7th International Conference on Web Research
May 19-20, 2021; Tehran, IRAN ICWR2021

Presenting a Hybrid Model Based on Data Mining for Startup Failure Detection using Feature Selection and Classification

Shahram Almasi*, Mehrshad Khosraviani

Institute of Higher Education Mehr Alborz, Tehran, Iran
Shahram.alma30@gmail.com, Ite.admin@mehralborz.ac.ir

Abstract

A startup is a newly-founded company that is established by one or more entrepreneurs while presenting a raw idea to produce one or more products. In Iran, the most important reason for startup failure is the lack of sufficient fund. Given the fledgling entrepreneurship and venture capital industry in Iran, as well as the challenges of attracting foreign investment due to international sanctions, the lack of sufficient fund can be considered as the most important reason for failure. Therefore, in the present study, three machine learning methods including decision tree, naive Bayes, and support vector machine were used on a database obtained from Iranian startups in order to predict the success or failure of the startups. First, the weight of effective attribute was calculated by the information gain method and in the classification step, by comparing the accuracy and the amount of area under ROC curve, info gain+svm reported 88.97% accuracy. The results showed that the type of industry, creativity, skill, and innovation have great impact on the success or failure of the startup.

Keywords: Classification, Startup failure, Success prediction, Information gain, Decision tree, Support vector machine, Feature selection.

7th International Conference on Web Research
May 19-20, 2021; Tehran, IRAN ICWR2021

ارائه یک مدل ترکیبی داده‌کاوی جهت بررسی شکست و یا موفقیت استارتاپ‌های ایرانی با انتخاب ویژگی و طبقه‌بندی

شهرام الماسی*[۱]، مهرشاد خسرویانی[۲]

[۱] دانشجو کارشناسی ارشد،گروه مهندسی فناوری اطلاعات، گرایش تجارت الکترونیک، موسسه آموزش عالی مهر البرز،تهران
Shahram.alma30@gmail.com

[۲] استاد یار،گروه مهندسی فناوری اطلاعات، گرایش تجارت الکترونیک، موسسه آموزش عالی مهر البرز،تهران
Ite.admin@mehralborz.ac.ir

چکیده

استارتاپ در حقیقت یک شرکت نوپا است که در جریان ارائه یک ایده خام و برای تولید یک و یا چند محصول، توسط یک یا چند کارآفرین، تأسیس و به بازار عرضه می‌شود. در ایران مهم‌ترین دلیل شکست، نبود سرمایه کافی عنوان‌شده است. با توجه به نوپا بودن صنعت کارآفرینی و سرمایه‌گذاری خطرپذیر در ایران، و هم‌چنین چالش‌های موجود برای ورود سرمایه‌ی خارجی به دلیل تحریم‌های بین‌المللی، انتخاب این گزینه به‌عنوان مهم‌ترین دلیل شکست می‌تواند قابل درک باشد.

در تحقیق پیش رو، بر روی مجموعه داده استارتاپ‌های ایرانی از تکنیک‌های طبقه‌بندی درخت تصمیم، نایو بیز، و ماشین بردار پشتیبان برای پیش‌بینی موفقیت و یا شکست استارتاپ‌ها استفاده شد. ابتدا وزن شاخص‌های تأثیرگذار با روش شاخص بهره اطلاعاتی محاسبه شد و در گام طبقه‌بندی، با مقایسه دقت کل و مقدار سطح زیر نمودار ROC، تکنیک ترکیبی ماشین بردار پشتیبان به همراه انتخاب ویژگی بهره اطلاعاتی بالاترین دقت (۸۸/۹۷ درصد) و کمترین خطا را در مقایسه با سایر روش‌ها گزارش داد. نتایج نشان داد، نوع صنعت ایده، خلاقیت و مهارت افراد، نوآوری، و نوع سرمایه‌گذار تأثیر زیادی در موفقیت و یا شکست استارتاپ‌های ایرانی دارند.

کلمات کلیدی

طبقه‌بندی، موفقیت استارتاپ‌ها، پیش‌بینی موفقیت، شاخص بهره اطلاعاتی، درخت تصمیم، ماشین بردار پشتیبان، انتخاب ویژگی

۱- مقدمه

تعاریف مختلفی برای مفهوم استارتاپ بیان شده است. در ابتدا چند تعریف اولیه بر مفهوم استارتاپ را از نگاه‌های مختلف بیان می‌کنیم.
فرهنگ لغت وبستر، استارتاپ را به صورت «راه‌اندازی یک کسب‌وکار جدید» تعریف کرده است.

مرکز کسب و کارهای کوچک آمریکا(USSBA)، استارتاپ را کسب و کاری معرفی کرد که عموماً ریشه در فناوری داشته و پتانسیل توسعه و رشد بالایی دار د[۱]. به زبان ساده و تجربی، استارتاپ در حقیقت یک شرکت نوپا است که در جریان ارائه یک ایده خام و برای تولید یک و یا چند محصول منحصربه‌فرد، توسط یک یا چند کارآفرین تأسیس و به بازار عرضه می‌شود [۲]. از نظر ماهیت، اکثر استارتاپ‌ها با کمترین هزینه و سرمایه، توسط افراد با تخصص‌های متفاوت ایجاد می‌شوند. چالش اصلی این تیم‌ها، جذب سرمایه‌گذار و سهام‌دار در جریان ارائه محصول جدید به بازار است. بسیاری از

تیم‌های استارتاپی ممکن است در جریان جذب پشتوانه مالی از صحنه رقابت حذف شوند؛ اما این تنها یکی از دلایل شکست تیم‌ها است. دلایل زیادی برای موفقیت و یا عدم موفقیت استارتاپ‌ها وجود دارد که در این تحقیق به آن می‌پردازیم.

یکی از اولین فعالیت‌های اعضای تیم، ارائه استدلال محکم در قالب یک نمونه اولیه است که ادعا کنند محصول و ایده آن‌ها واقعاً جدید است و یا برای بهبود محصولات قبلی، ارزش جدیدی را پیشنهاد می‌دهد [۳]. به همین جهت شرط لازم و کافی یک استارتاپ، تولید یک ایده جدید است که قرار است به بازار ارزش تازه‌ای ارائه دهد. پس در گام اول نبود یک ارزش و ایده ناب و یا کپی‌برداری از محصولات رقیب، از دلایل اصلی شکست یک استارتاپ است.

اما دلایل اصلی و عمده دیگری نیز وجود دارد که ۸۰ درصد استارتاپ‌های شکست‌خورده با آن مواجه شده‌اند. طبق نظرسنجی انجام‌شده توسط CB insights دلایل شکست را می‌توان در ۶ دلیل عمده زیر بیان کرد: ۱- سرمایه ناکافی، ۲- هم‌تیمی‌های ضعیف، ۳- اطلاعات ناکافی از بازار، ۴- مدل کسب‌وکار اشتباه، ۵- رقابت شدید و بازاریابی ضعیف، و ۶- محصول بی‌کیفیت؛ اما نوع جوامع و رشد اقتصادی کشورها، تأثیر بسیار زیادی بر موفقیت و یا عدم موفقیت این تیم‌ها دارد. در ایران مهم‌ترین دلیل شکست تیم‌های استارتاپی، نبود سرمایه کافی عنوان‌شده است. با توجه به نوپا بودن صنعت کارآفرینی و سرمایه‌گذاری خطرپذیر در ایران، تعداد کم شرکت‌ها و صندوق‌های سرمایه‌گذاری خطرپذیر، و هم‌چنین چالش‌های موجود برای ورود سرمایه‌ی خارجی به دلیل تحریم‌های بین‌المللی، این گزینه می‌تواند مهم‌ترین دلیل شکست استارتاپ‌ها باشد.

ازاین‌رو در این تحقیق، با استفاده از تکنیک‌های انتخاب ویژگی و طبقه‌بندی داده‌کاوی، به بررسی موفقیت و یا عدم موفقیت تیم‌های استارتاپی در ایران می‌پردازیم. در حال حاضر، داده‌کاوی مهم‌ترین فناوری جهت بهره‌برداری مؤثر از داده‌های حجیم است و اهمیت آن رو به فزونی است [۴]. بر اساس تخمین‌های انجام‌شده، مقدار داده‌ها در جهان هر بیست ماه تقریباً دو برابر می‌شود. هر چه حجم داده‌ها بیشتر و روابط میان آن‌ها پیچیده‌تر باشد، دسترسی به اطلاعات نهفته در میان داده‌ها مشکل‌تر می‌شود و نقش داده‌کاوی، به‌عنوان یکی از روش‌های کشف دانش، روشن‌تر می‌گردد.

در بخش دوم این مقاله، تحقیقات انجام‌شده در این حوزه را مرور می‌کنیم. در بخش سوم، مفاهیم و روش‌های کارشده تعریف می‌شوند. در بخش چهارم، روش تحقیق را بر اساس فرآیند پروژه انجام شده تعریف می‌کنیم و مراحل درک داده، انتخاب فاکتورها و طبقه‌بندی اجراشده در نرم‌افزار را گزارش می‌دهیم. در بخش پنجم نیز، بر اساس پارامترهای ارزیابی، نتایج را ارزیابی و تحلیل می‌کنیم.

۲- مقالات کارشده

مقالات معتبر محدودی با استفاده از روش‌های داده‌کاوی به بررسی دلایل شکست استارتاپ‌ها پرداخته‌اند.

محققان کره جنوبی برای پیش‌بینی بودجه و مسائل مالی استارتاپ‌های این کشور از یک سیستم اعتباردهی امتیاز مبتنی بر درخت تصمیم‌گیری استفاده کردند[۵]. آن‌ها با استفاده از ویژگی‌های خاصه شرکت‌ها و شاخص‌های اقتصادی کشور، مدلی را با دقت ۷۴ درصد پیشنهاد دادند که در مقایسه با شبکه عصبی و رگرسیون دقت بیشتری داشت.

مقالات زیادی از تکنیک‌های درختی جهت پیش‌بینی و ارزیابی سیستم تصمیم‌گیری استفاده کرده‌اند. محققان گزارش داده‌اند که به دلایل مختلف از جمله کمبود بودجه و نبود سرمایه‌گذار مناسب، از هر ۱۰ استارتاپ، ۹ تیم دچار شکست می‌شوند [۶]. آن‌ها پس از پیش‌پردازش، از تکنیک‌های داده‌کاوی درخت تصمیم متناوب (ADTree) و جنگل تصادفی استفاده و دقت کل و مقدار سطح زیر نمودار ROC را با شبکه بیزین مقایسه کردند. نتایج آن‌ها نشان داد که یک استارتاپ برای بررسی عوامل موفقیتش می‌تواند بر روی چه فاکتورهای مهمی تمرکز کند.

محققان به توسعه محصولات مرتبط در تیم‌های استارتاپی نیز پرداخته‌اند [۲]. تمرکز آن‌ها بر این اصل بود که چرا ۹۰ درصد استارتاپ‌ها با شکست مواجه می‌شوند. روش آن‌ها بر پایه تحلیل رفتار مشتری و تحلیل نقشه راه مشتری است. آن‌ها از تکنیک‌های داده‌کاوی برای تجزیه و تحلیل استفاده کردند. نتایج آن‌ها با زنجیره مارکوف و الگوریتم‌های فرا ابتکاری به پیشنهاداتی برای بهبود وفاداری مشتری و توسعه محصول منجر شد.

عنصر اولیه موفقیت استارتاپ‌ها، توانایی تیم‌ها در تأمین بودجه و یافتن سرمایه‌گذار اولیه است. محققان اطلاعات کارآفرینان، ارتباط کارآفرینان و سرمایه گذاران و سطح تعامل اجتماعی را به مدت ۷ تا ۱۰ ماه از طریق شبکه‌های اجتماعی مختلف جمع‌آوری کردند [۷]. تجزیه و تحلیل آن‌ها نشان داد که مشارکت فعال در شبکه‌های اجتماعی ارتباط زیادی با موفقیت در تأمین مالی جمعی دارد. در برخی موارد، سطح تعامل اجتماعی برای شرکت‌های موفق به‌وضوح بالاتر است. آن‌ها همچنین از الگوریتم k-نزدیکترین همسایه و درخت تصمیم برای پیش‌بینی توانایی شرکت در تأمین مالی تیم استفاده کرده و مدل پیش‌بینی را با دقت بالای ۸۴ درصد ارائه دادند.

رشد نمایی و فراگیر شدن اینترنت کارآفرینان تازه وارد را به سمت راه‌های جدیدی برای جذب حمایت مالی و تأمین بودجه استارتاپ خود سوق داده است. علاوه بر بودجه، سه عامل خلاقیت، پشتکار و تجربه افراد تیم نیز به‌عنوان عوامل اساسی نام برده شده‌اند و جهت پیش‌بینی عامل شکست روی نتایج شرکت‌ها، از یادگیری عمیق و تکنیک‌های تجمعی روی ویژگی‌های متنی و عددی استفاده شده است [۸]. مدل پیشنهادی با دقت ۹۳ درصد برای پیش‌بینی شکست تیم استفاده شد.

محققان توسعه مدل پیش‌بینی موفقیت یا شکست شرکت‌های نوپای شتاب‌دهی‌شده را به منظورکاهش عدم قطعیت در انتخاب شرکت‌های موفق انجام دادند [۹]. محققان ایرانی با توسعه و بسط مدل پیش‌بینی موفقیت و شکست لوزیر و ادغام آن با مدل تجربی بیلگراس، مدل جدیدی پیشنهاد دادند. آن‌ها از انواع مدل‌های انتخاب ویژگی فرا ابتکاری و روش‌های تجمعی

آدابوست و استکینگ استفاده کردند و به روش استکینگ با دقت ۸۹ درصد دست یافتند.

۳- تعاریف

در این بخش، سه روش طبقه‌بندی درخت تصمیم، بیز ساده، و ماشین بردار پشتیبان معرفی می‌شود. روش‌های طبقه‌بندی در داده‌کاوی بر اساس متغیر هدف، داده‌ها را به دو یا چند کلاس طبقه‌بندی کرده و یک مدل یادگیری جهت پیش‌بینی نتایج ساخته می‌شود.

۳-۱- درخت تصمیم C4.5

درخت ID3 نوعی درخت تصمیم بالا به پائین است. ابتدا بر اساس میزان بهره اطلاعات نمونه‌های موجود در مجموعه داده آزمایش، درخت تصمیم می‌گیرد که کدام ویژگی در ریشه قرار داشته باشد. مقادیر حاصل از میزان بهره اطلاعات[۱] به ازای هر ویژگی نشان‌دهنده توانایی آن ویژگی در دسته‌بندی نمونه‌های آزمایشی است. با انتخاب این ویژگی، برای هر یک از مقادیر ممکن آن، یک شاخه ایجاد شده و مثال‌های آموزشی بر اساس ویژگی هر شاخه مرتب می‌شوند. سپس عملیات فوق برای مثال‌های قرارگرفته در هر شاخه تکرار می‌شوند تا بهترین ویژگی برای گره بعدی انتخاب شود. این الگوریتم یک جستجوی حریصانه است که در آن انتخاب‌های قبلی هرگز مورد بررسی قرار نمی‌گیرد. الگوریتم درختID3، هر انشعاب از درخت را تا زمانی که تمام نمونه‌ها را دسته‌بندی نماید، ادامه می‌دهد که منجر به عمق زیاد درخت می‌شود. بنابراین می‌تواند عامل بروز بیش‌برازش[۲] شود. دلایل اصلی بیش‌برازش، وجود نویز در داده‌های آموزشی و تعداد کم مثال‌های آموزشی است. ID3 و C4.5 هر دو از معیار میزان خلوص در تابع جداساز استفاده می‌کنند [۱۰].

درخت C4.5 الگوریتم درخت ID3 است که از معیار نسبت بهره[۳] جهت انتخاب ویژگی استفاده می‌کند. این الگوریتم زمانی متوقف می‌شود که تعداد نمونه‌ها کمتر از مقداری مشخص باشد. در درخت C4.5، پس از ساخت درخت، از روش هرس کردن استفاده می‌شود. در این مدل از داده‌های عددی نیز می‌توان استفاده نمود.

۳-۲- بیز ساده

در آمار، طبقه‌بند نایو بیز خانواده‌ای از طبقه‌بندهای ساده «احتمال پذیر» است که مبتنی بر کاربرد قضیه بیز با پیش‌فرض استقلال بین ویژگی‌ها تعریف می‌شود. این تکنیک‌ها یکی از ساده‌ترین مدل‌های شبکه بیزی هستند. اما می‌توان روش را با برآورد تراکم هسته همراه کرد و به سطح دقت بالاتری رسید. این روش بسیار مقیاس‌پذیر و به تعدادی از پارامترهای خطی مستقل و پیش‌بینی‌گر در یک مشکل یادگیری نیاز دارد [۱۱].

۳-۳- ماشین بردار پشتیبان

استفاده از بردارهای پشتیبان خطی در مسائل طبقه‌بندی، رویکرد جدیدی است که در چند سال اخیر مورد توجه بسیاری قرارگرفته است. روش به این صورت است که در مرحله آموزش، سعی می‌شود مرز تصمیم‌گیری به‌گونه‌ای انتخاب شود که حداقل فاصله آن با هریک از دسته‌های موردنظر بیشینه باشد. این بیشینه‌سازی با استفاده از یک الگوریتم بهینه‌سازی و نمونه‌هایی که مرز کلاس‌ها را تشکیل می‌دهند، بدست می‌آید. به این نمونه‌ها بردارهای پشتیبان[۴] می‌گویند. تعدادی از نقاط آموزشی که کمترین فاصله با مرز تصمیم‌گیری دارند را می‌توان به عنوان این بردارها در نظر گرفت. یکی از مزایای ماشین بردار پشتیبان (SVM) نسبت به سایر تکنیک‌های طبقه‌بندی این است که تنها بردارهای جداکننده را در نظر می‌گیرد (یعنی داده‌های روی مرز) و بنابراین می‌تواند تعمیم بهتری در مقایسه با سایر تکنیک‌ها مانند رگرسیون لجستیک ارائه دهد. این ویژگی باعث می‌شود که SVM کمتر دچار بیش‌برازش یا به‌خاطرسپاری شود. فضای ویژگی ورودی از دو کلاس تشکیل‌شده و کلاس‌ها در مجموع دارای x_i نقطه آموزشی می‌باشند که $i=1,2,\ldots,N$تعریف می‌شود. برای محاسبه مرز تصمیم‌گیری دو کلاس کاملاً جدا از هم، از روش حاشیه بهینه استفاده می‌شود [۱۰].

٤- روش تحقیق

در این پژوهش، به منظور تحلیل داده و بررسی شکست و یا موفقیت تیم‌های استارتاپی در ایران، از روند اجرای پروژه‌های داده‌کاوی (فرآیند کریسپ) بهره خواهیم برد. این فرآیند شامل مراحل زیر می‌شود:

- استخراج داده
- استخراج اطلاعات آماری
- کشف نیازمندی‌های پروژه و انتخاب مدل
- مراحل اجرایی شامل پیش‌پردازش و پاک‌سازی داده‌ها در صورت نیاز
- تحلیل ویژگی‌های تأثیرگذار
- طبقه‌بندی
- ارزیابی و اعمال

در ادامه، بر اساس گام‌های تعریف‌شده، مراحل پروژه را به تفصیل بیان می‌کنیم.

٤-۱- شناخت داده‌ها

مجموعه داده شامل اطلاعات ۱۶۵ استارتاپ ایرانی با ۳۹ متغیر ورودی و ۱ متغیر پیش‌بینی برای موفقیت و یا عدم موفقیت استارتاپ‌ها می‌شود. از این میان، ۱۳۱ استارتاپ شکست‌خورده و تنها ۳۳ استارتاپ موفق وجود دارد. ۳۹

ویژگی ورودی، شامل سؤالاتی از نحوه ایجاد تیم، تعداد اعضا و و سؤالات فنی، که از مدیران تمام استارتاپ‌های ایرانی از طریق پرسشنامه تهیه شده است، در شکل ۱ نشان داده شده است. اطلاعات زیر از طریق بررسی نمودارهای آماری فراوانی و پراکندگی داده‌ها به‌دست آمده است:

- ✓ در ۴۷ درصد استارتاپ‌ها تعداد اعضا بین ۲ تا ۵ نفر، ۳۴ درصد بیش از ۵ نفر، و تنها در ۱۸درصد (۲۸ شرکت) تعداد اعضا ۲ نفر است.
- ✓ در ۵۳ درصد تیم‌ها تحصیلات مدیر، لیسانس و ۳۸ درصد فوق‌لیسانس است. تنها ۱ شرکت داریم که تحصیلات مدیر فوق‌دیپلم و حتی دیپلم تعریف شده است.

- بیشتر اعضای شرکت‌ها (۵۵ درصد) جوان و میانگین سنی ۲۵ تا ۳۰ سال دارند.
- ۸۷ درصد استارتاپ‌ها مدیر مرد و مابقی مدیر زن دارند و تنها ۱ تیم، مدیر مشترک زن و مرد داردکه جزو استارتاپ‌های شکست‌خورده است.
- شرکت‌های بررسی‌شده بسیار نوپا هستند، به طوری که ۶۸ درصد شرکت‌ها بین ۱ تا ۳ سال کار را شروع کرده، ۱۴ درصد کمتر از یک سال و ۱۳ درصد بین ۳ تا ۵ سال شروع به فعالیت نموده‌اند و ۰/۰۲ درصد، یعنی تنها ۴ شرکت، در مجموع بیش از ۵ سال سابقه دارند؛ شرکت‌های با سابقه‌ی بیشتر جزو استارتاپ‌های موفق هستند.
- در ۴۴ درصد از استارتاپ‌ها افراد پیش از شروع استارتاپ با یکدیگر همکار بوده و به تجربه هم اعتماد داشته‌اند. در ۳۳ درصد تیم‌ها افراد شناخت قبلی از هم نداشته اما به مدت یک سال قبل از شروع کار، برای شناخت یکدیگر روی کار متمرکز شده‌اند که این استارتاپ‌ها کمتر شکست‌خورده‌اند.
- ۹۴ درصد استارتاپ‌ها از نوع سرویس و مابقی محصول هستند که استارتاپ‌های نوع محصول شکست‌خورده نبودند.
- تنها در ۵ تیم رقابتی در بازار محصول وجود ندارد که بیشتر استارتاپ‌های شکست‌خورده از این دسته هستند، و مابقی در حال رقابت با استارتاپ‌های رقیب هستند، بطوریکه ۱۷تیم رقبا را بسیار موفق‌تر از استارتاپ خود ارزیابی کردند. ۵۰ تیم (۳۹ درصد) نیز سطح رقابتی خود را خوب ارزیابی کردند.
- ۵۶ درصد تیم‌ها برای بازاریابی استارتاپ خود فعالیت خاصی نداشتند؛ تیم‌هایی که بازاریابی انجام دادند، موفق‌تر گزارش شدند.
- ۵۸ درصد استارتاپ‌ها کاملاً یک ایده نو را در ابتدای راه ارائه و مابقی از طرح‌های دیگر استفاده و وارد بازار رقابتی شدند.
- ۷۹ درصد از تیم‌ها عمدتاً از زمان شروع بازخورد مثبتی داشته‌اند و تنها ۱ تیم بازخورد منفی گزارش داده است.

٤-۲- انتخاب فاکتورهای تأثیرگذار

به‌منظور بررسی عملکرد هر یک از ویژگی‌ها و میزان تأثیرگذاری آن‌ها روی موفقیت و یا عدم موفقیت تیم‌ها، از یکی از روش‌های شاخص وزن‌دهی داده‌کاوی و شاخص بهره اطلاعاتی با فرمول محاسبه آنتروپی در نرم‌افزار استفاده کردیم.

در این روش، عملگر «وزن‌دهی با شاخص بهره اطلاعاتی»[۵] رابطه بین متغیرها را بر اساس آنتروپی یا بهره اطلاعاتی روی مجموعه داده‌ها محاسبه می‌کند به‌گونه‌ای که وزن هر متغیر را با محاسبه به خروجی نرم‌افزار می‌دهد. در واقع، این روشی برای محاسبه ناخالصی متغیر است. متغیر نامتوازن متغیری است که کمترین میزان ناخالصی را داشته باشد. و گره متوازن بیشترین ناخالصی را دارد و بهتر می‌تواند تفکیک را بین دو کلاس در بخش دسته‌بندی داشته باشد. به عبارت دیگر، بیشترین بهره اطلاعاتی را دارد [۱۰].

فرمول محاسبه بهره اطلاعاتی بر پایه آنتروپی شانون طبق فرمول(۱) تعریف می‌شود:

$$\mathrm{Gain(A)} = \mathrm{Entropy(D)} - \mathrm{Entropy_A(D)} \qquad (۱)$$

که در آن A ویژگی و D بر مجموعه داده‌های آموزشی دلالت دارد. فرمول آنتروپی شانون طبق فرمول (۲) تعریف می‌شود که در آن c تعداد برچسب کلاس‌های موجود در داده‌های آموزشی، p_i احتمال اینکه نمونه‌ای از داده‌ها متعلق به کلاس iام باشد، v تعداد اعضای دامنه صفت خاصه A، و D_j قسمتی از داده‌های اولیه است. همچنین، |D| اندازه مجموعه داده است.

$$Entropy(D) = -\sum_{i=1}^{c} p_i \times \log(p_i)\ Entropy_A(D) = \sum_{j=1}^{v} \frac{|Dj|}{|D|} \times Entropy(D) \qquad (۲)$$

ابتدا نتایج وزن شاخص‌ها را گزارش می‌دهیم. وزن‌ها اعدادی بین صفر و یک را گزارش می‌کنند که هر چه به ۱ نزدیک‌تر باشد تأثیر بیشتری در موفقیت و یا عدم موفقیت تیم‌ها دارد. شکل(۱) وزن شاخص‌ها(متغیرها) را به ازای اجرای روش شاخص آنتروپی گزارش می‌دهد. نتایج نشان می‌دهند که:

- تأثیرگذارترین ویژگی با وزن ۱ که به‌طور مستقیم در موفقیت و یا عدم موفقیت استارتاپ‌ها تأثیر داشته، نوع صنعت ایده (خدمات پزشکی، مدیریتی، حمل‌ونقل و ...) است.
- سایر فاکتورهای تأثیرگذار عبارتند از زمان شروع به کار، مرحله رشد، انگیزه افراد، مقایسه محصولات با رقبا از نظر نوآوری ایده، سرمایه‌گذار، و خلاقیت افراد.
- سال تأسیس در شکست استارتاپ‌ها تأثیر زیادی داشته است. این ادعا با توجه به تغییرات شرایط اقتصادی در ایران قابل‌قبول است.

- بیشتر فاکتورهای تأثیرگذار عبارتند از نوع ایده ، مهارت مدیر، و تجربه افراد. همچنین، سن، تعداد اعضا، و جنسیت مدیر و افراد تأثیری در شکست و یا موفقیت ندارد.
- با توجه به نتایج وزن‌ها، سرمایه و نوع سرمایه‌گذار در چرخه عمر محصول نیز تأثیر زیادی دارد.

٤-٣- مدل‌سازی (طبقه‌بندی)

در این پژوهش، در گام طبقه‌بندی از تکنیک‌های درخت تصمیم، بیز ساده و ماشین بردار پشتیبان برای پیش‌بینی موفقیت و یا عدم موفقیت تیم‌ها استفاده شد.

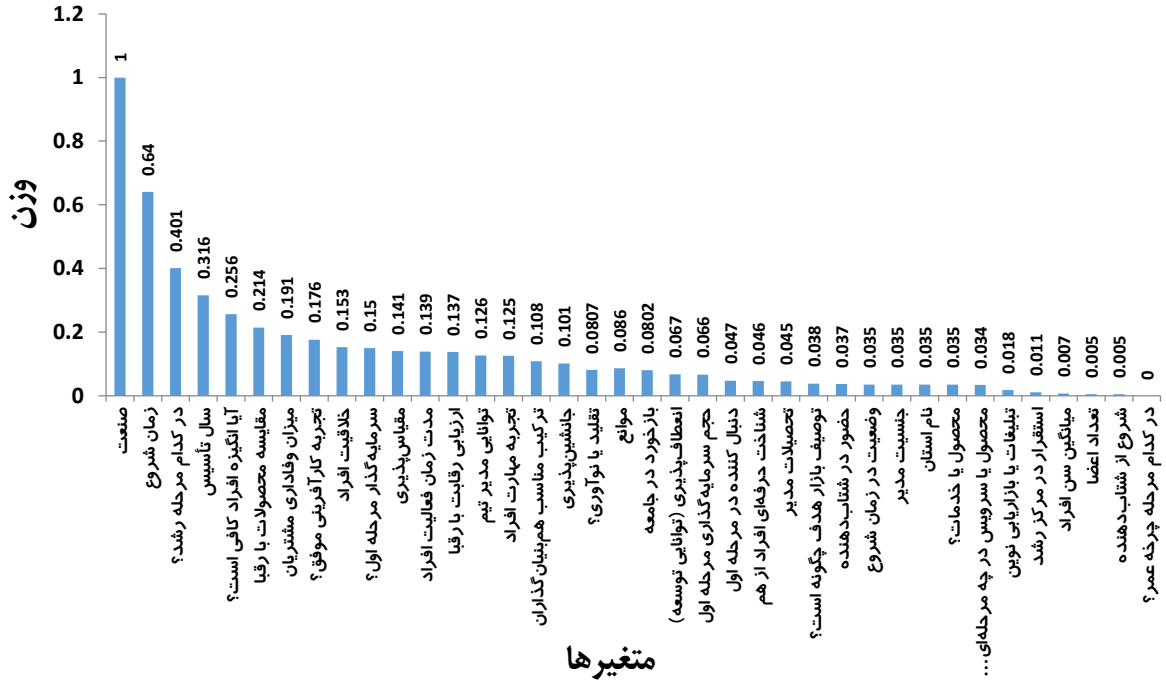
جهت اعتبارسنجی و تقسیم داده‌ها به دو قسمت آموزشی و آزمایشی از روش هوشمندانه اعتبارسنجی ضربدری به‌جای اعتبارسنجی ساده استفاده کردیم. در این نوع اعتبارسنجی داده‌ها به K زیرمجموعه افراز می‌شوند؛ از این K زیرمجموعه، هر بار یکی برای اعتبارسنجی و K-1تای دیگر برای آموزش بکار می‌روند. این روال Kبار تکرار می‌شود و همه داده‌ها دقیقاً یک بار برای آموزش و یک بار برای اعتبارسنجی بکار می‌روند. در نهایت، میانگین نتیجه این Kبار اعتبارسنجی به‌عنوان یک تخمین نهایی برگزیده می‌شود. پس از مراحل پیش‌پردازش مجموعه داده‌ها، طبقه بندهای مذکور در نرم‌افزار رپیدماینر روی داده‌های آموزش اعمال و پس از تست، دقت و ماتریس درهم‌ریختگی گزارش شد.

قبل از آن که به بررسی انواع مهم معیارهای طبقه‌بندی بپردازیم، لازم است که مفهوم ماتریس درهم‌ریختگی را، که در این بخش از آن استفاده خواهیم کرد، تشریح کنیم. این ماتریس چگونگی عملکرد الگوریتم طبقه‌بندی را با توجه به مجموعه داده ورودی به تفکیک انواع کلاس مساله طبقه‌بندی نشان می‌دهد. مفاهیم مثبت درست[٦]، مثبت غلط[٧]، منفی درست[٨] و منفی غلط[٩] در جدول (١) به شرح ذیل هستند.

- مثبت درست: این مقدار بیانگر تعداد رکوردهایی است که طبقه واقعی آن‌ها مثبت بوده و دسته‌بند نیز طبقه آن‌ها را به‌درستی مثبت تشخیص داده است.
- مثبت غلط: این مقدار بیانگر تعداد رکوردهایی است که طبقه واقعی آن‌ها منفی بوده و دسته‌بند طبقه آن‌ها را به نادرستی مثبت تشخیص داده است.

جدول ١: ماتریس اغتشاش

		نمونه‌های پیش‌بینی‌شده مدل	
نمونه‌های واقعی		طبقه منفی (−)	طبقه مثبت (+)
	طبقه منفی (−)	TN	FP
	طبقه مثبت (+)	FN	TP



شکل ١. وزن تأثیر فاکتورها با شاخص بهره اطلاعاتی

- منفی درست: این مقدار بیانگر تعداد رکوردهایی است که طبقه واقعی آن‌ها منفی بوده و دسته‌بند نیز طبقه آن‌ها را به‌درستی منفی تشخیص داده است.
- منفی غلط: این مقدار بیانگر تعداد رکوردهایی است که طبقه واقعی آن‌ها مثبت بوده و دسته‌بند طبقه آن‌ها را به نادرستی منفی تشخیص داده است.

در ادامه معیارهای ارزیابی را گزارش خواهیم داد. اولین معیار، دقت یا نرخ دسته‌بندی است و به این مفهوم است که دسته‌بند، چند درصد نمونه‌ها را به‌درستی چه در کلاس مثبت چه در کلاس منفی طبقه‌بندی کرده است. این دقت بر اساس مفاهیم ماتریس اغتشاش از فرمول(۳) تعیین می‌شود:

$$Accuracy = \frac{Tp+TN}{Tp+TN+Fp+FN} \qquad (۳)$$

معیار مهم دیگر که برای تعیین کارایی یک طبقه‌بند مخصوصاً در داده‌های نامتوازن بسیار مؤثر است، معیار سطح زیر نمودار ROC[۱۰] تحلیل ویژگی عامل‌هاست. منحنی ROC، منحنی‌های دوبعدی هستند که در آن‌ها محور افقی نرخ تشخیص صحیح رده مثبت و محور عمودی نرخ تشخیص غلط رده منفی[۱۱] است که با فرمول(۴) مشخص می‌شود.

$$FAR = \frac{FP}{TN+FP} \qquad (۴)$$

۵- نتایج و بحث

سه تکنیک مذکور را یک بار با انتخاب ویژگی و یک بار بدون انتخاب ویژگی روی داده با ۱۶۵ نمونه پیاده‌سازی کرده و نتایج دقت کل و مقدار سطح زیر نمودار مطابق جدول (۲) گزارش شدند.

- بالاترین دقت طبقه‌بندی با روش ماشین بردار با انتخاب ویژگی بهره اطلاعاتی با دقت کل ۸۸/۹۷ درصد و سطح زیر نمودار ۰/۹۱۳ بدست آمد.
- در تمام روش‌ها، دقت هر سه کلاس با انتخاب ویژگی (انتخاب ۱۳ متغیر از ۴۱ متغیر) افزایش داشته است.
- در مقدار سطح زیر نمودار، محور عمودی نرخ درست کلاس مثبت (شکست) و محور افقی نرخ منفی کلاس موفقیت است. در سیستم‌های تشخیص شکست تیم، هزینه خطای شکست (کلاس مثبت) بیشتر از موفقیت است و تشخیص درست کلاس مثبت اهمیت بیشتری دارد. در نمودار ROC هرچه دهانه نمودار به سمت محور عمودی باشد، نرخ تشخیص کلاس مثبت بیشتر و مقدار سطح زیر نمودار بیشتر است که Svm+info gain دقت بهتری نسبت به سایر روش‌ها دارد.

جدول ۲: نتایج دقت طبقه‌بندها

مدل	دقت کل (درصد)	مقدار سطح زیر نمودار (AUC)
C4.5	۸۳/۴۹	۰/۸۰۳
C4.5+info gain	۸۳/۴۹	۰/۸۱۴
NB	۸۷/۲۴	۰/۸۶۷
NB+info gain	۸۷/۷۳	۰/۸۹۶
svm	۸۶/۵۱	۰/۹۱۱
Svm+info gain	۸۸/۹۷	۰/۹۱۳

- تکنیک درخت تصمیم با دقت ۸۳/۴۹ درصد نتایج خوبی حاصل کرد. دلیل انتخاب این روش، که از تکنیک‌های جعبه سفید است، تولید قوانین تحلیلی است.

۶- نتیجه‌گیری

در این پژوهش، جهت ارزیابی موفقیت و یا عدم موفقیت استارتاپ‌ها، از مجموعه داده شامل ۱۶۵ استارتاپ ایرانی با ۳۹ متغیر ورودی و ۱ متغیر پیش‌بینی و تکنیک‌های طبقه‌بندی استفاده شد. به دلیل ابعاد بالا، روش آنتروپی بهره اطلاعاتی جهت انتخاب ویژگی روی داده‌ها اعمال و وزن تأثیرگذاری هر یک از فاکتورها روی شکست و یا موفقیت تیم‌ها گزارش شد. نتایج نشان داد نوع ایده، مهارت مدیر و تجربه افراد از عوامل تأثیرگذار است و سن و تعداد اعضا و جنسیت مدیر و افراد تأثیری در شکست و یا موفقیت ندارد.

در گام طبقه‌بندی، داده‌ها به دو دسته آموزشی و تست تقسیم شدند و سه تکنیک درخت تصمیم، بیز ساده، و ماشین بردار پشتیبان روی داده‌ها اعمال شد. بر اساس دقت کل و نتایج ماتریس درهم‌ریختگی، مدل ماشین بردار پشتیبان+شاخص بهره اطلاعاتی (انتخاب ویژگی با ۳۰ فاکتور) دقت بالایی (۸۸/۹۷ درصد) را گزارش داد.

تفاوت زیاد دقت در حالت بدون انتخاب و انتخاب ویژگی، لزوم استفاده از روش‌های شاخص وزن‌دهی را اثبات کرد. جهت توسعه داده با توجه به ابعاد بالا، سایر تکنیک‌های تشخیص فاکتورهای تأثیرگذار مانند روش‌های فرا ابتکاری پیشنهاد می‌شود.

مراجع

[1] [1] U.S. Small Business Administration (USSBA), U.S.B.C., https://www.sba.gov/.
[2] [2] Morozov, V., *Product Development of Start-up Through Modeling of Customer Interaction Based on Data Mining* in *International Conference on Data Stream Mining and Processing*, Springer, 2020.
[3] [3] Sharchilev, B., Roizner, M., *Web-based startup success prediction*, Proceedings of the 27th ACM

International Conference on Information and Knowledge Management, Torino, Italy, October 2018.

[4] [4] Tripathi, D., *"Evolutionary Extreme Learning Machine with novel activation function for credit scoring"*, Engineering Applications of Artificial Intelligence, vol. 96, pp.103980, 2020.

[5] [5] Sohn, S.Y., Kim, J.W., *"Decision tree-based technology credit scoring for start-up firms: Korean case"*, Expert Systems with Applications, vol. 39, no. 4, pp. 4012-4007, 2012.

[6] [6] Krishna, A., Agrawal, A., Choudhary A., *Predicting the outcome of startups: Less failure, more success*, 2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW), Barcelona, Spain, December, 2016.

[7] [7] Zhang, Q., *Predicting startup crowdfunding success through longitudinal social engagement analysis*, Proceedings of the 2017 ACM Conference on Information and Knowledge Management, Singapore, November, 2017.

[8] [8] Srinivasan, A., *"An Ensemble Deep Learning Approach to Explore the Impact of Enticement"*, Engagement and Experience in Reward Based Crowdfunding, papers.ssrn.com, 2020.

[9] [9] Sadatrasoul, S.M., Ebadati, O., Saedi, R., " *A Hybrid Business Success Versus Failure Classification Prediction Model: A Case of Iranian Accelerated Start-ups"*, Journal of AI and Data Mining, vol. 8, no. 2, pp. 287-279, 2020.

[10] [10] Han, J., Pei, J., Kamber, M., *Data mining: concepts and techniques*, Elsevier, 2011..

[11] [11] Mirzakhanov, V.E., *"Value of fuzzy logic for data mining and machine learning: A case study"*, Expert Systems with Applications, vol. 162, pp. 113781, 2020.

زیرنویس‌ها

[1] information gain
[2] overfitting
[3] Gain ratio
[4] Support Vector
[5] Weight by information gain
[6] True positive-TP
[7] False Positive-FP
[8] True Negative-TN
[9] False Negative-FN
[10] Area Under Curve
[11] False Positive Rate(FPR)