوامل موفقیت در سودآوری خود می‌باشد. پیش‌بینی سود باید اطلاعاتی را فراهم کند که منطقی و به موقع باشد تا بتواند نیازهای استفاده‌کنندگان را برطرف نماید و به تبع آن، تاثیر خود بر شرکت را به نحو مناسبی نشان دهد (Ying Chan, 2016). میزان اطلاعات افشا شده برای مشارکت‌کنندگان در بازار سرمایه، بر دقت پیش‌بینی سود تحلیلگران تاثیر می‌گذارد، یافته‌ها بیانگر آن است که شرکت‌هایی که اطلاعات بیشتری برای استفاده‌کنندگان خارجی منتشر می‌کنند، دقت پیش‌بینی سود بیشتری دارند (کردستانی، ۱۳۸۹).

اما افزایش حجم اطلاعات درباره شرکت‌ها، عملاً کار شناسایی عوامل تاثیرگذار در سودآوری شرکت‌ها به روش رایج و سنتی را سخت و دشوار کرده و مسلماً با خطاهای انسانی و مشکلاتی همراه است. از جمله این مشکلات، طولانی بودن فرآیند شناسایی و یا سطحی بودن امر شناسایی یا اعمال سلیقه‌های کارشناسان می‌باشد. افزایش این حجم اطلاعات نیز این خطاها را بیش‌از پیش افزایش می‌دهد. بنابراین، استفاده از روش‌هایی که به کمک آن بتوان شناخت بیشتر و ارزیابی دقیق‌تری از سودآوری شرکت‌ها بدست آورد، مهم می‌باشد (افسر و همکاران، ۱۳۹۳). برای اینکه شناسایی بهتری در این حجم انبوه از اطلاعات داشته باشیم و نیز جهت کاهش زمان فرآیند، می‌توان با خوشه‌بندی؛ شرکت‌های همگن با ارزش یکسان را در خوشه‌های کوچک‌تر مشابه قرارداد تا حل مساله راحت‌تر شود. خوشه‌بندی در واقع یک روش برای دسته‌بندی داده‌ها است که داده‌ها را با توجه به میزان شباهتشان در دسته‌هایی قرار می‌دهد.

باتوجه به مطالب ذکر شده، مساله مورد نظر این پژوهش این است که آیا می‌توان

با استفاده از اطلاعات مالی، عملکرد و وضعیت سودآوری شرکت‌ها، و به شکل قابل‌اتکایی نتیجه عملکرد آتی آنها را پیش‌بینی نمود؟ در این صورت خوشه بهینه سودآوری شرکت‌های پذیرفته شده در بورس اوراق بهادار چه تعداد خواهد بود؟

مبانی نظری، ادبیات و سوالات

پژوهش‌هایی که مستقیماً حوزه این تحقیق را پوشش دهد، به ندرت یافت می‌شوند. با این حال در این بخش، جهت آشنائی با موضوع پیش‌بینی سود و عوامل تاثیرگذار بر آن و روش‌هایی که جهت این امر پیشنهاد شده است، نزدیک‌ترین پژوهش‌های انجام شده در خارج و داخل ارائه و مرور می‌شوند.

پژوهش‌های خارجی

گراهام و همکاران (۱۹۶۲) در پژوهش خود جهت پیش‌بینی سود آتی، از مطالعه و بررسی متوسط سودهای گذشته در طول زمان استفاده کردند و اعتقاد داشتند که تنها راه پیش‌بینی سود، استفاده از متوسط سودهای گذشته است و این نظر توسط پژوهشگران دیگر نیز مورد توجه قرار گرفت و الگوهای قابل‌قبولی نیز در این زمینه ارائه شد. اما محقق دیگری به نام لیو (۱۹۹۳) در بررسی‌های خود دریافت که غیر از سری‌های زمانی سودهای گذشته، اطلاعات دیگری نیز می‌تواند در پیش‌بینی سودهای آتی موثر باشد، ولی در صورت عدم دسترسی به سایر اطلاعات، سودهای گذشته بهترین پارامتر در پیش‌بینی سود آتی واحد تجاری است. در راستای تحقیق لیو، محقق دیگری به نام فینگر (۱۹۹۴)، به بررسی سود در پیش‌بینی سودها و جریان‌های نقدی آتی پرداخت. در این تحقیق که دوره زمانی سال‌های ۱۹۳۵-۱۹۸۷ را پوشش می‌داد و نمونه‌ای متشکل از ۵۰ شرکت پذیرفته شده در بورس نیویورک را دربرمی‌گرفت، این نتیجه حاصل شد که برای ۸۸ درصد شرکت‌های نمونه، سودهای گذشته را می‌توان به عنوان یک پیش‌بینی‌کننده خوب برای سودهای آتی دانست. از طرف دیگر چاریتو و همکارانش (۲۰۰۰) پژوهشی با عنوان "مربوط بودن ارزش سودها و گردش وجوه نقد" در ژاپن انجام دادند. نتایج پژوهش آنها نشان داد که سودها و گردش وجوه نقد هر دو در پیش‌بینی سودهای آتی موثرند.

توماس و ژانگ (۲۰۰۲) در پژوهشی با ۳۹۳۱۵ مشاهده بین سال‌های ۱۹۷۰ تا ۱۹۹۷، به بررسی ارتباط بین تغییرات موجودی کالا و بازدهی آتی شرکت‌ها پرداختند. نتایج پژوهش آنان نشان داد که ارتباط معکوس بین ارقام تعهدی و بازدهی غیرعادی آتی

شرکت‌ها، کاملاً از تغییرات در موجودی کالا ناشی می‌شود. آنها بیان کردند که موجودی کالا عامل تعیین‌کننده بسیار مهمی برای عملکرد و ارزش شرکت است. در راستای تحقیقات لیو، کاسکی و هانلن (۲۰۰۵) با بررسی رابطه بازده سال جاری سهام و سود آتی برای واحدهای تجاری که در سال جاری سود پرداخت می‌کنند، درمقایسه با واحدهای تجاری که سود پرداخت نمی‌کنند، به این نتیجه رسیدند که رابطه میان سود سال آتی و بازده سال جاری در شرکت‌های توزیع‌کننده سود، رابطه مستقیم و معناداری است.

بینون و همکاران (۲۰۰۸) در کار مطالعاتی خود با استفاده از متغیرهای حسابداری در شرکت‌های انگلستان و با بکارگیری تکنیک داده‌کاوی به پیش‌بینی طبقه سودآور و غیرسودآور پرداختند که نتایج ارائه‌شده، صحت الگوی اندرس و همکاران را نشان می‌داد. میوجنگ و همکاران (۲۰۱۰)، در بعد دیگری فعالیت کرده و نتایج تحقیق آنها در خصوص شناسایی عوامل تعیین‌کننده برتری صحت نسبی پیش‌بینی سود مدیریت بر پیش‌بینی سود توسط تحلیل‌گران عبارتند از: ۱. زمانی که کیفیت سود افزایش یابد، برتری صحت نسبی پیش‌بینی سود مدیریت بر پیش‌بینی سود توسط تحلیل‌گران کاهش می‌یابد. ۲. زمانی که پیش‌بینی سود با دشواری همراه باشد، برتری صحت پیش‌بینی سود مدیریت بر پیش‌بینی سود توسط تحلیل‌گران کاهش می‌یابد. در راستای استفاده از نرم‌افزارها برای پیش‌بینی، اگنس و همکاران (۲۰۱۱) با بکارگیری یک سیستم هوشمند نرم‌افزاری، اقدام به طراحی الگو برای پیش‌بینی عملکرد شرکت‌ها نمودند. نتایج کار آنها انعکاس‌گویای میزان خطای پایین این الگوریتم در این پیش‌بینی‌ها بود. در ادامه تحقیقات اگنس در زمینه پیش‌بینی، گوارا و همکاران (۲۰۱۴) توانایی پیش‌بینی مدل‌های تحلیل تمایزی چندگانه و شبکه‌های عصبی را در پیش‌بینی ورشکستگی شرکت‌ها بررسی کردند. نتایج این پژوهش نشان داد هر دو مدل قادر به پیش‌بینی ورشکستگی شرکت‌ها می‌باشد اما شبکه‌های عصبی به طور معنی‌داری از توانایی بالاتری برخوردار است.

چوی، کلی و سدکا (۲۰۱۶) در پژوهشی با عنوان «اخبار سود، سودهای موردانتظار و بازده انباشته سهام» به بررسی رابطه بین تغییرات سود و بازده سهام انباشته پرداختند. نتایج آنان نشان داد بین اخبار سود و بازده سهام همزمان رابطه مستقیمی وجود دارد. همچنین مورامیا و تاکادا (۲۰۱۷) در پژوهشی، عنوان می‌کنند که عدم دقت در پیش‌بین سود انجام شده توسط مدیریت، می‌تواند ناشی از در دسترس نبودن اطلاعات اولیه موردنیاز پیش‌بینی (ضعیف بودن اطلاعات ورودی مورد نیاز پیش‌بینی) و یا بکارگیری یا تفسیر نادرست این

اطلاعات توسط مدیریت باشد.

پژوهش‌های داخلی

در داخل کشور نیز مطالعاتی در خصوص کمک به پیش‌بینی بهتر سود شرکت‌ها انجام شده است. خالقی مقدم و آزاد (۱۳۷۷) دقت پیش‌بینی سود شرکت‌ها که توسط مدیریت اعلام می‌شود را در ۴۵ شرکت بورس اوراق بهادار تهران مورد بررسی قرار دادند. ارتباط چهار متغیر قیمت سهام، اندازه، عمر شرکت و درجه اهرم مالی با دقت پیش‌بینی سود، با استفاده از رگرسیون‌های یک و چند متغیره مورد آزمون قرار گرفت. نتایج آزمون فقط ارتباط بین دقت پیش‌بینی و متغیرهای قیمت سهام و اندازه شرکت را تایید نمود. در مطالعه ای دیگر، جنت رستمی (۱۳۷۸) نقش سود در پیش‌بینی جریان‌های نقدی و سودهای آتی را بررسی کرد. نمونه این تحقیق ۵۱ شرکت پذیرفته شده در بورس اوراق بهادار تهران و روش شناسی آن برگرفته از روش تحقیق فینگر در سال ۱۹۹۴ بود و نتایج نشان داد که برای ۹۲/۱۶ درصد شرکت‌های نمونه، سودهای تاریخی را می‌توان به عنوان یک پیش‌بینی کننده خوب برای سودهای آتی دانست. برخلاف تحقیقات محققینی چون لیو، نوروش و غلام زاده (۱۳۸۲) در بررسی رفتار سود حسابداری با استفاده از سری‌های زمانی باکس-جنکیز نشان دادند که سودهای گذشته در مورد سودهای آینده اطلاعات چندانی ارائه نمی‌کند.

در یکی از پژوهش‌ها با استفاده از اطلاعات دوساله ۸۱ و ۸۲ و ۱۲۰ شرکت و ۶ متغیر اقدام به بررسی و پیش‌بینی شرکت‌های موفق و ناموفق کردند. متغیرهای مورد بررسی آنها متشکل از ۳ متغیر مالی و ۳ متغیر غیرمالی بود. در مدل نهایی آنها، تنها دو متغیر بازده حقوق صاحبان سهام و رشد فروش به عنوان متغیرهای توضیح‌دهنده شرکت‌های موفق و ناموفق استفاده شد و سایر متغیرها رابطه مهمی با طبقه‌بندی انجام شده نداشتند (مهرانی و همکاران، ۱۳۸۳). در این میان بهرامفر و ساعی (۱۳۸۵) در پژوهشی نشان دادند که نسبت‌های مرتبط با سنجش فعالیت، و نسبت‌های مرتبط با اندازه‌گیری وضعیت بدهی‌ها، اندازه شرکت و نوع صنعت، در پیش‌بینی رتبه عملکرد شرکت‌ها مفید می‌باشند. برخلاف نتایج حاصل از تحقیق نوروش و همکاران، مدرس و عباس زاده (۱۳۸۷) در پژوهشی به این نتیجه رسیدند که سودهای گذشته می‌توانند سود آتی را با کمترین خطای ممکن پیش‌بینی کنند و همچنین ورود یکی از اجزای سود (نقدی و تعهدی) به مدل‌ها، پیش‌بینی را بهبود می‌بخشد و در ضمن مشخص شد تاثیر جزء نقدی در مدل‌های پیش‌بینی از اقلام تعهدی

بیشتر است. در ادامه تحقیقات خالقی مقدم، نمازی و شمس الدینی (۱۳۸۷) در بررسی سازه های موثر بر دقت پیش بینی سود، با استفاده از نمونه ۷۷ تایی از شرکت ها در طی سال های ۱۳۷۹ تا ۱۳۸۳ به این نتیجه رسیدند که بین رشد سود، رشد فروش، رشد دارایی ها، سود پیش بینی شده گذشته، اهرم مالی، قیمت سهام و دقت پیش بینی سود، رابطه وجود دارد؛ اما بین سود سهام پرداختی و اندازه شرکت با دقت پیش بینی سود، رابطه ای نیست. همچنین یکی از پژوهش ها برای بررسی عوامل موثر بر میزان دقت پیش بینی سود شرکت، ۵ عامل افق زمانی پیش بینی، دفعات تجدیدنظر، نوع اظهار نظر حسابرس، نوع صنعت و اندازه شرکت را مورد توجه قرار داد. نتایج تحقیق حاکی از آن بود که از بین عوامل گفته شده، افق زمانی و نوع صنعت بر دقت پیش بینی سود شرکت ها موثرند و بین سایر عوامل با دقت پیش بینی سود رابطه معناداری مشاهده نشد (حقیقت و همکاران، ۱۳۹۰).

اعتمادی و همکاران (۱۳۹۱) با بکارگیری شبکه عصبی مصنوعی با استفاده از اطلاعات هفت ساله ۸۰ - ۸۶، ۹۰ شرکت و ۴۲ متغیر اقدام به بررسی و پیش بینی سودآوری شرکت های بورس اوراق بهادار تهران کردند. در مدل نهایی آنها، تنها ۹ متغیر به عنوان توانمندترین متغیرها در تشخیص و تفکیک دو گروه شرکت های سودآور و زیان آور استفاده شده است. نتایج کار آنها با ۸۶ درصد صحت نشان از کارایی مناسب این تکنیک داشت. در پژوهش دیگری حاکی از تاثیرات پیش بینی سود، حسینی نسب و همکاران (۱۳۹۴) به بررسی واکنش بازار پرداختند. هدف آن ها بررسی واکنش بازار به اخبار خوب یا بد شرکت ها با توجه به نوع پیش بینی سود سال قبل آن ها، تمایل مدیران به حفظ یا اصلاح شهرت خود در پیش بینی و مقایسه تمایل مدیران به مدیریت افزایشی سود از طریق انواع مختلف مدیریت سود با توجه به نوع پیش بینی سال قبل آن ها بود. آن ها صورت های مالی ۸۶ شرکت را طی سال های ۱۳۸۵-۱۳۹۱ مورد تجزیه و تحلیل قرار دادند. نتایج پژوهش بیانگر آن بود که واکنش بازار به پیش بینی سود با محتوای اخبار خوب (بد)، زمانی که پیش بینی دوره قبل بدبینانه (خوشبینانه) است، مثبت تر (منفی تر) از حالتی است که پیش بینی دوره قبل خوشبینانه (بدبینانه) است. همچنین مدیران در پیش بینی های خود از ثبات رفتاری برخوردار می باشند. علاوه بر این، نتایج حاکی از آن بود که پیش بینی خوشبینانه در دوره قبل می تواند عاملی جهت مدیریت افزایشی سود دوره جاری باشد. رحمانی و حیاتی (۱۳۹۵) با بررسی تاثیر دقت برآوردی پیش بینی سود مدیران بر رانش پس از اعلان سود نشان دادند که ویژگی های پیش بینی سود به طور معناداری می تواند

دقت واقعی پیش‌بینی سود مدیران را توضیح دهد. از طرفی سود غیرمنتظره نمی‌تواند بازده‌های غیرعادی تعدیل شده را تبیین کند. همچنین مهربان‌پور و همکاران (۱۳۹۶) نیز بادر نظر گرفتن ۱۱ متغیر، اقدام به شناسایی عوامل موثر بر سودآوری کردند. در پژوهش آنها، متغیر بازده حقوق صاحبان سهام به عنوان معیار سودآوری مطرح شد و سپس با استفاده از اطلاعات مالی ۱۰ ساله بانک‌ها؛ ساختار دارایی، تنوع درآمدی، رشد اقتصادی و تورم به عنوان متغیرهای موثر بر سودآوری شناسایی شدند.

علاوه بر موارد بیان شده در فوق، تحقیقات انجام شده در زمینه خوشه‌بندی را می‌توان به صورت خلاصه در جدول ۱ مشاهده نمود:

جدول ۱. تحقیقات انجام گرفته در زمینه خوشه‌بندی

پژوهش	سال	محقق
ایده جداسازی گروه‌ها در یک جمعیت	۱۹۳۶	فیشر
توسعه اولین سیستم ارزیابی تقاضانامه‌های اعتباری را با بکارگیری ۵ معیار	۱۹۳۸	آلتمن
ارائه آنالیز ممیزی چند متغیره برای رتبه‌بندی اعتباری مشتریان	۱۹۶۳	مایرز و هنرجی
رتبه‌بندی شرکت‌ها با استفاده از روش آنالیز ممیزی چند متغیره	۱۹۶۸	مور و کلن
طبقه‌بندی مشتریان وام‌های بین‌المللی در سه کشور آمریکا، آلمان و استرالیا	۱۹۹۸	دسای و همکاران
تقسیم مشتریان را به دودسته خوب و بد با استفاده از شبکه عصبی	۲۰۰۰	وست
طبقه‌بندی مشتریان بانک به سه گروه عمده سودآور با استفاده از شبکه عصبی	۲۰۰۴	هسیه

از جمله مقالات معدودی که در حوزه بورس برای تقسیم‌بندی انجام شد، تحقیقی است که شین و سون (۲۰۰۴) انجام داده‌اند. در این تحقیق با استفاده از روش K-means و SOM توانستند نرخ کمیسیون کارگزاری را برای مشتریان پیشنهاد دهند. نتایج حاصل از تجزیه و تحلیل آنها نشان داد که روش k-means قویترین روش برای تقسیم‌بندی است. به همین منظور در این پژوهش نیز از این روش برای خوشه‌بندی شرکت‌های مذکور استفاده شده است.

پژوهش حاضر، فاقد فرضیه می‌باشد. در عوض سوالاتی به شرح زیر طرح شده که در بخش‌های بعدی به آنها جواب داده شده است: ۱. آیا با خوشه‌بندی و پیش‌بینی سودآوری شرکت‌ها می‌توان به شکل قابل‌اتکایی، نتیجه عملکرد آتی آنها را پیش‌بینی نمود؟ ۲. خوشه‌بینه سودآوری شرکت‌های پذیرفته شده در بورس اوراق بهادار چه تعداد است؟

۳. عوامل موثر بر سودآوری شرکت‌های عضو بورس اوراق بهادار تهران کدام هستند؟

روش تحقیق

داده‌کاوی به بررسی و تجزیه تحلیل مقدار عظیمی از داده‌ها به منظور کشف الگوها و قوانین معنی‌دار اطلاق می‌گردد. داده‌کاوی در دو نوع ظاهر می‌شود: داده‌کاوی هدایت شده و داده‌کاوی هدایت نشده. در داده‌کاوی هدایت شده هدف دسته‌بندی اطلاعات براساس برخی پارامترها مشخص می‌باشد اما در داده‌کاوی هدایت نشده هدف یافتن الگوها یا تشابهات بین گروه‌هایی از اطلاعات، بدون استفاده از هیچگونه پیش‌زمینه‌ای در مورد اطلاعات می‌باشد. از نمونه روش‌های داده‌کاوی هدایت نشده و هدایت شده می‌توان به خوشه‌بندی و دسته‌بندی اشاره نمود. خوشه‌بندی، به عمل تقسیم جمعیت ناهمگن به تعدادی از زیر مجموعه‌ها یا گروه‌های همگن گفته می‌شود. در دسته‌بندی هر داده به دسته‌ای از پیش تعیین شده براساس دانش قبلی اختصاص می‌یابد؛ اما در خوشه‌بندی هیچ دسته‌ای از پیش تعیین شده‌ای وجود ندارد. در واقع خوشه‌بندی راهی برای یافتن ساختار داده‌های پیچیده فراهم می‌کند (باقرزاده و همکاران، ۱۳۸۷). خوشه‌بندی در واقع تقسیم کردن داده‌ها به گروه‌های مشابه است و هر گروه را اصطلاحاً یک خوشه می‌نامند. داده‌ها در هر خوشه به یکدیگر شبیه و در عین حال با داده‌های موجود در خوشه‌های دیگر متفاوت اند (Abraham, 2006). به همین دلیل در این پژوهش در مرحله اول، خوشه‌بندی داده‌ها انجام می‌گیرد.

روش‌های متعددی برای خوشه‌بندی وجود دارد. یکی از الگوریتم‌های رایج خوشه‌بندی، الگوریتم K-means است. روش K-Means با وجود سادگی آن، یک روش پایه برای بسیاری از روش‌های خوشه‌بندی دیگر محسوب می‌شود (Alpaydin, 2004). ولی تعیین تعداد خوشه بهینه در ابتدای امر از معایب آن می‌باشد و نتایج به انتخاب اولیه خوشه‌ها وابسته است. بنابراین می‌توان گفت که تعیین تعداد خوشه‌ها از اهمیت زیادی برخوردار بوده و بر نتیجه کار تاثیر خواهد گذاشت. از این رو، برای تعیین تعداد خوشه بهینه از روشی به نام SSE استفاده شده است. در این روش، نخست مراکز خوشه در نظر گرفته می‌شود و سپس فاصله نقطه مورد نظر از مراکز خوشه محاسبه می‌گردد. خوشه‌ای که SSE پایین‌تر دارد، نشان‌دهنده بهترین خوشه‌بندی (تعداد خوشه‌ها) است (مطیعیان و نعیمی، ۱۳۹۲). برای این امر از رابطه ۱ استفاده می‌شود:

$$SSE = \sum_{i=1}^k \sum_{p \in C_i} d(P, m_i) \quad \text{رابطه ۱}$$

الگوریتم K-Means به شرح زیر می باشد:

۱. انتخاب K داده به عنوان مراکز خوشه
۲. تعیین فواصل بقیه داده ها با مراکز خوشه‌ها
۳. قرار گیری داده هایی که به مرکز هر خوشه نزدیک ترند در آن خوشه
۴. محاسبه میانگین هر خوشه به عنوان مرکز جدید خوشه
۵. تکرار مرحله دوم تا چهارم تا رسیدن با عدم تغییر در خوشه‌ها (دهقان وهمکاران، ۱۳۹۲).

معیار سنجش کیفیت خوشه ها

همانطور که اشاره شد، گونه ای از خوشه‌بندی بهتر است که فاصله درون خوشه‌ها در آن کمترین و فاصله بین خوشه‌ها بیشترین باشد. برای مقایسه خوشه‌بندی‌های انجام شده به کمک الگوریتم‌های خوشه‌بندی، این معیار سنجش کیفیت مطرح شده است:

اگر $O = \{c^n | n = 1, \dots, k\}$ مجموعه مراکز خوشه‌ها و c^n مراکز خوشه ها باشد و $O^n = \{c_i | i = 1, \dots, |T^c - O|\}$ مجموعه باقیمانده شرکت ها که به عنوان مرکز خوشه انتخاب نشده‌اند و T^c مجموعه کلیه شرکت‌هایی که خوشه‌بندی روی آنها صورت گرفته است؛ باشد، کیفیت نتایج خوشه‌بندی با K خوشه می‌تواند بصورت رابطه ۲ تعریف شود:

شود:

(۲)

$$\rho(k) = \frac{1}{k} \sum_{n=1}^k (\max \left\{ \frac{\gamma_n + \gamma_m}{\delta_{nm}} \right\}) \quad (۳)$$

$$\gamma_m = \frac{1}{||O^m||} \sum_{c_j \in O^m} \text{dist}(c_j, c^m) \quad (۴)$$

$$\gamma_n = \frac{1}{||O^n||} \sum_{c_i \in O^n} \text{dist}(c_i, c^n) \quad (۵)$$

$$\delta_{nm} = \text{dist}(c^n, c^m)$$

معادله (۳)، γ_m را به عنوان میانگین فاصله بین مراکز خوشه c^m و همه شرکت ها خوشه O^m را تعریف می‌کند. معادله (۴) نیز توضیحاتی مشابه معادله مذکور دارد. معادله (۵) δ_{nm} را به عنوان فاصله c^n و c^m تعریف می‌کند. معیار فوق هرچه کمتر باشد تراکم

درون خوشه‌ای بیشتر و فاصله بین خوشه‌ای بیشتر است، در نتیجه کیفیت خوشه بندی صورت گرفته بهتر است (Tsai and chiu, 2004).

برای اجرای این تحقیق، پس از پالایش و آماده سازی داده ها، برای آن که بتوان سودآوری شرکت‌ها را به صورت دقیق تری پیش‌بینی کرد، از سیستم های هوشمند و تکنیک درخت تصمیم C5 استفاده شد. این روش به خاطر قابلیت برتر این روش پیش‌بینی، نسبت به انواع مسائل مدل سازی پیش‌بینی و محبوبیت در انتشارات اخیر ادبیات داده کاوی انتخاب شده است.

درخت تصمیم به صورت بازگشتی، مشاهدات مجزا را در یک شاخه قرار می‌دهد تا یک درخت به منظور دستیابی به بالاترین دقت پیش‌بینی ممکن ساخته شود. در انجام این کار، الگوریتم های ریاضی مختلفی استفاده می شود تا یک ویژگی و حد آستانه مربوط برای آن ویژگی شناسایی شود، تا مخزن مشاهدات به دو یا چند زیر گروه تقسیم شود. این مرحله تا گره برگ تکرار می شود تا درخت کامل ساخته شود. تعداد شاخه برای هر گره ایجاد شده به الگوریتم خاص استفاده شده و تعداد مقادیر ویژگی انتخاب شده وابسته است. در نهایت باتوجه به امکانات موجود در نرم‌افزارهای Excel و Clementine به پیش‌بینی سودآوری شرکت‌های مذکور پرداخته می شود.

متغیرها

با توجه به تحقیقات پیشین، نسبت ها و متغیرهای مربوطه و تاثیرگذار بر موفقیت شرکت ها و سودآوری یا ورشکستگی آنها در جدول ۲ از کار تحقیقاتی فرج زاده دهکردی (۱۳۸۶) مستخرج شده است.

جدول ۲. متغیرها و نسبت های مورد استفاده در پژوهش های پیشین

نسبت‌ها و متغیرهای مورد استفاده	سال	محقق / محققان
گردش وجوه نقد به کل بدهی‌ها، سودخالص به کل دارایی‌ها، کل بدهی‌ها به کل دارایی‌ها، سرمایه در گردش به کل دارایی‌ها، دارایی‌های جاری به بدهی‌های جاری	۱۹۶۶	بیور
فروش، سودخالص، گردش وجوه نقد، اوراق قرضه قابل معامله، حساب‌های دریافتی، دارایی‌های آنی، موجودی کالا، دارایی‌های جاری، سرمایه در گردش، کل دارایی‌ها، بدهی‌های جاری، جمع بدهی‌ها و سهام ممتاز، ارزش ویژه، وجوه نقد به کل بدهی‌ها، سودخالص به کل دارایی‌ها، کل دارایی‌ها، بدهی‌ها به کل دارایی‌ها، دارایی‌های جاری، آنی به بدهی‌های جاری، وجوه نقد به بدهی‌های جاری، دارایی‌های جاری به فروش، دارایی‌های آنی به فروش، سرمایه در گردش به فروش، وجوه نقد به فروش	۱۹۶۸	بیور
سرمایه در گردش به کل دارایی‌ها، سود (زیان) انباشته به کل دارایی‌ها، سود قبل از بهره و مالیات به کل دارایی‌ها، ارزش بازار حقوق صاحبان سهام به ارزش دفتری بدهی‌ها، فروش به کل دارایی‌ها	۱۹۶۸	آلتمن
الگوی تصادفی از توابعی بر مبنای ارقام صورت‌های مالی	۱۹۷۱	ویلکاکس
جریان وجوه سالانه به بدهی‌های جاری، حقوق صاحبان سهام به فروش، سرمایه در گردش به فروش، بدهی‌های جاری به حقوق صاحبان سهام، موجودی کالا به فروش	۱۹۷۲	ادمیستر
نگاه کنید به بیور ۱۹۶۸	۱۹۷۲	دیکن
گردش وجوه نقد، نقدینگی، اهرم مالی، گردش حقوق صاحبان سهام	۱۹۸۴	منساح
گردش وجوه نقد عملیاتی	۱۹۸۴	کیسی و بارتچک
گردش وجوه نقد عملیاتی، گردش وجوه نقد عملیاتی به بدهی‌های جاری، گردش وجوه نقد عملیاتی به کل بدهی‌ها	۱۹۸۵	کیسی و بارتچک
وجوه حاصل از عملیات، سرمایه در گردش، تامین مالی، هزینه‌های ثابت، مخارج سرمایه‌ای، سود سهام، سایر جریان‌های دارایی‌ها و بدهی‌ها، تغییر در اوراق بهادار نقدی و قابل معامله	۱۹۸۵	گنتری، نیوبلد، ویتنفورد
گردش وجوه نقد عملیاتی، خالص سرمایه‌گذاری ثابت، مالیات پرداختی، تغییرات نقدینگی، بهره پرداختی، بدهی‌های میان‌مدت یا بلندمدت افزایش یافته یا بازپرداخت شده، سود سهام نقدی، حقوق صاحبان سهام افزایش یافته یا بازپرداخت شده	۱۹۸۸	عزیز، امانوئل، لاوسن
گردش وجوه نقد عملیاتی، خالص سرمایه‌گذاری ثابت، مالیات پرداختی، تغییرات نقدینگی، بهره پرداختی، بدهی‌های میان‌مدت یا بلندمدت افزایش یافته یا بازپرداخت شده، سود سهام نقدی، حقوق صاحبان سهام افزایش یافته یا بازپرداخت شده	۱۹۸۹	عزیز و لاوسن
وجوه نقد به کل دارایی‌ها، گردش وجوه نقد عملیاتی به بدهی‌های جاری، گردش وجوه نقد عملیاتی به کل بدهی‌ها، گردش وجوه نقد عملیاتی به کل دارایی‌ها، دارایی‌های جاری به بدهی‌های جاری، دارایی‌های جاری به کل دارایی‌ها، سود قبل از بهره و مالیات به کل	۱۹۹۰	گیلبرت، منون و

اسجوارتز	دارایی‌ها، حقوق صاحبان سهام به کل بدهی‌ها، سودخالص به کل دارایی‌ها، سودانباشته به کل دارایی‌ها، فروش به دارایی‌های جاری، فروش به کل دارایی‌ها، سرمایه در گردش به کل دارایی‌ها	
پلات و پلات	۱۹۹۰	گردش وجوه نقد به کل دارایی‌ها، سودخالص به کل دارایی‌ها، سودخالص به ارزش ویژه، سود عملیاتی به کل دارایی‌ها، گردش وجوه نقد به فروش، سود خالص به فروش، دارایی‌های جاری به کل دارایی‌ها، فروش به کل دارایی‌ها، کل بدهی‌ها به کل دارایی‌ها، ارزش ویژه، فروش به سرمایه در گردش، دارایی‌های جاری به فروش، حساب‌های دریافتی به موجودی کالا، حساب‌های دریافتی به فروش، حساب‌های دریافتی و موجودی کالا به کل دارایی‌ها، دارایی‌های جاری به بدهی‌های جاری، خالص دارایی‌های ثابت به کل دارایی‌ها، وجه نقد به بدهی‌های جاری، وجه نقد به کل دارایی‌ها، درصد تغییرات در فروش، سود قبل از بهره و مالیات به کل هزینه بهره، گردش وجوه نقد به کل هزینه بهره، درصد تغییر در بازده دارایی‌ها، گردش وجوه نقد به بدهی‌ها
فوستر و وارد	۱۹۹۷	گردش وجوه نقد حاصل از فعالیت‌های عملیاتی، گردش وجوه نقد حاصل از فعالیت‌های تامین مالی، گردش وجوه نقد حاصل از سرمایه‌گذاری
سانگ، چانگ و لی	۱۹۹۹	نرخ رشد کل دارایی‌ها، نرخ رشد اموال و ماشین‌آلات و تجهیزات، نرخ رشد دارایی‌های جاری، نرخ رشد فروش، نرخ رشد سودخالص، سودناخالص به فروش خالص، سود عملیاتی به فروش خالص، سود عملیاتی به کل دارایی‌ها، سود عادی به کل دارایی‌ها، سود خالص به کل دارایی‌ها، نسبت سود سهام، سود سهام به سود خالص، سود هر سهم، گردش وجوه نقد هر سهم، حقوق صاحبان سهام به کل دارایی‌ها، نسبت دارایی‌های ثابت به حقوق صاحبان سهام و بدهی‌های بلندمدت، نسبت جاری، نسبت آنی، نسبت بدهی، نسبت پوشش بدهی، گردش وجوه نقد به بدهی‌ها، گردش وجوه نقد به کل دارایی‌ها، نسبت گردش کل دارایی‌ها، نسبت گردش حقوق صاحبان سهام، نسبت گردش موجودی کالا، میانگین دوره ناخلص، بهره وری نیروی کار، سود عادی هریک از کارکنان، کل دارایی‌های هر ریال سرمایه، بهره وری سرمایه، ارزش افزوده ناخالص به فروش خالص
شاه و مورتازا	۲۰۰۰	دارایی‌های جاری به بدهی‌های جاری، فروش به وجه نقد، گردش حساب‌های دریافتی، بهره کسب شده، بدهی به حقوق صاحبان سهام، کل دارایی‌ها به حقوق صاحبان سهام، سود خالص به فروش خالص، فروش خالص به حقوق صاحبان سهام عادی
گریس و اینگرام	2001	نگاه کنید به آلتمن ۱۹۹۶
شین و لی	۲۰۰۲	نسبت سریع، بدهی جاری به کل دارایی، هزینه‌های مالی به فروش، نسبت‌های نقدینگی، درآمد خالص به حقوق صاحبان سهام، درآمد عملیاتی به هزینه عملیاتی، سودانباشته به کل دارایی، حقوق صاحبان سهام به کل دارایی، ارزش افزوده به

کل هزینه		
جریان نقد به دارایی‌های در گردش محدود شده، نسبت سرمایه (سرمایه به کل دارایی‌ها)، مالیات منقضی شده، سود انباشته به کل دارایی‌ها، موجودی کالا، درآمد ناخالص، پوشش بدهی، درآمد خالص، نسبت سریع، نسبت بدهی	۲۰۰۴	سیلن
هزینه بدهی، وضعیت بدهی، رشد، بدهکاری، سهم هزینه نیروی انسانی، نقدینگی کوتاه مدت و گردش دارایی‌ها	۲۰۰۵	اندرس

متغیر وابسته در این پژوهش، سودآوری (زیان آوری) شرکت‌ها در سال آینده است. منظور از این دو واژه مورد تاکید در این تحقیق، موارد زیر است:

۱. سودآوری: شرکتی که خالص سود و زیان پس از کسر مالیات آن در سال مالی آتی مثبت باشد.

۲. زیان آوری: شرکتی که خالص سود و زیان پس از کسر مالیات آن در سال مالی آتی منفی باشد (اعتمادی و همکاران، ۱۳۹۲).

متغیرهای مستقل با توجه به کارهای تحقیقاتی قبلی (متغیرهایی که قابل محاسبه در محیط ایران بوده و برای پیش‌بینی موفقیت شرکت‌ها طی تحقیقات پیشین به کار رفته اند) در پیش‌بینی سودآوری (یا زیان آوری) شرکت‌ها استفاده شده است. فرج زاده دهکردی (۱۳۸۶) پس از بررسی مبسوط تحقیقات پیشین، متغیرهایی را که قابل محاسبه در محیط ایران هستند، برای پیش‌بینی ورشکستگی به بوته آزمایش گذاشت. این پژوهش نیز توانمندی هر یک از ۳۲ متغیر را که طی پژوهش‌های گذشته در پیش‌بینی‌های عملکرد از جمله ورشکستگی و موفقیت شرکت‌ها به کار رفته بودند، بررسی نمود. این متغیرها در جدول شماره ۳ آمده است.

جدول ۳. متغیرهای مستقل مورد استفاده در پژوهش

ردیف	متغیر	ردیف	متغیر
۱	حقوق صاحبان سهام به بدهی بلند مدت	۱۷	بدهی جاری به کل دارایی‌ها
۲	سود عملیاتی به فروش خالص	۱۸	فروش به دارایی‌های جاری
۳	کل بدهی به کل حقوق صاحبان سهام	۱۹	سودخالص به کل دارایی‌ها
۴	کل بدهی به کل دارایی	۲۰	سود خالص به حقوق صاحبان سهام
۵	دارایی‌های جاری به کل دارایی‌ها	۲۱	فروش به حقوق صاحبان سهام
۶	سود عملیاتی به کل دارایی‌ها	۲۲	فروش به کل دارایی‌ها

۷	سود خالص به فروش خالص	۲۳	دارایی های آنی به بدهی جاری
۸	موجودی کالا به کل دارایی ها	۲۴	دارایی ثابت به کل دارایی ها
۹	بدهی به ارزش ویژه	۲۵	فروش به وجه نقد
۱۰	وجه نقد به بدهی جاری	۲۶	سود خالص به دارایی ثابت
۱۱	دارایی های آنی به کل دارایی ها	۲۷	سود و زیان انباشته به کل دارایی ها
۱۲	حساب های دریافتی به موجودی کالا	۲۸	فروش به حساب های دریافتی
۱۳	وجه نقد به کل دارایی ها	۲۹	سود و زیان انباشته به حقوق صاحبان سهام
۱۴	حقوق صاحبان سهام به کل دارایی	۳۰	دارایی جاری به بدهی جاری
۱۵	هزینه بهره به سود ناخالص	۳۱	بدهی جاری به حقوق صاحبان سهام
۱۶	سود ناخالص به فروش	۳۲	بدهی جاری به کل بدهی

جامعه و نمونه آماری

جامعه آماری مورد استفاده در این تحقیق، داده های مربوط به اطلاعات شرکت های پذیرفته شده در بورس اوراق بهادار تهران در قلمرو زمانی ۱۳۸۷ تا ۱۳۹۵ می باشد. جامعه آماری شامل تمام شرکت های پذیرفته شده در بورس اوراق بهادار تهران می باشند که حائز معیارهای زیر باشند:

۱. شرکت هایی که طی سالهای ۸۷ تا ۹۵ عضو بورس اوراق بهادار بوده اند.
 ۲. در تمامی ۹ سال مذکور به بورس اوراق بهادار تهران صورت مالی ارائه کرده باشند.
 ۳. اطلاعات آنها در دسترس باشد.
 ۴. سال مالی آنها منتهی به پایان اسفندماه باشد.
- پس از تشکیل جامعه آماری و تعیین اعضای واجد شرایط، ۱۸۶ شرکت شرایط ذکر شده را داشتند. در تحقیق حاضر به دلیل استفاده از داده کاوی و لزوم داشتن هرچه بیش تر تعداد نمونه ها در این روش، از نمونه گیری استفاده نشد و تمام داده های موجود به شرح فوق به عنوان داده های این تحقیق به کار برده شده اند.

یافته های تحقیق

آماده سازی داده ها یکی از مراحل اصلی است که داده ها را برای مراحل دیگر آماده می کند. بنابراین ابتدا متغیرهای تحقیق از بانک اطلاعاتی بورس اوراق بهادار تهران استخراج شد. روشهای متعددی برای خوشه بندی وجود دارد. در این مطالعه روش K-Means

جهت خوشه‌بندی شرکت‌ها بکار گرفته شد. روش K-Means با وجود سادگی آن یک روش پایه برای بسیاری از روش‌های خوشه‌بندی دیگر است (Alpaydin, 2004). ولی تعیین تعداد خوشه بهینه در ابتدای امر از معایب آن می‌باشد و نتایج به انتخاب اولیه خوشه‌ها وابسته است، بنابراین می‌توان گفت که تعیین تعداد خوشه‌ها از اهمیت زیادی برخوردار بوده و بر نتیجه کار تاثیر خواهد گذاشت. از این رو برای تعیین تعداد خوشه بهینه از روشی به نام SSE استفاده شد. خوشه ای که SSE پایین تر دارد، نشان دهنده بهترین خوشه‌بندی است. برای محاسبه SSE از نرم‌افزار متلب نسخه ۲۰۱۵ استفاده شد که نتایج طبق جدول ۴ ارائه شده است.

جدول ۴. تعداد خوشه‌ها در K-Means

SSE	تعداد خوشه‌ها
۷,۸۹۲۹	۱
۲,۴۵۳۹	۲
۱,۴۳۵۹	۳
۹,۹۱۲۳	۴

با توجه به نتیجه داده های جدول بالا، تعداد خوشه بهینه $K=3$ می باشد و همچنین در شکل ۱ تعداد ثبت های موجود در هر خوشه مشاهده می‌شوند که بیشترین شرکت‌ها در خوشه ۱ قرار گرفته اند.



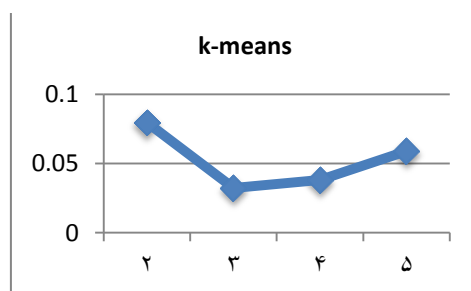
شکل ۱. مقادیر ثبت ها در هر خوشه

پس از محاسبه متغیرها، با استفاده از الگوریتم K-means شرکت‌ها خوشه‌بندی شدند. برای سنجش کیفیت خوشه‌بندی از معیار مذکور برای خوشه‌بندی بین ۲ تا ۵ خوشه محاسبه شده است. اکنون نتایج حاصله بررسی می‌شوند:

معیار: باتوجه به رابطه (۲)، کیفیت نتایج خوشه‌بندی با استفاده از الگوریتم K-mean در جدول (۵) نشان داده شده است. همانطور که پیشتر اشاره شد، کمتر بودن معیار ذکر شده نشان می‌دهد که خوشه‌های تشکیل شده، کیفیت بهتری دارند. همانطور که در نمودار (۱) مشاهده می‌شود، کیفیت خوشه‌های تشکیل شده با استفاده از این الگوریتم مطلوب است و همچنین خوشه‌بندی با سه خوشه بهترین کیفیت را دارد.

جدول ۵. الگوریتم خوشه‌بندی براساس معیار کیفیت

تعداد خوشه	k-means
۲	0/079767389
۳	0/032394225
۴	0/038239037
۵	0/058896799



نمودار ۱. کیفیت خوشه‌بندی الگوریتم k-means با استفاده از معیار کیفیت

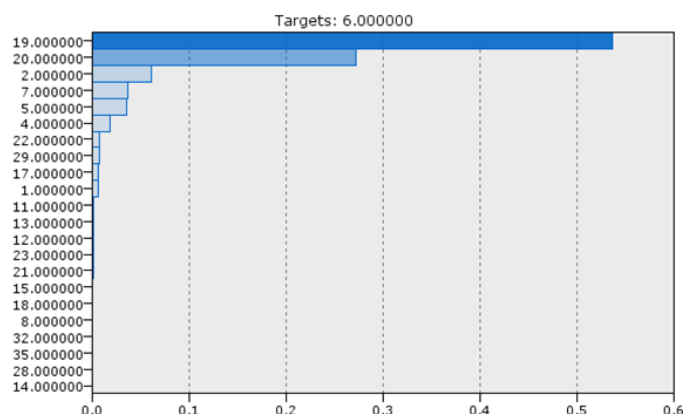
در این مرحله با استفاده از درخت تصمیم به تحلیل بخش‌های تشکیل شده توسط الگوریتم خوشه‌بندی k-means پرداخته شده است، که وضعیت خوشه‌ها به شرح زیر است:

خوشه ۱: شامل شرکت‌هایی می‌باشند که دارای خصوصیات زیر هستند:
نسبت دارایی جاری به کل دارایی آنها قابل ملاحظه و نسبت بدهی جاری به کل دارایی آنها اندک بوده و از طرفی نسبت فروش به کل دارایی آنها زیاد بوده است.
خوشه ۲: شامل شرکت‌هایی می‌باشند که همگی نسبت کل بدهی به کل حقوق صاحبان

سهام آنها در حد متوسط می باشد. خوشه ۳: شامل شرکت‌هایی هستند که نسبت سودخالص به فروش خالص آنها بسیار بالا بوده و همچنین نسبت کل بدهی به کل حقوق صاحبان سهام آنها نیز بسیار ناچیز است، ولی در عین حال نسبت بدهی جاری به کل دارایی‌های آنها بالا است. در نهایت با اجرای درخت تصمیم، طبق شکل ۲ از بین ۳۲ متغیر، ۸ متغیر طبق جدول ۶ به عنوان متغیرهای تاثیرگذار بر سودآوری شرکت‌ها شناخته شدند که بدین شرح هستند:

جدول ۶. متغیرهای تاثیرگذار بر سودآوری

۱. سودخالص به کل دارایی‌ها	۵. دارایی جاری به کل دارایی‌ها
۲. فروش به کل دارایی‌ها	۶. سودخالص به حقوق صاحبان سهام
۳. سود عملیاتی به فروش خالص	۷. سود و زیان انباشته به حقوق صاحبان سهام
۴. سودخالص به فروش خالص	۸. کل بدهی‌ها به کل دارایی‌ها



شکل ۲. متغیرهای تاثیرگذار بر سودآوری

پس از شناسایی متغیرهای تاثیرگذار بر سودآوری شرکت‌ها به عنوان ورودی درخت تصمیم C5 به پیش‌بینی سودآوری آنها در سال ۹۶ در هر خوشه پرداخته شد. اما از آنجایی که یکی از محدودیت‌های درخت تصمیم در نرم‌افزار Clementine این است که نمی‌تواند مستقیماً از داده‌های استخراج شده از پایگاه داده بورس اوراق بهادار تهران استفاده کند، بنابراین نیاز به تبدیل این داده‌ها وجود دارد که براساس طیف لیکرت داده‌ها