Industrial Management Perspective

Print ISSN: 2251-9874, Online ISSN: 2645-4165
Volume 13, Serial Number 49, Spring 2023, Pages 281-322 (Original Article)
DOI: 10.52547/JIMP.13.1.281

Prediction Model of the Gas Pipeline Critical Risk Using Data Mining Algorithms

Mohammad Sadegh Behrouz[*], Mohammad Ali AfsharKazemi[**], Adel Azar[***], Ezatolah Asgharizadeh[****]

Abstract

Predictive approaches play an important role in detecting events, controlling risks and reducing maintenance and repair costs. The purpose is to provide a model for predicting critical and prioritized risks based on data mining algorithms. Data mining method was planned based on the CRISP methodology. Data modeling has been done in two parts: "descriptive" and "predictive" data mining and the use of "clustering" and "classification" algorithms."Sillhouette index" is considered for clustering and the K-Means, Kohnen, Two Step algorithm is used; the best value is based on the K-Means algorithm. Silhouette is equal to 0.6446 with the number of clusters 5. Next, Neural Network Algorithms, C.5 tree, Nearest Neighbor and Support Vector have been used for classification. These techniques recognizing data classification patterns and their integration increases the amount of data learning. The results showed learning in 97.56% of the agreed data and the accuracy and validity of the combined model for data classification was estimated at 92.86%. Based on the results, 13 critical risks have been identified; "release of polluting gases and chemicals" and "lack of training and justification of contractors regarding the pipeline" have the highest and lowest priority, respectively.

Keywords: Risk Assessment; Maintenance; Modeling; Data Mining; Gas Pipeline.

Received: Jul. 06, 2022; Accepted: Sep. 13, 2022.

* Ph.D. Student of Industrial Management, Faculty of Management and Economics, Branch of Science and Research, Islamic Azad University, Tehran, Iran.
** Associate Professor, Faculty of Management and Economics, Branch of Science and Research, Islamic Azad University, Tehran, Iran. (Corresponding Author).
Email: dr.mafshar@gmail.com
*** Professor, Management and Accounting Faculty, Tarbiat Modares University, Tehran, Iran.
**** Associate professor, Faculty of management, Tehran University, Tehran, Iran.

چشم‌انداز مدیریت صنعتی

شاپای چاپی: ۹۸۷۴-۲۲۵۱، شاپای الکترونیکی: ۴۱۶۵-۲۶۴۵

سال سیزدهم، شماره ۴۹، بهار ۱۴۰۲، ص‌ص ۲۸۱ – ۳۲۲ (**نوع مقاله: پژوهشی**)

DOI: 10.52547/JIMP.13.1.281

ارائه مدل پیش‌بینی ریسک‌های بحرانی شبکه انتقال گاز با استفاده از الگوریتم‌های داده‌کاوی

محمدصادق بهروز*، محمدعلی افشارکاظمی، عادل آذر***، عزت‌اله اصغری‌زاده******

چکیده

باتوجه به نقش مهم رویکردهای پیش‌بینانه در کاهش هزینه‌های نگهداری تعمیرات، هدف از انجام پژوهش، ارائه مدل پیش‌بینی ریسک‌های بحرانی و اولویت‌دار بر پایه الگوریتم‌های داده‌کاوی است. روش داده‌کاوی پژوهش بر اساس روش CRISP طرح‌ریزی شده است. مدل‌سازی داده‌ها بر پایه داده‌کاوی «توصیفی» و«پیش‌بینی» و استفاده از الگوریتم‌های خوشه‌بندی و طبقه‌بندی است. شاخص سیلوئیت مبنای خوشه‌بندی در نظر گرفته شده و از الگوریتم‌های Two Step، Kohnen و K-Means استفاده شده است. بهترین مقدار، مبتنی بر الگوریتم K-Means برابر ۰/۶۴۴۶ با تعداد خوشه ۵ بود و ویژگی‌های اصلی برای انجام طبقه‌بندی و پیش‌بینی ریسک‌ها تعیین شد. الگوریتم‌های شبکه عصبی، درخت C.5، نزدیک‌ترین همسایگی و بردار پشتیبان برای طبقه‌بندی استفاده شده است. در این پژوهش، الگوریتم ترکیبی پیش‌بینی به‌صورت تکاملی به‌کارگیری شده و در هر مرحله، هدف تقویت میزان صحت و اعتبار مدل طبقه‌بندی و افزایش یادگیری داده‌ها است. نتایج پژوهش، یادگیری در۹۷/۵۶ درصد از داده‌های موردتوافق را نشان داده و میزان صحت و اعتبار مدل ترکیبی برای طبقه‌بندی داده‌ها، ۹۲/۸۶ درصد برآورد شده است. بر اساس نتایج، ۱۳ ریسک، بحرانی تشخیص داده شده‌اند که در این میان «انتشار گازهای آلاینده و مواد شیمیایی» و «عدم‌آموزش و توجیه‌نبودن پیمانکاران نسبت به موقعیت شبکه» به‌ترتیب بیشترین و کمترین اولویت را دارد.

کلیدواژه‌ها: ارزیابی ریسک؛ نگهداری و تعمیرات؛ مدل‌سازی؛ داده‌کاوی؛ شبکه انتقال گاز.

تاریخ دریافت مقاله: ۱۴۰۱/۰۴/۱۵، تاریخ پذیرش مقاله: ۱۴۰۱/۰۶/۲۲.

* دانشجوی دکتری مدیریت صنعتی، گروه مدیریت صنعتی، دانشکده مدیریت و اقتصاد، واحد علوم و تحقیقات، دانشگاه آزاد اسلامی، تهران، ایران.

** دانشیار، گروه مدیریت صنعتی، دانشکده مدیریت و اقتصاد، واحد علوم و تحقیقات، دانشگاه آزاد اسلامی، تهران، ایران. (نویسنده مسئول).

Email: dr.mafshar@gmail.com

*** استاد، گروه مدیریت صنعتی، دانشکده مدیریت و اقتصاد، دانشگاه تربیت مدرس، تهران، ایران.

**** دانشیار، گروه مدیریت، دانشکده مدیریت، دانشگاه تهران، تهران، ایران.

۱. مقدمه

یکی از صنایع پرخطر با ریسک عوامل متنوع، شبکه‌های انتقال گاز هستند و استفاده از فناوری‌های روز دنیا در همه فعالیت‌های مرتبط با این صنعت از اهمیت زیادی برخوردار است. از مهم‌ترین اقدامات برای کاهش مخاطرات، افزایش سطح ایمنی و کنترل سطح هزینه‌های نگهداری و تعمیرات، اجرای برنامه‌های مدیریت و ارزیابی ریسک است. مدیریت ریسک، فرایندی نظام‌یافته برای شناسایی، تحلیل و واکنش به ریسک است که متضمن بیشینه‌کردن احتمال و پیامدهای رویدادهای مثبت و کمینه‌کردن احتمال و پیامدهای رویدادهای منفی است [۱۵]. ارزیابی ریسک یکی از مهم‌ترین ابزارها در نگرش سیستمی به مدیریت ایمنی است؛ به‌گونه‌ای که آن را «قلب سیستم‌های مدیریت ایمنی» نامیده‌اند. ریسک تابعی از احتمال وقوع حادثه و شدت پیامدهای آن است [۷]. مدیران سازمان‌ها می‌کوشند که دیدگاه همه‌جانبه، جامع و یکپارچه‌ای به‌منظور مدیریت ریسک و ارزیابی سطح استقرار آن داشته باشند تا نسبت به تدوین برنامه‌های بهبود اقدام کنند [۲۰]. اجرای برنامه‌های مدیریت ریسک دقیق، تأثیر بسزایی در کاهش هزینه‌های نگهداری تعمیرات دارد و نقش آن در افزایش سطح قابلیت اطمینان دستگاه‌ها و تجهیزات انکارناپذیر است. نگهداری و تعمیرات، یکی از ارکان مهم و اصلی بهره‌وری است؛ بنابراین می‌توان به آن به‌عنوان فرهنگی که سعی در بهبود شرایط موجود دارد، نگریست. افزایش عمر مفید تجهیزات و کارایی سیستم مستلزم یک نظام مناسب برنامه‌ریزی، تجزیه‌وتحلیل، کنترل و اعمال روش‌های صحیح مدیریت است [۲۲]. تحلیل ریسک در هنگام تصمیم‌گیری‌ها و بررسی مشکلات و فرصت‌ها، به‌منظور پشتیبانی از تصمیم‌گیری در سیستم‌های نگهداری و تعمیرات به‌کار می‌رود. در نگهداری مبتنی بر ریسک، برنامه‌ریزی نت بر اساس تحلیل ریسک صورت می‌گیرد. در این روش فعالیت‌های بازرسی و نگهداری بر اساس ریسک‌های کمّی‌شده منتج از شکست تجهیزات اولویت‌بندی می‌شوند؛ به‌طوری‌که ریسک کل سیستم کمینه شود [۲۱]. استفاده از سوابق و اطلاعات برنامه‌های نت و ارزیابی ریسک‌های اجراشده قبلی از اقداماتی است که در تحلیل ریسک‌ها و خطرهای شناسایی‌شده همواره موردتوجه متخصصان است؛ اما چگونگی بهره‌برداری از این داده‌ها و اطلاعات، مسئله‌ای است که امروزه با توجه به ابزارها و فناوری‌های نوین در زمینه بهره‌برداری از پایگاه‌های داده، دارای اهمیت است. با توجه به اینکه در دنیای امروز تصمیم‌گیری‌ها و انتخاب‌های درست، نتایج را رقم می‌زنند و وجود داده‌های زیاد در فناوری‌های جدید، نیاز به منابع برای جمع‌آوری آن‌ها را ضروری کرده است، برای کشف دانش و پاسخ به سؤال‌های ذهنی تصمیم‌گیران باید این داده‌ها تحلیل و بررسی شوند؛ ازاین‌رو «داده‌کاوی» و روش‌های به‌کاررفته در آن برای تصمیمات صحیح در مواقع عادی و کنترل و مدیریت جهت‌دار امور برای مدیران سازمان‌ها از ابزارهای اساسی محسوب می‌شود [۲۲]. در این پژوهش از داده‌کاوی برای پیش‌بینی ریسک‌های بحرانی

در شبکه انتقال گاز از مجموعه داده‌های موجود در زمینه نگهداری تعمیرات و ارزیابی ریسک استفاده شده و مدل‌سازی داده‌ها در دو بخش بر پایه استفاده از الگوریتم‌های «خوشه‌بندی» و »طبقه‌بندی» انجام شده است. تشخیص شرایطی که ظرفیت وقوع حوادث و خطرها را دارد، قبل از اینکه حادثه‌ای به وقوع پیوندد، به میزان زیادی از خسارات مستقیم و غیرمستقیم حوادث می‌کاهد. این موضوع ضرورت بهره‌مندی از روش‌های پیش‌نگر و پیش‌بینانه در شناسایی ریسک‌های بحرانی و دارای اولویت را ایجاب می‌کند؛ مسئله‌ای که کمتر در برنامه‌های مدیریت ریسک به‌صورت داده‌محور به آن توجه شده و در پژوهش‌های صورت‌گرفته، بیشتر با رویکرد پسینی (بررسی داده‌ها بعد از وقوع حادثه) و با استفاده از ابزار و تکنیک‌های سنتی ارزیابی ریسک، به آن پرداخته شده است؛ ازاین‌رو استفاده از ابزار داده‌کاوی و ارائه مدل پیش‌بینی ریسک‌های بحرانی به‌صورت داده‌محور با بهره‌گیری از الگوریتم‌های داده‌کاوی و یادگیری ماشین در این پژوهش موردتوجه قرار گرفته است؛ به‌صورتی که هم‌زمان آموزش داده‌ها و افزایش اعتبار نتایج از طریق ترکیب الگوریتم‌ها به‌صورت یکپارچه (و نه مقایسه تکنیک‌ها با یکدیگر) در راستای هدف پژوهش حاصل شود. داده‌های پژوهش، داده‌های تاریخی مربوط به پیاده‌سازی برنامه‌های نگهداری و تعمیرات در بازه زمانی سه‌ساله در صنعت موردمطالعه است. این داده‌ها مربوط به ۲۵۰ کیلومتر از خط لوله گاز اهواز به سنندج است که از اجرای برنامه نت و بررسی سوابق مربوط به برنامه‌های ارزیابی ریسک پیاده‌سازی شده قبلی به‌دست آمده است. پژوهش حاضر در پنج بخش شامل مقدمه، مبانی نظری و پیشینه پژوهش، روش‌شناسی، تحلیل داده‌ها و یافته‌ها و نتیجه‌گیری و ارائه پیشنهادها برای مطالعات آتی تنظیم شده است.

۲. مبانی نظری و پیشینه پژوهش

ارزیابی ریسک فرایندی شامل شناسایی و طبقه‌بندی فعالیت‌ها، شناسایی خطرها، تخمین میزان ریسک و تصمیم‌گیری در مورد قابلیت تحمل و پذیرش ریسک است. تصمیم‌گیری در مورد قابلیت تحمل یا پذیرش ریسک را «ارزشیابی ریسک» می‌گویند [۱۶].

در ارزیابی ریسک، دو رویکرد ارزیابی ریسک احتمالی و قطعی معرفی و به‌کار برده شده است. در رویکرد قطعی، ورودی‌های مدل با یک برآورد نقطه‌ای مانند میانگین و میانه نمایش داده می‌شوند و همچنین رویکرد قطعی، رویکردی قیاسی است که مشخص می‌کند یک علت تنها به یک پیامد منجر می‌شود و اثر علت‌ها تحلیل می‌شود. رویکرد احتمالی رویکردی آینده‌نگر و پیش‌بینی‌کننده است. در این رویکرد علت ممکن است پیامدهای مختلفی داشته باشد و رویدادها با احتمال وقوع تشخیص داده می‌شوند [۱۳]. برخی از روش‌ها مانند حالات خطا و

تجزیه‌وتحلیل اثرات آن[1]، در بعضی از منابع در دسته روش‌های قطعی و در برخی دیگر در روش‌های احتمالی قرار گرفته است. این روش در صنایع فرایندی کاربرد بسیار دارد و با هر دو رویکرد قابلیت بحث و بررسی است [۲۵]. داده‌های استفاده‌شده در این پژوهش نیز حاصل بررسی‌های قبلی و محاسبات صورت‌گرفته با این تکنیک است که در یک دوره سه‌ساله توسط متخصصان ارزیابی ریسک در صنعت موردمطالعه، انجام شده است.

داده‌کاوی. داده‌کاوی فرآیند کشف دانش مطلوب از مقدار بزرگی از داده است که در پایگاه داده، انبار داده و دیگر مخازن اطلاعات ذخیره شده است [۱۱]. داده‌کاوی پل ارتباطی میان علم آمار، علم رایانه، هوش مصنوعی، الگوشناسی، فراگیری ماشین و بازنمایی بصری داده است و به‌صورت یک محصول قابل خریداری نیست؛ بلکه یک رشته علمی و فرآیندی است که باید به‌صورت یک پروژه اجرا شود [۱۹]. داده‌کاوی درحقیقت فرایند کاملی از اعمال روش‌های مبتنی بر رایانه شامل تکنیک‌های جدید برای اکتشاف دانش از داده‌ها است. فرایند داده‌کاوی به‌صورت کلی گام‌های بیان مسئله، جمع‌آوری داده‌ها، انجام پیش‌پردازش داده‌ها، کاوش داده‌ها، مدل‌سازی، برآورد و تفسیر مدل و رسیدن به نتایج را شامل می‌شود [۳].

الگوریتم‌های داده‌کاوی به دو دسته کلی نظارتی و غیرنظارتی یا پیش‌بینی و توصیفی تقسیم شده است. در الگوریتم‌های پیش‌بینی، هدف، پیش‌بینی یک ویژگی خاص بر مبنای ویژگی‌های دیگر است. ویژگی پیش‌بینی‌شونده «متغیر وابسته» و بقیه متغیرها «متغیر مستقل» نامیده می‌شوند. در الگوریتم‌های توصیفی، هدف استخراج الگو از داده‌ها است که نیاز به تحلیل نتایج دارد و به دنبال کشف راهی برای آگاهی از خصوصیات داده است [۱۱].

روش‌های مربوط به الگوریتم‌های پیش‌بینی

طبقه‌بندی یا کلاس‌بندی. موجب شناسایی ویژگی‌های گروهی که هر مورد به آن تعلق دارد می‌شود.

رگرسیون. با استفاده از مقادیر موجود، مقادیر دیگر را پیش‌بینی می‌کند. رگرسیون معمولاً از تکنیک‌های آماری استاندارد مانند رگرسیون خطی استفاده می‌کند.

تحلیل زمانی. مقادیر ناشناخته را با استفاده از پیش‌بینی‌کننده‌های متغیر با زمان، مثل رگرسیون، پیش‌بینی می‌کند.

1. FMEA

روش‌های مربوط به داده‌کاوی توصیفی

کشف توالی. کشف مواردی که اتفاق‌افتادن آن‌ها در وجود یا عدم‌وجود موارد دیگر نقش دارد.

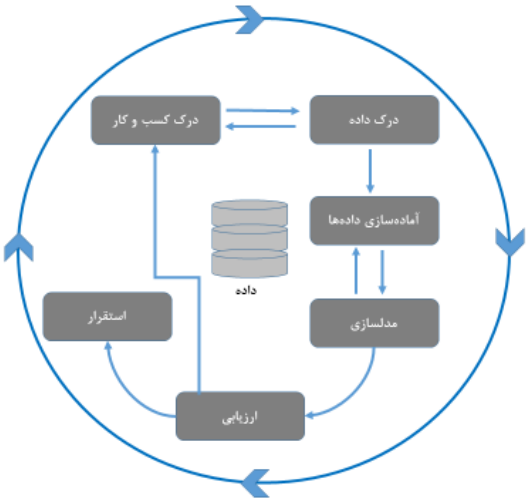
قواعد انجمنی. یافتن قواعد در مواردی که باهم اتفاق می‌افتند را کشف قواعد وابستگی می‌نامند.

خوشه‌بندی. در این روش داده‌های موجود در یک گروه بسیار شبیه به هم هستند و داده‌های گروه‌های مختلف بیشترین تفاوت را باهم دارند.

خلاصه‌سازی. برای شناخت دقیق داده‌ها و انجام کاوش روی آن‌ها از ابزارهایی برای خلاصه‌سازی نتایج استفاده می‌شود [۲۶].

روش‌های مختلفی برای پیاده‌سازی و اجرای پروژه‌های داده‌کاوی وجود دارد. یکی از این موارد روش CRISP-DM[1] است. موفق‌ترین پروژه‌های داده‌کاوی، در چارچوب این فرآیند استاندارد اجرا می‌شود و به‌دلیل گستردگی کاربرد در نمونه مطالعات مشابه، در این پژوهش از این روش بهره‌گیری شده است. در این روش، نتیجه هر مرحله به نتیجه مراحل قبلی وابستگی دارد. نکته قابل‌توجه دیگر در این مدل، امکان بازگشت به هر یک از مراحل قبل و اعمال اصلاحات موردنیاز است. این روش از گام‌های درک کسب‌وکار[2]، درک داده[3]، آماده‌سازی داده[4]، مدل‌سازی[5]، ارزیابی[6] و پیاده‌سازی[7]، تشکیل شده است که هر یک شامل زیربخش‌های مربوط به خود است. شکل ۱، ارتباط بین مراحل این روش را نشان می‌دهد.

1. Cross Industry Standard Process For Data Mining
2. Business understanding
3. Data Understanding
4. Data Preparation
5. Modeling
6. Evaluation
7. Deployment



شکل۱. گام‌های روش CRISP-DM/۴، ۱۹/

فاز درک کسب‌وکار. در نخستین مرحله این روش، تمرکز بر درک اهداف پروژه و نیازمندی‌های آن است. پس از تعیین اهداف باید به شناخت وضعیت موجود پرداخت و منابع موجود، نیازمندی‌ها و محدودیت‌ها تعیین می‌شود. در ادامه طرح امکان‌سنجی تدوین خواهد شد.

فاز درک داده. این فاز با جمع‌آوری داده اولیه آغاز می‌شود و طی فعالیت‌های آشنایی با داده مشکلات کیفیتی داده شناسایی می‌شود. شناخت داده‌ها عبارت است از: جمع‌آوری داده‌های اولیه؛ توصیف داده‌ها؛ بازرسی و بررسی داده‌ها و اعتبارسنجی کیفی داده‌ها.

فاز آماده‌سازی داده. این فاز تمام فعالیت‌هایی که برای ساخت مجموعه داده‌ی نهایی از داده‌های خام اولیه لازم است را دربرمی‌گیرد .آماده‌سازی داده شامل جدول‌بندی، ثبت، انتخاب مشخصه و همچنین انتقال و پاک‌سازی داده برای ابزار مدل‌سازی است.

فاز مدل‌سازی. با انتخاب و به‌کاربستن تکنیک‌های مدل‌سازی مناسب و روش‌های داده‌کاوی معین، نتایج مدل‌سازی بهینه می‌شود. در نخستین گام از مدل‌سازی باید روش مناسب را انتخاب کرد. پارامترهای موردنیاز مدل نیز پس از تعیین روش مورداستفاده مشخص می‌شود. در هر مرحله برای تضمین کیفیت مدل طراحی‌شده، مدل باید آزمون شود و در صورت عدم‌دقت یا صحت نتایج باید پارامترها تغییر کنند.

فاز ارزیابی. قبل از اجرای نهایی مدل، باید تعیین شود که آیا این مدل امکان دستیابی به اهداف تعیین‌شده در مرحله نخست را فراهم می‌کند و در صورت نیاز مراحل اجراشده برای ایجاد این مدل بازبینی شود. در این گام اعتبار مدل بررسی می‌شود.

فاز پیاده‌سازی و توسعه. به‌طورکلی خلق مدل انتهای پروژه نیست. حتی اگر هدف مدل افزایش دانش داده باشد، دانش به‌دست‌آمده نیاز دارد که سازمان‌دهی شود و به‌صورتی که برای مشتری قابل‌استفاده باشد ارائه شود. بسته به نیازمندی‌ها، فاز پیاده‌سازی می‌تواند به‌سادگی ایجاد یک گزارش یا به پیچیدگی اجرای یک فرآیند داده‌کاوی قابل‌تکرار باشد. پس از اجرای اقدامات، دورنمایی از طرح توسعه ایجاد می‌شود. جزئیات هر یک از زیربخش‌های روش CRISP-DM در شکل ۲، ارائه شده است [۱۴، ۱۹].

شکل ۲. زیربخش‌های مراحل روش CRISP/۱۴]

انجام پیش‌بینی‌های مبتنی بر تکنیک‌های داده‌کاوی در زمینه‌های متعدد مطالعاتی کاربرد داشته و پژوهشگران زیادی در این باره پژوهش کرده‌اند. این مطلب در مورد موضوع مدیریت و ارزیابی ریسک نیز صادق است؛ اما ویژگی اصلی متمایزکننده این پژوهش از موارد مشابه، آموزش داده‌ها با یادگیری ماشین است. در این پژوهش، الگوریتم ترکیبی پیش‌بینی برای بهینه‌سازی طبقه‌بندی داده‌ها، به‌صورت تکاملی به‌کارگیری شده و در هر مرحله، هدف تقویت میزان صحت و اعتبار مدل طبقه‌بندی و افزایش میزان یادگیری داده‌ها است. درحالی‌که سایر پژوهش‌های مشابه بر رویکرد مقایسه‌ای در به‌کارگیری تکنیک‌ها متمرکز است و کارایی و صحت آن‌ها در مقایسه با یکدیگر مورد قضاوت و ارزیابی قرار می‌گیرد. همچنین استفاده هم‌زمان از رویکردهای توصیفی و پیش‌بینی داده‌کاوی برای دستیابی به هدف پژوهش و تلفیق ویژگی‌ها و ابعاد دو مقوله «نگهداری تعمیرات» و «ارزیابی ریسک» برای جامعیت پایگاه داده و پیشگیری از خطا در انتخاب ویژگی‌های اصلی و کاهش ابعاد داده‌ها از موارد دیگری است که

مورتوجه قرار گرفته است. در ادامه چند نمونه از پژوهش‌های مشابه با پژوهش حاضر ذکر شده است.

پیشینه پژوهش. با توجه به گستردگی کاربرد و زمینه‌های فراوانی که برای بهره‌برداری از علم داده‌کاوی وجود دارد، پژوهش‌های زیادی در حوزه‌های مختلف دانشی با موضوع استفاده از قابلیت‌ها و ابزار مرتبط با داده‌کاوی انجام شده است. به‌خصوص در زمینه ارزیابی ریسک می‌توان به پژوهش چنگ[1] و همکاران (۲۰۱۵)، اشاره کرد. آن‌ها با استفاده از شبکه عصبی و الگوریتم درخت تکاملی، داده‌ها که درواقع نتایج به‌دست‌آمده از پیاده‌سازی ارزیابی ریسک با FMEA است را تصویرسازی و خوشه‌بندی کردند و استفاده از این مدل را در ادامه تکنیک FMEA پیشنهاد دادند. بایگ[2] و همکاران (۲۰۱۳)، از تکنیک ارزیابی ریسک تحلیل درخت خطا برای تحلیل قابلیت اطمینان استفاده کرده و هدف از انجام آن را توسعه قابلیت اطمینان در سیستم‌های ایمنی ذکر کردند. در پژوهشی مشابه عبدالعزیز و هلال[3] (۲۰۱۲)، کاربرد FMEA و FTA را در برنامه‌ریزی نگهداری تعمیرات مبتنی بر قابلیت اطمینان موردبررسی قرار داده‌اند. یانگ[4] و همکاران (۲۰۱۵)، از داده‌کاوی به‌عنوان ابزاری برای جداسازی خطاها و اعتبارسنجی روش ارزیابی ریسک FMEA بهره گرفتند. استین وینکل[5] و همکاران (۲۰۲۱)، با استفاده از دو رویکرد دانش‌محور و داده‌محور مدلی را ارائه کردند که با بهره‌گیری از یادگیری ماشین، ناهنجاری‌ها در جریان داده در حسگرها را تشخیص دهد. همتا و همکاران (۲۰۱۷)، بهبود قابلیت اطمینان و افزایش اثربخشی برنامه‌ریزی تعمیرات را در «مجتمع پتروشیمی شازند) با تلفیق رویکردهای داده‌کاوی و تکنیک‌های ارزیابی ریسک بررسی کردند. آن‌ها برای این منظور از روش خوشه‌بندی k میانگین و تکنیک‌های ارزیابی ریسک FMEA و FTA بهره گرفتند و هدف در این پژوهش بهبود میزان قابلیت اطمینان بود. در زمینه‌های بسیار دیگری مانند ارزش‌گذاری مشتریان بانکی، پیش‌بینی عوامل مؤثر در مصرف انرژی، تشخیص بیماری‌ها و پیش‌بینی عوامل بیماری‌زا با استفاده از مجموعه اطلاعات بالینی بیماران و بسیاری از زمینه‌های موضوعی و دانشی دیگر، تکنیک‌ها و الگوریتم‌های داده‌کاوی برای ارائه مدل‌های پیش‌بینی، استخراج و کشف الگوی دانش، خوشه‌بندی، طبقه‌بندی و یا مقایسه الگوریتم‌ها با یکدیگر کاربرد داشته و مورداستفاده قرار گرفته است. در این پژوهش، الگوریتم ترکیبی پیش‌بینی برای بهینه‌سازی طبقه‌بندی داده‌ها، به‌صورت تکاملی به‌کار رفته و در هر مرحله، هدف تقویت میزان صحت و اعتبار مدل طبقه‌بندی و افزایش میزان یادگیری داده‌ها است؛ درحالی‌که سایر پژوهش‌های مشابه بر رویکرد مقایسه‌ای

1. Chang
2. Bayg
3. Abdel-aziz & Helal
4. Yang
5 Steenwinckel

در به‌کارگیری تکنیک‌ها تمرکز دارند و کارایی و صحت آن‌ها در مقایسه با یکدیگر مورد قضاوت و ارزیابی قرار می‌گیرد. استفاده هم‌زمان از رویکردهای توصیفی و پیش‌بینی داده‌کاوی برای دستیابی به هدف پژوهش و تلفیق ویژگی‌ها و ابعاد دو مقوله «نگهداری تعمیرات» و «ارزیابی ریسک» برای جامعیت پایگاه داده و پیشگیری از خطا در انتخاب ویژگی‌های اصلی و کاهش ابعاد داده‌ها از موارد دیگری است که موردتوجه قرار گرفته است.

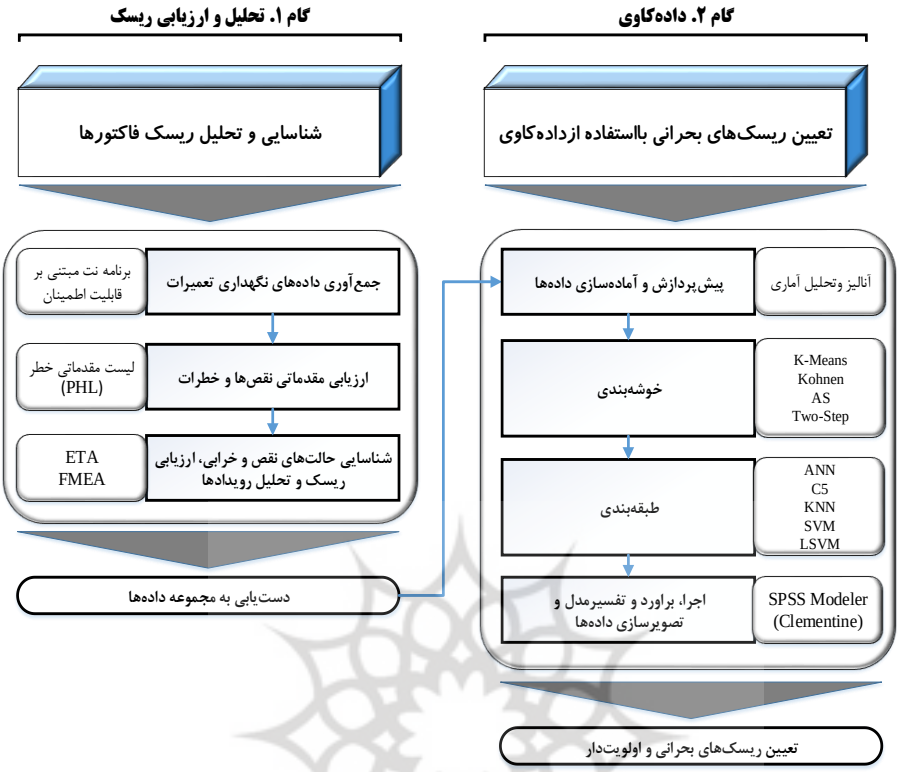
۳. روش‌شناسی پژوهش

پژوهش حاضر از نوع توصیفی با رویکرد کاربردی است. ابزار استفاده‌شده در پژوهش، خوش‌بندی و طبقه‌بندی داده‌ها مب‌باشد که از تکنیک‌های یادگیری ماشین است و در دو دسته داده‌کاوی «پیش‌بینی» و «توصیفی» قرار می‌گیرد. در گام نخست پژوهش، داده‌های تاریخی مربوط به اجرای برنامه‌های نگهداری و تعمیرات و ارزیابی ریسک در بازه زمانی سه‌ساله در صنعت موردمطالعه جمع‌آوری شده است. خروجی گام نخست پژوهش بر اساس تحلیل و ارزیابی صورت‌گرفته با بهره‌گیری از داده‌های موجود، دستیابی به پایگاه داده است.

در گام دوم، فعالیت‌های مربوط به پیش‌پردازش، آماده‌سازی داده‌ها، مشخص‌کردن داده‌های پرت و داده‌های مفقوده، نرمال‌سازی داده‌ها، ایجاد داده تمیز و سازمان‌دهی مجموعه داده انجام شده است. برای این منظور از نرم‌افزار SPSS Modeler 18.0 استفاده شد. گره‌های[1] مربوط به انتخاب ویژگی‌های اصلی، کاهش ابعاد داده‌ها، انجام تحلیل‌های آماری و مدیریت داده‌های پرت و مفقوده در نرم‌افزار در این بخش موردتوجه بوده و از آن استفاده شده است. شاخص‌های چولگی[2] و کشیدگی[3] برای آزمون نرمال‌بودن داده‌ها و استفاده از گره‌های Anomaly Detect مربوط به مدیریت داده‌های مفقوده، Data Audit و Feature selection از اقداماتی است که آماده‌سازی مجموعه داده را سبب شده است. در ادامه برای تعیین ریسک‌های بحرانی و اولویت‌دار، داده‌های مربوط به هر ریسک عامل در قالب روش‌های خوشه‌بندی و طبقه‌بندی تحلیل و ارزیابی شده است. در این مرحله برای خوشه‌بندی داده‌ها از الگوریتم‌های K-Means، Kohnen و Two-step و برای طبقه‌بندی از الگوریتم‌های شبکه عصبی، درخت C.5، نزدیک‌ترین همسایگی و بردار پشتیبان استفاده شد و برای سنجش اعتبار آن، نتایج به‌دست‌آمده با نتایج خوشه‌بندی و طبقه‌بندی پیش‌فرض نرم‌افزار[4] و نتایج گره آنالیز[5] تطبیق داده شده و درنهایت پس از اجرا و تفسیر الگوریتم‌ها، داده‌ها مدل‌سازی و تصویرسازی شدند. نتیجه

1. Nodes
2. Kurtosis
3. Skewness
4. Auto cluster & Auto Classify
5. Analysis Node

مدل‌سازی داده‌ها، شناسایی و تعیین ریسک‌های بحرانی و اولویت‌دار است که مبتنی بر طبقه‌بندی و خوشه‌بندی آن‌ها است. شکل ۳، مراحل انجام پژوهش را نشان می‌دهد.



شکل ۳. مراحل انجام پژوهش

ارزیابی ریسک. داده‌های استفاده‌شده در پژوهش نتایج اجرای فرایند ارزیابی ریسک و شناسایی نقص‌ها و خرابی‌ها در یک بازه زمانی سه‌ساله است. برای این منظور دو رویکرد ارزیابی ریسک احتمالی و قطعی در مطالعات این حوزه توسط متخصصان در صنعت موردمطالعه به‌کار برده شده است که عبارت‌اند از: تحلیل درخت رویداد و تجزیه‌وتحلیل حالت‌های نقص و اثرات. واکاوی و تحلیل درخت رویداد روشی برای شناسایی و ارزیابی توالی وقایع و رویدادها در یک سناریوی حادثه بالقوه و به دنبال وقوع یک رویداد ابتدایی (شروع‌کننده) است. این تکنیک از یک ساختار درخت منطقی که «درخت واقعه»[1] نامیده می‌شود، بهره می‌گیرد. بیشتر درخت‌های رویداد بر اساس یک منطق دوحالتی (موفقیت یا شکست) طراحی می‌شوند و در این مسیر وقایع میانی بررسی می‌شوند و درنهایت احتمالات هر مسیر به‌صورت جداگانه محاسبه می‌شود. نتیجه یک

1. ET

ETA دستیابی به پیامدهای ممکن مختلف ناشی از یک واقعه اولیه و به‌دست‌آوردن احتمال هر پیامد است [۱، ۲۷]. هدف ETA تعیین چگونگی پیشرفت یک رویداد اولیه به یک حادثه جدی و یا تعیین چگونگی کنترل یک واقعه به‌وسیله دستورالعمل‌های ایمنی است [۱۸، ۲۵]. FMEA تکنیکی تحلیلی است که برای تشخیص، کاهش و حذف خطاها و مشکلات بالقوه و بالفعل موجود در سیستم، استفاده می‌شود [۶، ۲۷]. در نخستین گام در فرایند تجزیه‌وتحلیل حالات خرابی و آثار آن، سیستم به عناصر مجزا تقسیم‌بندی می‌شود. در ادامه و پس از احصای ریسک‌ها، برآورد ریسک با محاسبه عدد اولویت ریسک[1] برای هر حالت بالقوه خطا و اثر آن انجام می‌گیرد. این متغیر با ضرب سه عامل «شدت پیامد»، «احتمال وقوع» و «قابلیت کشف خطا» محاسبه می‌شود. این سه عامل با مقیاسی از یک تا ۱۰ درجه‌بندی می‌شوند. عدد اولویت ریسک، مبنای اولویت‌بندی حالات خرابی است. با توجه به اینکه سه شاخص بالا می‌توانند اعدادی بین ۱ تا ۱۰ اختیار کنند، RPN مقداری بین ۱ تا ۱۰۰۰ خواهد بود [۱۸، ۲۵]. قبل از استفاده از روش‌های ارزیابی ریسک، تهیه فهرست مقدماتی خطر[2] سبب تسهیل فعالیت‌ها و اقدامات در زمینه مدیریت و ارزیابی ریسک می‌شود.

فرایند داده‌کاوی پژوهش. در این پژوهش داده‌ها و اطلاعات، درواقع اطلاعات مربوط به رویدادها و نقص‌هایی هستند که در جریان فرایند تولید در مطالعه موردی صورت‌گرفته ثبت می‌شود. این داده‌ها مبتنی بر نتایج برنامه‌های نگهداری و تعمیرات پیاده‌شده و نتایج ارزیابی ریسک‌های انجام‌شده است. اطلاعات مربوط به برنامه‌های نگهداری و تعمیرات با استفاده از اصول مبتنی بر علم داده و به‌صورت داده‌محور[3] و داده‌های مربوط به تحلیل و ارزیابی ریسک با استفاده از تلفیق منطقی برخی از ویژگی‌ها و ثبت نتایج جدید مبتنی بر دانش ارزیابی ریسک[4] ثبت شده است؛ به‌عبارت‌دیگر در زمینه ارزیابی ریسک، تلفیق برخی از داده‌ها مبتنی بر روابط ریاضی باعث ایجاد داده‌های جدید یا فراداده[5] شده است. مجموعه داده[6] ایجادشده متشکل از ۲۴۰ نمونه (رکورد) و ۱۵ ویژگی (خصیصه) است که ۱۱ منطقه شامل ۴۰ ناحیه را در محیط موردمطالعه شامل می‌شود. منطقه‌بندی انجام‌شده بر اساس ناحیه‌بندی صورت‌گرفته قبلی برای اجرای برنامه‌های مدیریت و ارزیابی ریسک است. ویژگی‌ها و تعریف مربوط به هر ویژگی در جدول ۱، مشاهده می‌شود.

1. RPN
2. PHA
3. Data Driven
4. Knowledge Driven
5. Meta Data
6. Data Set

جدول ۱. ویژگی‌های مجموعه داده [۱۶، ۲۴]

عنوان	تعریف
نرخ فراوانی (FR)	احتمال یا به‌عبارت‌دیگر شمارش (فراوانی) تعداد شکست‌ها
شدت نقص (SR)	نتیجه ارزیابی و سنجش نتیجه شکست (نقص) و سطح پیامدها
نرخ کشف (OR)	احتمال بازیابی (کشف) نقص قبل از آنکه اثر وقوع آن مشخص شود.
عدد اولویت ریسک	از ضرب سه عامل نرخ فراوانی، شدت نقص و نرخ کشف به‌دست می‌آید.
تعداد خرابی	تعداد دفعاتی که نقص/ خرابی در بازه زمانی سه‌ساله در یک تجهیز اتفاق‌افتاده
میانگین زمانی وقوع نقص (MTBF)	متوسط زمان بین دو خرابی متوالی
متوسط زمان لازم برای تعمیر (MTTR)	متوسط زمان لازم برای تعمیر
احتمال وقوع نقص (FMP)	قضاوت متخصصان ارزیابی ریسک بر اساس تواتر وقوع نقص و نتایج ارزیابی ریسک موجود در بازه زمانی سه‌ساله
نرخ کیفیت	نسبت محصول تولیدشده قابل‌قبول به کل مقدار محصول تولیدی توسط ماشین
نرخ قابلیت دسترسی	نسبت زمان بهره‌برداری و تولید به زمان اشغال یک تجهیز
نرخ کارایی	مدتی که بر اساس سرعت اسمی برای تولید مقدار مشخص محصول صرف می‌شود.
اثربخشی کلی تجهیزات (OEE)	از ضرب سه عامل نرخ کیفیت، نرخ قابلیت دسترسی و نرخ کارایی به‌دست می‌آید.
سطح هزینه نگهداری و تعمیرات	قضاوت متخصصان نگهداری و تعمیرات در سه سطح کلی «کم»، «متوسط» و «زیاد» با توجه به نتایج مربوط به هزینه‌های نت صورت‌گرفته برای یک نقص
هزینه نگهداری و تعمیرات (ریال)	مجموع هزینه‌های صورت‌گرفته برای اجرای اقدامات اصلاحی و برنامه نت
برآورد کلی اهمیت نقص	قضاوت متخصصان ارزیابی ریسک در سه سطح کلی «کم»، «متوسط» و «زیاد»

برای انجام فرایند داده‌کاوی از نرم‌افزار SPSS Modeler 18(Clementine) بهره‌گیری شده است. در گام نخست، ایجاد جریان داده[1] و انجام عملیات پیش‌پردازش و آماده‌سازی داده‌ها صورت گرفته است. برای این منظور آزمون نرمال‌بودن داده‌ها و استفاده از گره‌های Feature Selection, Data Audit, Anomaly Detection, Filtering, Type از اقداماتی است که موردتوجه پژوهشگران قرار گرفته و تشخیص داده‌های مفقوده و پرت، انتخاب ویژگی‌های اصلی، تلفیق داده‌ها و کاهش ابعاد، پاک‌سازی داده‌ها و ایجاد مجموعه داده تمیز برای مدل‌سازی از نتایج این بخش است. در مرحله مدل‌سازی، خوشه‌بندی و طبقه‌بندی داده‌ها انجام شد که برای این منظور از الگوریتم‌های Two-Step، Kohnen و K-Means برای خوشه‌بندی و از

1. Data Streame

الگوریتم‌های SVM، C5، KNN و ANN برای طبقه‌بندی استفاده شده است. برای ارزیابی صحت و دقت نتایج شاخص‌های حساسیت، شفافیت، دقت و صحت موردارزیابی قرار گرفت.

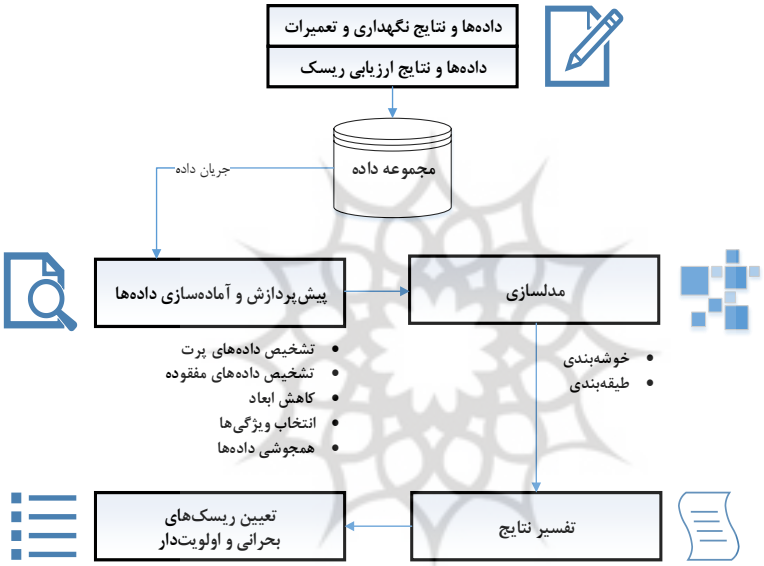
ـ حساسیت= تعداد داده‌های عضو کلاس که درست دسته‌بندی شده‌اند/ تعداد کل داده‌های عضو کلاس

ـ شفافیت= تعداد داده‌های غیرعضو کلاس که درست دسته‌بندی شده‌اند/ تعداد کل داده‌های غیرعضو کلاس

ـ دقت= تعداد داده‌های عضو کلاس که درست دسته‌بندی شده‌اند/ تعداد داده‌های عضو کلاس که درست دسته‌بندی شده‌اند + تعداد داده‌های غیرعضو کلاس که به اشتباه دسته‌بندی شده‌اند.

ـ صحت = حساسیت* (تعداد داده‌های عضو کلاس/ تعداد کل داده‌ها) + شفافیت*(تعداد داده‌های غیرعضوکلاس/ تعداد کل داده‌ها) [۱۱].

مبتنی بر نتایج خوشه‌بندیِ رویدادها با توجه به شاخص‌های برگزیده، خوشه بحرانی انتخاب و در ادامه پس از تعیین شاخص هدف به‌عنوان خروجی و تعیین ویژگی‌های ورودی، رویدادها (ریسک‌ها یا نقص‌ها) طبقه‌بندی شده‌اند. خلاصه نمای فرایند داده‌کاوی پژوهش در شکل ۴، ارائه شده است.



شکل ۴. روش‌شناسی فرایند داده‌کاوی پژوهش

مدل‌سازی، تحلیل و پردازش داده. انتخاب و اجرای تکنیک مناسب داده‌کاوی اقدام اصلی در این مرحله است؛ به شکلی که بتوان بهترین تصمیم را در مورد ورودی‌ها، خروجی‌ها و ترتیب و توالی به‌کارگیری هر الگوریتم اتخاذ و فرایند کشف دانش را به بهترین نحو ممکن اجرا کرد.

الگوریتم خوشه‌بندی K-Means . الگوریتم خوشه‌بندی K-Means یکی از ساده‌ترین و البته مشهورترین الگوریتم‌های یادگیری بدون نظارت است. ایده اصلی در این الگوریتم تعریف

K مرکز برای هریک از خوشه‌ها است. بهترین انتخاب برای مراکز خوشه‌ها در الگوریتم K-Means قراردادن آن‌ها (مراکز) در فاصله هرچه بیشتر از یکدیگر است. در این نوع خوشه‌بندی تابع هدفی وجود دارد که تابع خطا است و حداقل‌سازی آن مدنظر است. مشهورترین معیارهای محاسبه فاصله رکوردها در روش‌های خوشه‌بندی معیارهای فاصله اقلیدسی و فاصله همینگ هستند. در گام بهبود الگوریتم جابه‌جایی انتقال عضوی که بیشترین فاصله را با مرکز خوشه خودش دارد، به خوشه‌ای که کمترین فاصله را با آن دارد صورت گرفته و دستور توقف، زمان تغییرنیافتن اعضای خوشه‌ها یا کاهش نیافتن مقدار تابع خطا است. در منابع مختلف شاخص‌های متفاوتی برای سنجش میزان دقت و صحت مدل پیشنهاد شده است که شاخص دیویس ـ بولدین[1]، دان و سیلهوئت[2] از جمله آن‌ها است. در این پژوهش به‌دلیل گستردگی کاربرد، شاخص سیلهوئت به کار رفته است [۱۱، ۲۳].

الگوریتم خودسازمان‌ده کوهنن[3]. در این الگوریتم آموزشی گره‌های شبکه به‌صورت همسایه‌های محلی تنظیم می‌شود؛ به‌طوری‌که بتوانند ویژگی‌های داده‌های ورودی را طبقه‌بندی کنند. نقشه شبکه به‌طور خودکار با مقایسه دوره‌ای ورودی‌های هر گره با برداری که توسط آن گره در خطوط ارتباطی آن بردار ذخیره شده است، شکل می‌گیرد. هرگاه ورودی‌های یک گره با بردار ذخیره‌شده آن گره مطابقت کند آن ناحیه شبکه به‌طور گزینشی بهینه می‌شود. شبکه درحالی‌که ابتدا به‌طور تصادفی شکل گرفته است و به‌تدریج خود را تنظیم می‌کند، به حالت پایداری می‌رسد که به‌طور موضعی ویژگی‌های داده‌های ورودی را نمایش می‌دهد. مراحل الگوریتم کوهنن به‌طور خلاصه عبارت‌اند از: تعیین مقادیر اولیه که شامل تعداد ورودی‌ها،گره‌های خروجی و ضرایب وزنی اولیه است؛ واردکردن بردار ورودی بر اساس مقدار ورودی هر گره در زمان مشخص، محاسبه فواصل و انتخاب کوتاه‌ترین فاصله؛ اصلاح ضرایب وزنی و ادامه الگوریتم تا زمانی که کمترین میزان فاصله محاسبه شود [۲]

خوشه‌بندی سلسله‌مراتبی دومرحله‌ای[4]. این الگوریتم به‌جای تلاش برای پیش‌بینی یک خروجی، سعی در کشف الگوها از مجموعه‌ای از فیلدهای ورودی دارد. در این روش داده‌ها به‌گونه‌ای خوشه‌بندی می‌شوند که اعضای یک خوشه به هم شبیه و داده‌های درون خوشه‌های مجزا با هم متفاوت باشند. این روش یک روش خوشه‌بندی دومرحله‌ای است. در گام نخست با یک گذر از داده‌های ورودی، داده‌ها در قالب زیرخوشه‌ها فشرده‌سازی شده و در گام دوم با روش

1. Davies-Bouldin
2. Silhouette
3. Kohnen
4. Two-step

خوشه‌بندی سلسله‌مراتبی، به‌منظور تکامل خوشه‌بندی، این زیرخوشه‌ها ادغام و تبدیل به خوشه‌های بزرگ‌تر می‌شود [۴].

الگوریتم نزدیک‌ترین همسایگی[1]. این روش به‌عنوان یک الگوریتم ناپارامتری هیچ فرضی بر توزیع داده‌های ورودی ایجاد نمی‌کند. این روش بر اساس شباهت داده‌ها طبقه‌بندی را انجام می‌دهد. درواقع همسایه نزدیک را محاسبه کرده و برچسب k برای هر داده‌ی آزمایشی جدید، فواصل همسایه برای نقطه موردنظر را تعیین می‌کند. این الگوریتم از سه‌گام اصلی تشکیل شده است: محاسبه فاصله نمونه ورودی با تمام نمونه‌های آموزشی؛ مرتب کردن نمونه‌های آموزشی بر اساس فاصله و انتخاب k همسایه نزدیک‌تر؛ استفاده از دسته‌ای که اکثریت را در همسایه‌های نزدیک به‌عنوان تخمینی برای دسته نمونه ورودی دارد [۱۴]. یکی از مشکلاتKNN این است که به همه نمونه‌های آموزشی نیاز دارد که در زمان اجرا اصلی الگوریتم در حافظه باشند؛ به همین دلیل، «طبقه‌بندی مبتنی بر حافظه» نامیده می‌شود [۱۲].

الگوریتم درخت تصمیم C5. الگوریتم‌هایی که برای ساخت درخت تصمیم استفاده می‌شوند اساساً یک فرایند شبیه به هم را اجرا می‌کنند؛ بدین صورت که تمامی فیلدهای پایگاه داده را بررسی می‌کنند تا به ویژگی برسند که بهترین دسته‌بندی و پیش‌بینی را با تقسیم داده‌ها به زیرگروه‌ها انجام دهد. این فرایند به‌صورت بازگشتی تکرار می‌شود تا باز هم زیرگروه‌ها به زیرگروه‌های دیگری شکسته شود. متغیرهای هدف یا ورودی می‌توانند عددی یا طبقه‌ای باشند. اگر یک متغیر فاصله‌ای مورداستفاده قرار گیرد، نتیجه کار یک درخت رگرسیون و اگر ورودی‌ها به‌صورت طبقه‌ای باشد، نتیجه کار یک درخت دسته‌بندی خواهد بود. الگوریتم C5 از الگوریتم‌های پرکاربرد برای ساخت درخت‌های تصمیم است. این روش که توسعه‌یافته الگوریتم ID3 است، می‌تواند برای بیان دسته‌بندی به‌صورت درخت تصمیم و یا مجموعه قوانین به‌کار برده شود. در بسیاری از برنامه‌های کاربردی، مجموعه قوانین ترجیح داده می‌شوند؛ زیرا درک آن‌ها نسبت به درخت‌های تصمیم‌گیری، ساده‌تر است [۵]. درخت تصمیم یک توصیف صریح از شاخه‌زنی با استفاده از الگوریتم است. هر گره پایانی یا برگ یک زیرمجموعه از داده‌های آموزشی را توصیف می‌کند و هر نمونه در بخش آموزش دقیقاً به یک گره پایانی در درخت تعلق دارد [۴].

ماشین بردار پشتیبان[2]. یک ابزار ریاضی است که مبتنی بر اصل حداقل‌سازی خطای عملیاتی است و در کاربردهای امروزی یادگیری ماشین، ماشین بردار پشتیبان به‌عنوان یکی از

1. KNN
2. SVM

قدیمی‌ترین و دقیق‌ترین روش‌ها در میان الگوریتم‌های معروف شناخته می‌شود. این الگوریتم جزو الگوریتم‌های تشخیص الگوی دسته‌بندی است. از الگوریتم SVM در هر جایی که نیاز به تشخیص الگو یا دسته‌بندی اشیا در کلاس‌های خاص باشد، می‌توان استفاده کرد؛ همچنین ماشین بردار پشتیبان یکی از روش‌های یادگیری با نظارت است که از آن برای طبقه‌بندی و رگرسیون استفاده می‌کنند. مبنای این مدل، دسته‌بندی خطی داده‌ها است و در تقسیم خطی داده‌ها سعی بر آن است خطی انتخاب شود که حاشیه اطمینان بیشتری را داشته باشد [۱۲].

شبکه‌های عصبی. شبکه عصبی مصنوعی[1] و مغز انسان مشابه هم هستند؛ زیرا هر دوی آن‌ها شامل تعداد زیادی پردازش و واحدهای هوشمند هستند که «نورون‌ها یا سلول‌های مغزی» نامیده می‌شوند. هدف توسعه شبکه عصبی مصنوعی، یافتن رابطه بین داده‌های ورودی و داده‌های خروجی است. شبکه‌های عصبی مصنوعی را می‌توان ابرمجموعه تمامی شبکه‌های عصبی روبه‌جلو با قابلیت یادگیری خواند. پایه قانون یادگیری در شبکه‌های عصبی، روش گرادیان نزولی است که معمولاً بسیار کند بوده و در نقطه کمینه محلی گیر می‌کند؛ ازاین‌رو روش یادگیری پیشنهاد شده عموماً روش آموزش هیبرید است که درواقع ترکیبی از گرادیان نزولی و حداقل میانگین مربعات خطا است. مزیت آموزش هیبرید نسبت به گرادیان نزولی علاوه بر کاهش ابعاد فضای جست‌وجو، افزایش سرعت همگرایی سیستم نیز است [۸، ۲۳، ۱۲]. شبکه‌های عصبی به دو دسته پیش‌خور و برگشتی تقسیم‌بندی می‌شوند. در یک شبکه پیش‌خور، نورون‌ها به‌صورت لایه‌ای گروه‌بندی می‌شوند و سیگنال‌ها از لایه ورودی به طرف لایه خارجی با اتصالات تک‌جهته جریان پیدا می‌کنند. شبکه چندلایه پروسپترون[2] شناخته‌شده‌ترین نوع شبکه‌های پیش‌خور است. شبکه‌های آدلاین و مادلاین از جمله شبکه‌های پیش‌خور هستند. در بیشتر پژوهش‌های انجام‌شده در حوزه سیستم‌های غیرخطی از سیستم‌های پیش‌خور مثل MLP استفاده شده است. این شبکه‌ها قادرند با انتخاب مناسب تعداد لایه‌ها و سلول‌های عصبی که اغلب زیاد هم نیستند، یک نگاشت غیرخطی را با دقت دلخواه انجام دهند. از قانون یادگیری پس‌انتشار خطا برای آموزش شبکه‌های عصبی چندلایه پیش‌خور استفاده می‌شود. روش یادگیری پس انتشار خطا مبتنی بر قانون یادگیری اصلاح خطا است [۲۳].

1. ANN
2. MLP

۴. تحلیل داده‌ها و یافته‌های پژوهش

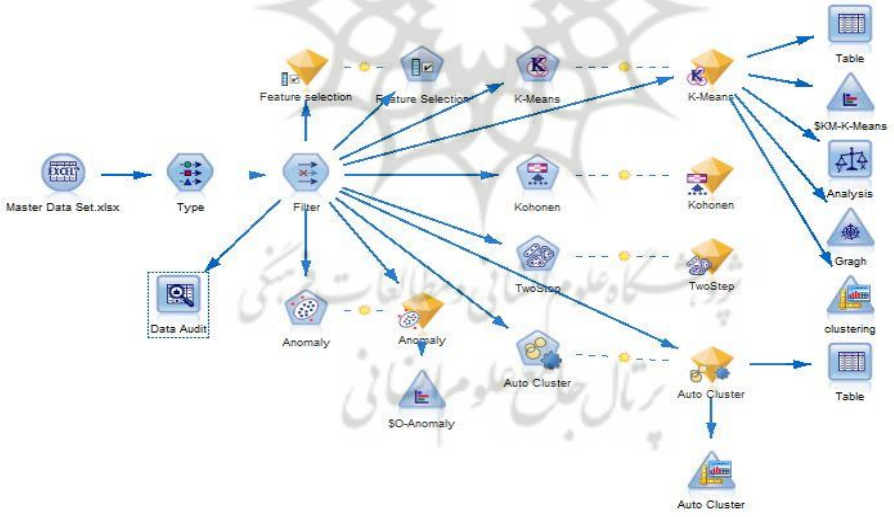
شناسایی ریسک‌ها و تشکیل پایگاه داده. شناسایی ریسک‌ها و خطرها بر اساس ارزیابی ریسک‌های صورت‌گرفته و برنامه‌های نگهداری و تعمیرات اجراشده قبلی توسط متخصصان در بازه زمانی سه‌ساله در صنعت موردمطالعه است. این داده‌ها مربوط به ۲۵۰ کیلومتر از خط لوله گاز اهواز به سنندج است که از اجرای برنامه‌های نت و ارزیابی ریسک به‌دست آمده است و مجموعه داده اولیه را تشکیل می‌دهد. در این راستا بررسی مستندات موجود، بررسی نتایج ارزیابی ریسک‌های صورت‌گرفته قبلی و گزارش‌های تدوین‌شده در ممیزی‌ها و به همراه ارزیابی پیامدها و نتایج احتمالی ناشی از وقوع هر یک از نقص‌ها و خطرها، از اقدامات اصلی انجام‌شده است. این موارد شامل عوامل ریسک و فهرست خطرها، حالت‌های نقص و خرابی، نرخ فراوانی، شدت نقص، نرخ کشف، عدد اولویت ریسک، احتمال وقوع نقص، تعداد خرابی، میانگین زمانی وقوع نقص و برآورد کلی اهمیت نقص بر اساس سطح هزینه نگهداری و تعمیرات ناشی از آن است (جدول ۲).

جدول ۲. ریسک‌ها و نقایص شناسایی‌شده خط لوله گاز

ردیف	عنوان نقص/ریسک	ردیف	عنوان نقص/ریسک
۱	ناکارآمدی تجهیزات کنترل فشار و نرخ جریان	۱۶	نشت میعانات گازی
۲	تنش در دیواره لوله‌های انتقال	۱۷	نوبت کاری نامنظم کارگران
۳	نقص در حفاری‌های ضروری		عوامل زیستی
۴	خوردگی لوله	۱۸	آبیاری ناایمن اراضی کشاورزی اطراف خط لوله
۵	تراکم جمعیت در منطقه	۱۹	عمق کم لوله
۶	خطای انسانی	۲۰	نقص در عایق‌کاری شبکه انتقال
۷	ناکارآمدی سامانه حفاظت کاتدیک خطوط	۲۱	کمبود امکانات و سیستم‌های اعلام و اطفای
۸	انبارش بیش از حد موردنیاز تجهیزات، قطعات، مواد اولیه قابل‌اشتعال و کپسول گازهای قابل‌انفجار و جابه‌جایی و انتقال ناایمن	۲۲	عدم‌آموزش مناسب و توجیه‌نبودن پیمانکاران نسبت به موقعیت محیطی خط لوله
۹	انتشار گازهای آلاینده و مواد سمی شیمیایی	۲۳	شرایط زیست‌محیطی نامناسب (رطوبت، آب‌وهوا، پوشش گیاهی و جانوری، اسیدی‌بودن خاک، دما و مقاومت خاک و غیره)
۱۰	کیفیت پایین و نقص شیرهای قطع جریان	۲۴	نبود فشارسنج حساس در طول خط لوله
۱۱	استفاده از مواد اولیه بی‌کیفیت در تجهیزات	۲۵	گزیدگی در عملیات حفاری
۱۲	خرابکاری (آسیب عمدی شخص ثالث)	۲۶	نقص مانیتورینگ
۱۳	آسیب مکانیکی (برخورد ماشین‌آلات مانند اسکریپر، پیکور، بلدوزر و غیره) با خط لوله	۲۷	خطر آسیب‌های مزمن ناشی از عوامل فیزیکی مانند صوت و ارتعاش ماشین‌آلات و تجهیزات
۱۴	چیدمان ناایمن تجهیزات کارگاه	۲۸	سقوط در عملیات لوله‌گذاری
۱۵	مدیریت پسماند نامناسب	۲۹	عوامل طبیعی (سیل، زلزله، رانش زمین و ...)

اطلاعات مورداستفاده در پژوهش، داده‌های تاریخی مربوط به اجرای برنامه‌های نگهداری و تعمیرات در بازه زمانی سه‌ساله در صنعت موردمطالعه است. این داده‌ها مربوط به ۲۵۰ کیلومتر از خط لوله گاز اهواز به سنندج است که از اجرای برنامه نت مبتنی بر قابلیت اطمینان به‌دست آمده است و مجموعه داده اولیه را تشکیل می‌دهد. پایگاه داده ایجادشده متشکل از ۲۴۰ نمونه (رکورد) و ۱۵ ویژگی (خصیصه) است که ۱۱ منطقه شامل ۴۰ ناحیه را در محیط موردمطالعه شامل می‌شود. منطقه‌بندی انجام‌شده بر اساس ناحیه‌بندی صورت‌گرفته قبلی برای اجرای برنامه‌های مدیریت و ارزیابی ریسک است. در پایگاه داده، هر یک از نقص‌ها و ریسک‌ها که اعضای مجموعه داده هستند، برای رعایت اختصار، کدگذاری و برچسب‌زنی شده است. برای مثال، کد B2F4 مربوط به نقص ۴ از ناحیه B منطقه ۲ شبکه انتقال خط لوله گاز است.

داده‌کاوی پژوهش. در آغاز فرایند داده‌کاوی، ابتدا ایجاد جریان داده[1] و انجام عملیات پیش‌پردازش و آماده‌سازی داده‌ها صورت گرفته است. برای این منظور آزمون نرمال‌بودن داده‌ها و استفاده از گره‌های Feature Selection, Data Audit, Anomaly Detection, Filtering, Type از اقداماتی است که انجام شده و تشخیص داده‌های مفقوده، تشخیص داده‌های پرت، انتخاب ویژگی‌های اصلی، تلفیق داده‌ها و کاهش ابعاد و درنهایت پاک‌سازی داده‌ها و ایجاد مجموعه داده تمیز برای خوشه‌بندی و طبقه‌بندی، نتایج این بخش است. برای انجام خوشه‌بندی داده‌ها نیز از الگوریتم‌های Two-Step، Kohnen و K-Means استفاده شده است. شکل ۵، مدل داده‌کاوی را در مرحله خوشه‌بندی نشان می‌دهد.



شکل ۵. مدل داده‌کاوی در مرحله خوشه‌بندی

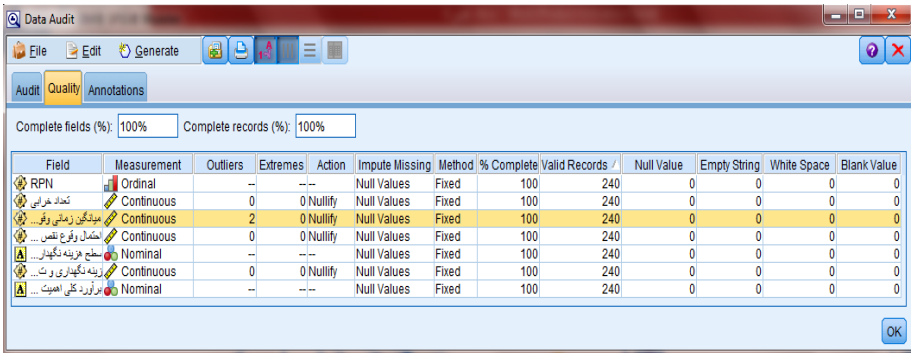
1. Data Stream

همان‌طور که در مدل شکل ۵، نشان داده شده است، پس از واردکردن و بارگذاری مجموعه داده، نوع داده موردبررسی با توجه به هر ویژگی مشخص می‌شود. نوع داده می‌تواند اسمی، ترتیبی، گسسته، پیوسته، دوتایی، مقادیر مجموعه‌ای یا مواردی باشد که به‌صورت خودکار توسط نرم‌افزار انتخاب شود. ارزش و مقادیر هر یک از ویژگی‌ها و محدوده‌ای که ویژگی‌ها در این محدوده دارای مقادیر معتبر هستند، مشخص شده و مقادیر مفقوده تعیین می‌شود. در این پژوهش مقادیر خارج از محدوده، بدون اثر[1] در نظر گرفته شده است. تعیین متغیرهای ورودی و خروجی هر یک از مدل‌های مورداستفاده برای داده‌کاوی، از موارد دیگری است که به آن پرداخته می‌شود. با استفاده از گره Filter می‌توان هر یک از ویژگی‌ها را تغییر داد و یا حذف کرد. با توجه به ماهیت هر ویژگی، مواردی از قبیل شماره سیستم، اجزای سیستم و حالت نقص که صرفاً جنبه برچسب‌گذاری داشته‌اند، حذف شده‌اند؛ همچنین پس از محاسبه عدد اولویت ریسک ویژگی‌های نرخ فراوانی، شدت نقص و نرخ کشف نیز فیلتر شده است. در ادامه با استفاده از گره Feature Selection ورودی‌های با ارزشِ بیش از ۹۵ درصد و با اهمیت بالا تعیین شده است. بر این اساس «هزینه نگهداری و تعمیرات»، «برآورد کلی اهمیت نقص»، «احتمال وقوع نقص»، «عدد اولویت ریسک»، «تعداد خرابی» و «میانگین زمانی وقوع نقص» برای تجزیه‌وتحلیل آماری و ارزش‌گذاری با استفاده از این گره انتخاب شده‌اند و آزمون مربوط به نرمال‌بودن داده‌ها در مورد آن‌ها انجام شده است. برای این منظور شاخص‌های «چولگی» و «کشیدگی» محاسبه و ارزیابی شده است. مبنای نرمال‌بودن داده‌ها بر اساس این دو شاخص قرارگرفتن اعضای مجموعه داده در بازه (۲+ و۲−) است. جدول۳، بر این اساس تمامی داده‌ها، موردارزیابی قرار گرفته و دو مورد داده پرت در ویژگی «میانگین زمانی وقوع نقص» تشخیص داده شده است که اقدام بی‌اثرسازی در مورد آن انجام شده و با استفاده از روش Fix که در آن مقادیر ثابت مانند میانگین و میانه یا هر مقدار ثابت دیگر جایگزین می‌شود، نسبت به آن رفتار شده است. شکل ۶، نتایج مربوط به ممیزی داده‌ها پس از انجام اقدامات مرتبط با آماده‌سازی و پیش‌پردازش داده‌ها را نشان می‌دهد.

جدول ۳. مقادیر پارامترهای آماری

ویژگی	مد	چولگی	کشیدگی	انحراف معیار	میانگین	واریانس
تعداد خرابی	۱۲	−۱/۱۵۴	−۰/۰۵۲	۳/۸۸۴	۷/۹۲۵	۱۵/۰۸۲
RPN	۷۲	−۰/۸۲۱	۱/۵۵۴	۱۹۶/۷۷۶	۱۹۵/۲۷۹	۳۸۷۲۰/۷۷۱
MTBF(min)	۶۵۶۶۷۲	۱/۲۲۵	۰/۵۷۸	۳۵۷/۳۳۵	۸۴۱۸۹۵/۸۳۷	۱۲۷۸۱۵۰
FMP (%)	۵۸	−۱/۰۱۰	−۰/۱۱۷	۲۱/۴۷۴	۴۵/۷۲۱	۴۶۱/۱۱۴
هزینه نت (ریال)	۳۷۰۶۱۱۶۶۵	۰/۰۴۴	۰/۶۱۴	۲۱۵۴۳۲	۳۴۲۷۶۸	۴۶۴۱۱۰۱۲۱۱

1. Nullify

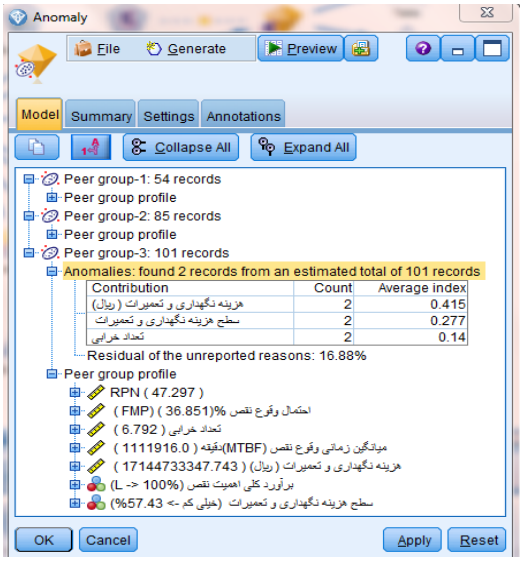


Field	Measurement	Outliers	Extremes	Action	Impute Missing	Method	% Complete	Valid Records	Null Value	Empty String	White Space	Blank Value
RPN	Ordinal	--	--	--	Null Values	Fixed	100	240	0	0	0	0
تعداد خرابی	Continuous	0	0	Nullify	Null Values	Fixed	100	240	0	0	0	0
میانگین زمانی وقو...	Continuous	2	0	Nullify	Null Values	Fixed	100	240	0	0	0	0
احتمال وقوع نقص ...	Continuous	0	0	Nullify	Null Values	Fixed	100	240	0	0	0	0
سطح هزینه نگهدار...	Nominal	--	--	--	Null Values	Fixed	100	240	0	0	0	0
زینه نگهداری و ت...	Continuous	0	0	Nullify	Null Values	Fixed	100	240	0	0	0	0
برآورد کلی اهمیت ...	Nominal	--	--	--	Null Values	Fixed	100	240	0	0	0	0

شکل ۶. نتایج گره Data Audit

در ادامه عملیات پیش‌پردازش داده‌ها، با استفاده از گره Anomaly، داده‌های پرت یا مقادیر غیرمعمول واکاوی شده است. این گره قسمتی از مدل کشف مغایرت است که اطلاعاتی راجع به ویژگی‌ها، تنظیمات و فرایند تخمین در مدل را نشان می‌دهد. مدل کشف مغایرت یک روش غیرنظارتی است و نیازی به مجموعه داده برای آموزش و یادگیری ندارد. در این مدل به‌منظور تعیین مغایرت‌های احتمالی هر رکورد با سایر رکوردها مقایسه و در گروه‌های همتا قرار می‌گیرد. روش مورداستفاده برای پذیرش یا عدم‌پذیرش مغایرت‌ها حد آستانه[1] است. میزان برش و حد آستانه به‌صورت خودکار تعیین و برحسب درصد به‌عنوان یک پارامتر در مدل محسوب می‌شود؛ همچنین زمینه‌هایی که به‌عنوان شاخص مغایرت باید مدنظر قرار گیرد، در تنظیمات اولیه مشخص می‌شود [۳]. بر اساس آن کشف مغایرت در سه گروه همتا با تعداد رکوردهای ۵۴، ۸۵ و ۱۰۱ صورت گرفته است که در گروه سوم ۲ رکورد به‌عنوان مقادیر مغایر تشخیص داده شده و شاخص میانگین با مقادیر مشخص جایگزین شده است. درنهایت نتایج نشان‌دهنده دو داده مغایر است که ۰/۸۳ درصد از کل رکوردها را شامل شده و ۹۹/۱۷ درصد داده‌ها معادل ۲۳۸ رکورد، عدم‌مغایرت داشته است (شکل ۷ و جدول ۴).

1. Cutoff

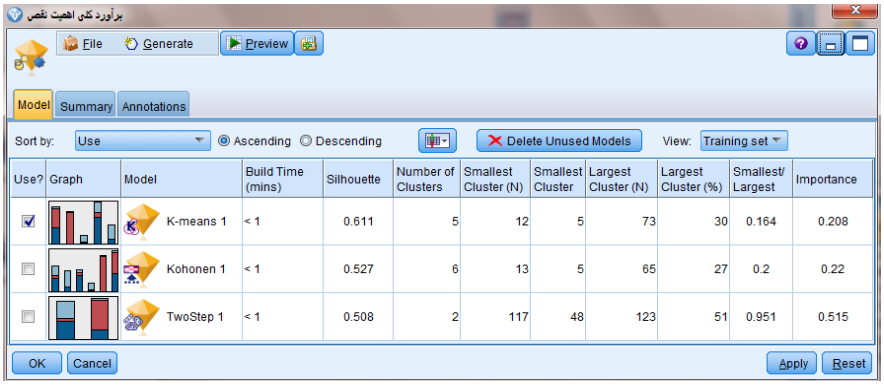


شکل ۷. نتایج گروه‌بندی الگوریتم Anomaly برای کشف مغایرت‌ها

جدول ۴. جدول نتایج اجرای الگوریتم Anomaly

ارزش	تعداد	درصد
مغایرت	۲	۰/۸۳
عدم مغایرت	۲۳۸	۹۹/۱۷

پس از انجام عملیات پیش‌پردازش داده‌ها و مشخص‌کردن ویژگی‌های اصلی و همچنین تعیین داده‌های پرت و مفقوده و جایگزین‌کردن آن‌ها، «شاخص سیلوئیت» برای انجام خوشه‌بندی به‌عنوان مبنا موردنظر قرار گرفته است و از سه الگوریتم Two Step، Kohnen و K-Means استفاده شده است که با تغییر مقادیر مربوط به تعداد خوشه (K)، بهترین مقدار برای این شاخص، مبتنی بر الگوریتم K-Means، ۰/۶۴۴۶ به‌دست آمد. برای سنجش اعتبار آن نیز از گره Auto Cluster بهره‌گیری شد که نتایج مشابه به‌دست آمده و منطبق بر شکل ۸، است. سایر ویژگی‌ها، نتایج و مختصات خوشه‌بندی با استفاده از این گره نیز در این شکل مشاهده می‌شود.

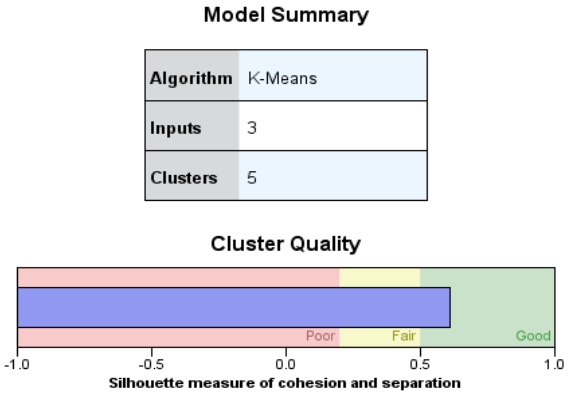


شکل ۸. نتایج خوشه‌بندی با گره Auto Cluster

بر اساس نتایج پیاده‌سازی الگوریتم‌ها، روش K-Means به‌عنوان روش برتر انتخاب و نتایج حاصل از پیاده‌سازی آن بررسی و تحلیل شد. بر این اساس داده‌ها در ۵ خوشه تقسیم‌بندی شد که تعداد اعضای هر یک از خوشه‌ها بر اساس نوع داده و درصد فراوانی آن‌ها در جدول ۵، مشخص شده است. شکل ۹ و جدول ۶، نشان‌دهنده محدوده قرارگیری شاخص سیلوئیت و مقدار آن است.

جدول ۵. فراوانی داده‌ها در هر خوشه

خوشه	تعداد اعضا	درصد
۱	۶۵	۲۷/۰۸
۴	۷۳	۳۰/۴۲
۵	۳۲	۱۳/۳۳
۳	۱۲	۵
۲	۵۸	۲۴/۱۷



شکل ۹. محدوده شاخص سیلوئیت

جدول ۶. مقدار شاخص سیلوئیت

مقدار شاخص سیلوئیت	محدوده
۰/۶۴۴۶	Good

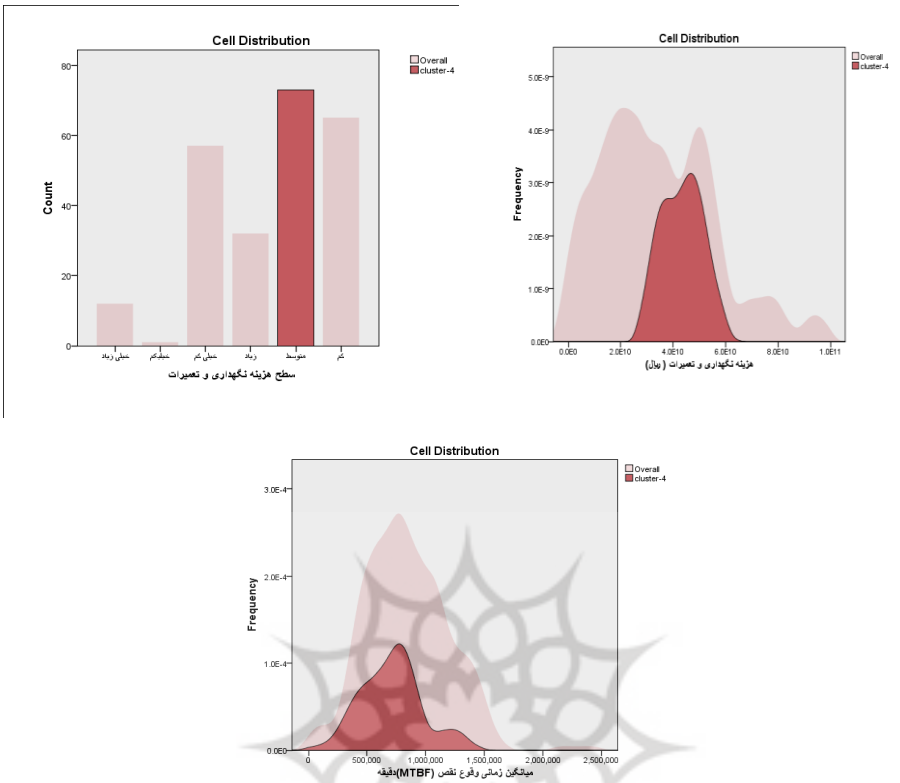
میزان اهمیت هر یک از ویژگی‌ها در الگوریتم k میانگین در جدول ۷، نشان داده شده است که بر این اساس، ویژگی‌های «سطح هزینه نگهداری و تعمیرات» و "«هزینه نگهداری و تعمیرات»، به‌ترتیب دارای بیشترین و کمترین اهمیت با مقادیر یک و ۰/۵۲۵۴ هستند.

جدول ۷. میزان اهمیت ویژگی‌ها در خوشه‌بندی با k-Means

ویژگی	اهمیت
میانگین زمانی وقوع نقص (MTBF) دقیقه	۰/۰۸۸۵
هزینه نگهداری و تعمیرات (ریال)	۰/۵۲۵۴
سطح هزینه نگهداری و تعمیرات	۱

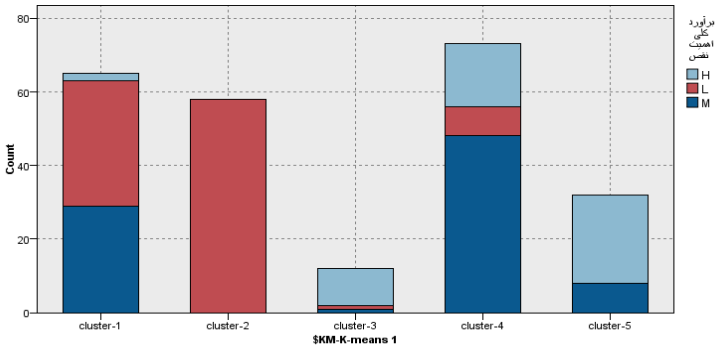
برای تشخیص خوشه بحرانی ویژگی‌های اصلی شامل هزینه نت، تعداد خرابی،عدد اولویت ریسک، MTBF و نرخ وقوع نقص مورد تجزیه‌وتحلیل آماری قرار گرفته است. برای این منظور با استفاده از گره Analysis تعداد اعضای هر خوشه، انحراف استاندارد، خطای استاندارد و میانگین هر یک از ویژگی‌ها در خوشه‌های پنج‌گانه بررسی شد. با توجه به تعداد اعضای بیشتر، انحراف استاندارد و درصد خطای کمتر در اغلب شاخص‌ها، میانگین عدد اولویت ریسک و تعداد خرابی نسبتاً بالا، میانگین وقوع نقص پایین، میانگین احتمال وقوع نقص نسبی تقریباً بالا به

نسبت اعضای خوشه، خوشه ۴ به‌عنوان خوشه بحرانی، انتخاب شد. شکل ۱۰، نشان‌دهنده وضعیت توزیع داده‌ها در خوشه ۴ نسبت به سایر خوشه‌ها بر اساس ویژگی‌های اصلی است.



شکل ۱۰. نمودار وضعیت توزیع داده‌ها در خوشه ۴ بر اساس سه ویژگی با اهمیت بالا

نمودار شکل ۱۱، وضعیت خوشه‌بندی‌ها بر اساس برآورد کلی اهمیت نقص را نشان می‌دهد. با توجه به نمودار، با درنظرگرفتن تعداد اعضا، تعداد ریسک با اهمیت متوسط و زیاد، خوشه ۴ را می‌توان خوشه بحرانی فرض کرد. اگرچه خوشه ۵ دارای ریسک با اهمیت بالای بیشتری است (۲۴ ریسک)، اما با تعداد ۳۲ عضو و ۸ ریسک با اهمیت متوسط به نسبت کل نقص‌های شناسایی‌شده و نیز بررسی سایر پارامترهای آماری که پیش‌تر ذکر شد، نمی‌توان آن را خوشه بحرانی دانست. اطلاعات مربوط به نمودار در جدول ۸، ارائه شده است.

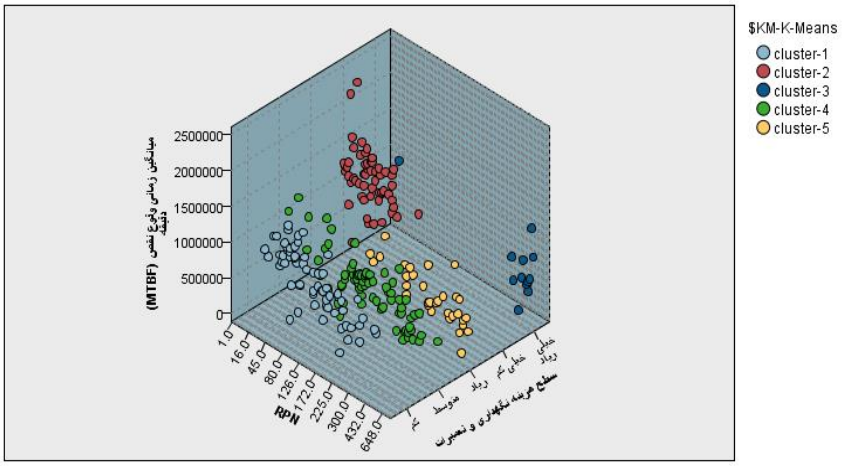


شکل ۱۱. نمودار میله‌ای خوشه‌های پنج‌گانه با توجه به اهمیت نقص

جدول ۸. اطلاعات مربوط به خوشه‌ها بر اساس اهمیت نقص

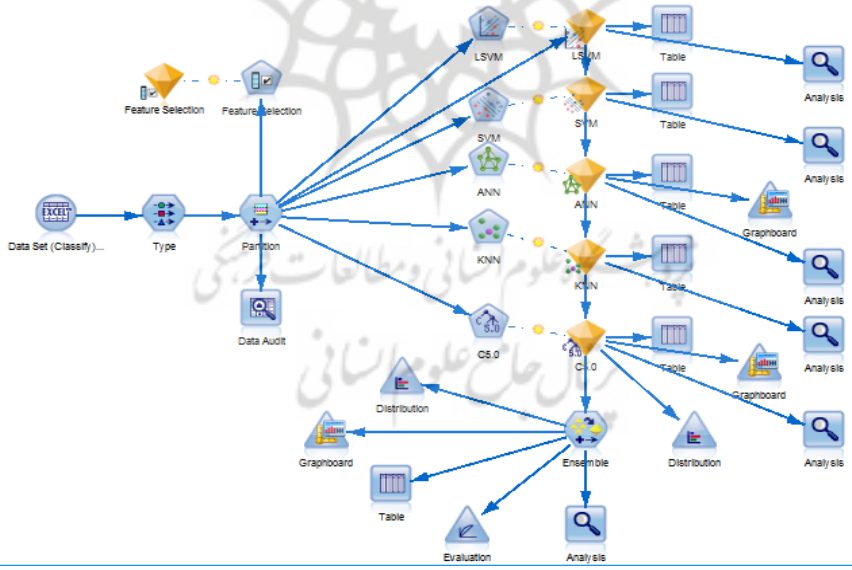
خوشه	۵		۴			۳			۲			۱	
تعداد	۸	۲۴	۸	۴۸	۱۷	۱	۱	۱۰	۵۸	۳۴	۲۹	۲	
اهمیت	M	H	L	M	H	M	L	H	L	L	M	H	

شکل ۱۲، نحوه خوشه‌بندی ریسک‌ها و نقص‌ها را با الگوریتم K-Means نشان می‌دهد. بر اساس سه ویژگی «عدد اولویت ریسک»، «سطح هزینه نگهداری و تعمیرات» و «میانگین زمانی وقوع نقص»، با توجه به شکل، خوشه ۴ دارای فراوانی بیشتر، سطح هزینه نت بین سطح زیاد و متوسط، RPN در مقادیر بالا دارای تراکم بیشتر و میانگین وقوع نقص نیز در بیشتر ریسک‌ها مقادیر کم را شامل می‌شود. درحالی‌که سطح هزینه نت در خوشه‌های دیگر نسبت به خوشه ۴ کمتر است؛ به‌جز خوشه ۳ که تعداد اعضای آن به نسبت سایر خوشه‌ها دارای فراوانی کمتر است و عناوین ریسک‌ها دارای جامعیت کافی برای بررسی نیستند. در خوشه ۱ بیشتر اعضا دارای عدد اولویت ریسک و سطح هزینه پایین هستند. در خوشه ۲ تقریباً همه اعضا هیچ‌یک از شرایط بحرانی‌بودن را ندارند.



شکل ۱۲. نتایج خوشه‌بندی با روش K-Means

در ادامه فرایند داده‌کاوی پژوهش از الگوریتم‌های شبکه عصبی، درخت C.5، نزدیک‌ترین همسایگی و بردار پشتیبان برای طبقه‌بندی بهره‌گیری شده است. این موارد از تکنیک‌های پرکاربرد در یادگیری ماشین و تشخیص الگوهای طبقه‌بندی داده‌ها هستند و در بسیاری از موارد تلفیق آن‌ها قدرت یادگیری را افزایش می‌دهد. شکل ۱۳، مدل داده‌کاوی را در مرحله طبقه‌بندی داده‌ها نشان می‌دهد.



شکل ۱۳. مدل طبقه‌بندی داده‌ها

در این مرحله اعضای خوشه بحرانی (خوشه ۴)، به‌عنوان مجموعه داده بارگذاری شده و با استفاده از گره Type تنظیمات اولیه در خصوص هر یک از ویژگی‌ها (نوع ویژگی، ورودی‌ها، خروجی و ارزش هر یک از ویژگی‌ها، بازه اعتبار، داده‌های مفقوده و پرت) مشخص می‌شود. در ادامه روند پیش‌پردازش داده‌ها، ویژگی‌های اصلی انتخاب شده‌اند. سه ویژگی «عدد اولویت ریسک»، «احتمال وقوع نقص» و «هزینه نگهداری و تعمیرات» به‌عنوان ویژگی‌های مهم و با اهمیت بالای ۹۵ درصد برگزیده شده‌اند. با استفاده از گره Data Analysis آزمون نرمال‌بودن داده‌ها مانند مرحله خوشه‌بندی صورت گرفت که برای رعایت اختصار از ارائه جدول‌ها در این بخش صرف‌نظر شده و به جدول‌های ارائه‌شده در فاز خوشه‌بندی اکتفا شده است. با توجه به اینکه در مدل طبقه‌بندی، ویژگی «برآورد کلی اهمیت نقص» به‌عنوان ویژگی هدف و خروجی انتخاب شده است، از گره «Partition» برای آموزش، آزمون و اعتبارسنجی مدل استفاده شده است. این گره برای ایجاد یک ویژگی بخش‌بندی‌شده به‌کار می‌رود که داده‌ها را به زیرمجموعه‌های جداگانه یا نمونه‌هایی برای آموزش، آزمون و اعتبارسنجی مدل جداسازی می‌کند. بر اساس بخشی از نمونه، مدل ایجاد می‌شود و بر اساس بخش‌های دیگری از نمونه‌ها مدل آزمون می‌شود؛ البته می‌توان بررسی کرد که مدل با چه کیفیتی می‌تواند به مجموعه داده‌های بزرگ‌تر که مشابه با داده‌های جاری هستند تعمیم داده شود [۴]. در مدل طبقه داده‌های پژوهش، ۷۰ درصد داده‌ها برای آموزش، ۲۰ درصد داده‌ها برای آزمون و ۱۰ درصد داده‌ها برای اعتبارسنجی در نظر گرفته شده است. در گام ابتدایی اجرای الگوریتم‌های طبقه‌بندی از الگوریتم ماشین بردار پشتیبان خطی و غیرخطی استفاده شده است که در مرحله نخست صحت طبقه‌بندی ۷۶/۸ درصد و پس از استفاده از مدل غیرخطی این الگوریتم این مقدار ۸۵/۷۱ درصد بوده است. برای الگوریتم SVM از تابع خطی نوع دوم و برای L-SVM، تابع سیگموئید با لاندای ۰/۱ و دقت رگرسیون ۰/۱ استفاده شده است. جدول ۹، ماتریس اغتشاش (درهم‌ریختگی) به‌دست‌آمده از پیاده‌سازی مدل L-SVM را نشان می‌دهد که بر اساس آن میزان صحت پیش‌بینی مدل ۷۷ درصد برآورد شده است.

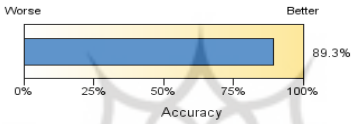
جدول ۹. ماتریس درهم‌ریختگی الگوریتم L-SVM

عنوان	پیش‌بینی صورت‌گرفته			
	میزان صحت	M	L	H
H	./۱۸	۹	۰	۲
L	۰	۴	۰	۰
M	۱	۴۱	۰	۰
میزان صحت	۰/۷۷	۰/۷۶	۰	۱

در ادامه اجرای مدل طبقه‌بندی، از شبکه عصبی پرسپترون چندلایه استفاده شده و تنظیمات اولیه شامل تعداد لایه‌ها و نورون‌ها، قوانین توقف الگوریتم، درصد داده‌هایی که در هر تکرار برای آموزش اختصاص داده می‌شوند و تصمیم در مورد داده‌های پرت، مشخص و مدل اجرا شده است. شکل ۱۴، ویژگی‌های الگوریتم و صحت طبقه‌بندی را نشان می‌دهد که این مقدار برابر ۸۹/۳ ٪ است و بنابراین استفاده از این الگوریتم در ادامه ماشین بردار پشتیبان، سبب ارتقای میزان صحت طبقه‌بندی داده‌ها شده است. میزان صحت پیش‌بینی مدل نیز بر اساس ماتریس درهم‌ریختگی در الگوریتم اجراشده ۹۶/۲ درصد است (جدول ۱۰).

Model Summary

Target	IPMORTANCE
Model	Multilayer Perceptron
Stopping Rule Used	Minimum accuracy exceeded
Hidden Layer 1 Neurons	2
Hidden Layer 2 Neurons	1

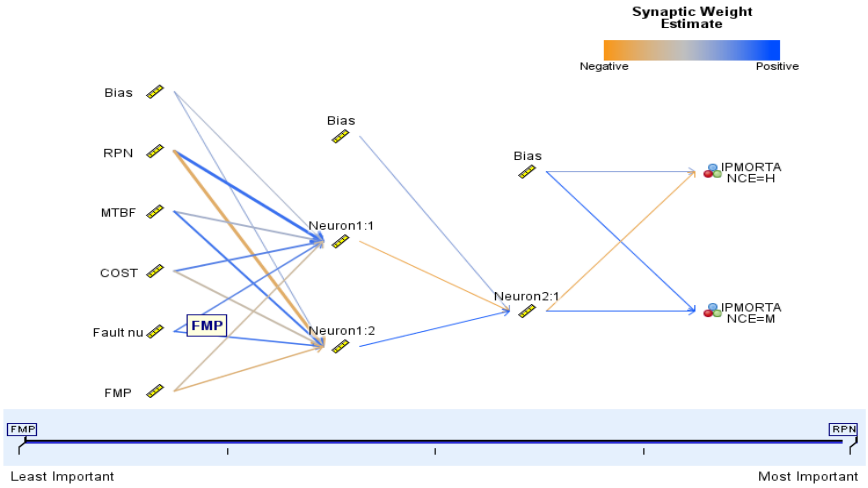


شکل ۱۴. ویژگی‌های الگوریتم و صحت طبقه‌بندی

جدول ۱۰. ماتریس درهم‌ریختگی الگوریتم ANN

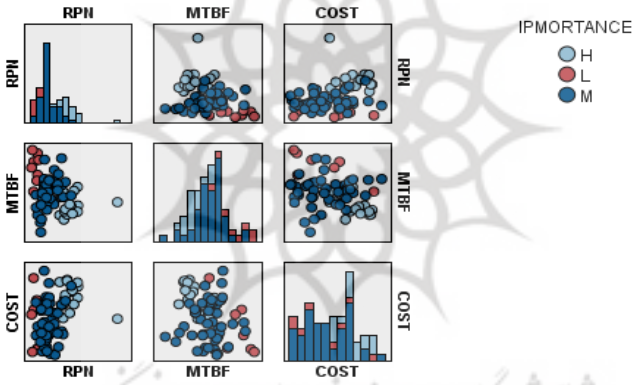
عنوان شناخته‌شده متغیر	پیش‌بینی صورت گرفته	
	M	H
H	٪ ۱۸/۲	٪ ۸۱/۸
M	٪ ۱۰۰	٪ ۰
میزان صحت	٪ ۹۶/۲	

شبکه عصبی استفاده‌شده از نوع پرسپترون دولایه است که در لایه اول دارای دو نورون و در لایه دوم دارای یک نورون است. ساختار شبکه عصبی مدل با درنظرگرفتن ۵ ورودی و انحراف و سطح کلی اهمیت نقص به‌عنوان خروجی به شکل ۱۵، است که در آن اهمیت وزن‌ها در طیف رنگی ارائه‌شده بر اساس میزان تأثیرگذاری مشخص شده است. در این ساختار ارتباطات آبی دارای بیشترین تأثیرگذاری و رنگ نارنجی، کمترین تأثیرگذاری را در یادگیری شبکه داشته است.



شکل ۱۵. ساختار شبکه عصبی پرسپترون چندلایه

نحوه طبقه‌بندی نقص‌ها در مدل پیاده‌سازی‌شده بر اساس ترکیب سه الگوریتم بالا، با توجه به سه ویژگی «عدد اولویت ریسک»، «احتمال وقوع نقص» و «هزینه نگهداری و تعمیرات» که دارای اهمیت بالا بوده‌اند، در نمودار ۱۶، نشان داده شده است.



شکل ۱۶. نمودار نتایج طبقه‌بندی داده‌ها با LSVM، SVM و ANN

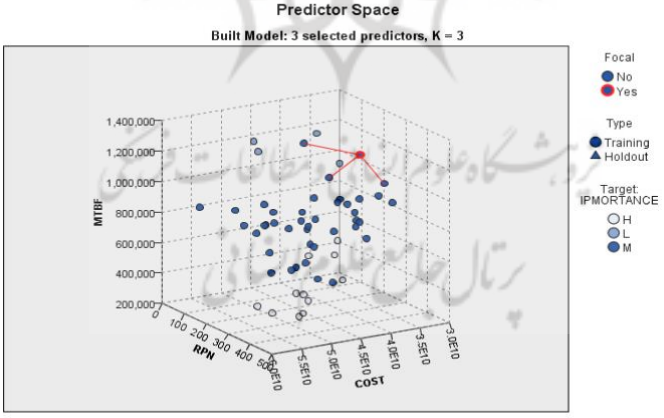
در این طبقه‌بندی ویژگی‌ها به‌صورت دوبه‌دو نسبت به یکدیگر قابل‌بررسی بوده و در سه سطح اهمیت نقص کم، متوسط و زیاد، طبقه‌بندی صورت گرفته است؛ همچنین نمودار هیستوگرام و چگونگی توزیع هر یک از داده‌ها بر اساس هر یک از ویژگی‌ها با درنظرداشتن ویژگی هدف در این نمودار ارائه شده است. نتایج به‌عنوان ورودی برای الگوریتم K نزدیک‌ترین همسایگی استفاده شده است. خروجی‌های حاصل از اجرای الگوریتم و محاسبه فواصل بر اساس

فاصله اقلیدسی، بهبود نتایج در دقت و صحت طبقه‌بندی را نشان می‌دهد؛ بدین ترتیب که پس از اجرای مدل، درصد داده‌های صحیح طبقه‌بندی‌شده ۹۱/۰۷ درصد است. در این الگوریتم حداقل k=3 و حداکثر K=5 در نظر گرفته شده است. در تکنیک K نزدیک‌ترین همسایگی، ویژگی‌های «عدد اولویت ریسک»، «احتمال وقوع نقص» و «هزینه نگهداری و تعمیرات» به‌ترتیب دارای بیشترین اهمیت هستند. در این تکنیک، فاصله اقلیدسی معیار محاسبه فواصل از یکدیگر است. برای مثال، اگر نقص ۲۶ از مجموعه داده‌ها انتخاب شود، همان‌طور که در جدول ۱۱ نشان داده شده است، نزدیک‌ترین نقاط به آن رکوردهای ۲۰، ۷ و ۴۶ هستند که به‌ترتیب مقادیر فاصله عبارت‌اند از: ۰/۲۱۶، ۰/۲۶۳ و ۰/۳۰۵.

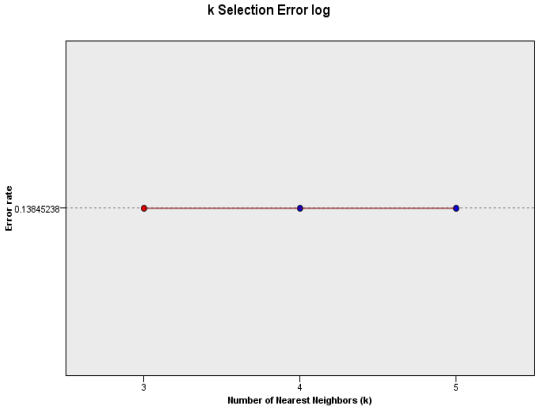
جدول ۱۱. نزدیک‌ترین داده‌ها به داده ۲۶ و فواصل

شماره داده	نزدیک‌ترین فواصل			نزدیک‌ترین نقاط همسایه		
	۳	۲	۱	۳	۲	۱
۲۶	۰/۳۰۵	۰/۲۶۳	۰/۲۱۶	۴۶	۷	۲۰

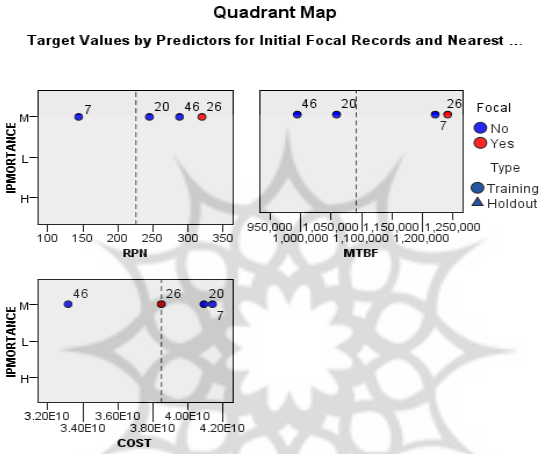
این فواصل بر حسب هر یک از ویژگی‌ها نیز در شکل ۱۷ که نمای سه‌بُعدی چگونگی پراکندگی داده‌ها در فضای پیش‌بینی مدل است، مشاهده می‌شود. برای مثال، نقص ۲۶ و نزدیک‌ترین نقاط به آن بر حسب فاصله اقلیدسی دارای عدد اولویت ریسک بین ۳۰۰ تا۳۵۰ هستند و همه این رکوردها سطح اهمیت متوسط دارند. این تفسیر برای ویژگی‌های میانگین وقوع نقص و هزینه نگهداری و تعمیرات نیز قابل‌ارائه است. میزان نرخ خطا در انتخاب K و فواصل اقلیدسی بر حسب K=3 در شکل‌های ۱۸ و ۱۹، ارائه شده است.



شکل ۱۷. نمای سه‌بعدی پراکندگی داده‌ها در فضای پیش‌بینی مدل



شکل ۱۸. نرخ خطا در انتخاب K



شکل ۱۹. فواصل اقلیدسی بر حسب $k=3$

ماتریس درهم‌ریختگی مربوط به الگوریتم K نزدیک‌ترین همسایگی در جدول ۱۲، مشاهده می‌شود و بر اساس آن ۸۷/۵ درصد میزان صحت پیش‌بینی مدل است و در طبقه‌بندی ۱۲/۵ درصد داده‌ها خطا وجود داشته است.

جدول ۱۲. ماتریس اغتشاش الگوریتم K نزدیک‌ترین همسایگی

عنوان شناخته‌شده متغیر	پیش‌بینی صورت گرفته			
	میزان صحت	M	L	H
H	٪۷۲/۷	۳	۰	۸
L	٪۵۰	۲	۲	۰
M	٪۹۵/۱	۳۹	۱	۱
میزان صحت	٪۸۷/۵	٪۷۸/۶	٪۵/۴	٪۱۶/۱
نرخ خطا در طبقه‌بندی داده‌ها				
٪۱۲/۵				

در ادامه اجرای مدل طبقه‌بندی داده‌های پژوهش، الگوریتم درخت تصمیم C.5 برای آموزش داده‌ها به مدل اضافه شده و نتایج نشان می‌دهد که صحت مدل طبقه‌بندی بر اساس درصد داده‌هایی که آموزش داده شده، ارتقا یافته است. این میزان پس از استفاده از الگوریتم درخت تصمیم C.5، به ۹۲/۸۶ درصد افزایش یافته است که میزان داده‌های صحیح طبقه‌بندی‌شده و صحت مدل را نشان می‌دهد. نتایج تحلیل مدل‌ها به‌صورت یکپارچه در جدول ۱۳، مشاهده می‌شود.

جدول ۱۳. نتایج تحلیل یکپارچه مدل ترکیبی طبقه‌بندی

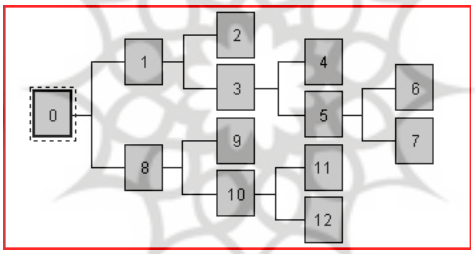
نتایج طبقه‌بندی با الگوریتم SVM						
اعتبارسنجی		آموزش		آزمون		Partition
٪ ۱۴/۲۹	۱	٪ ۶۰	صحیح	٪ ۷۶/۷۹	۴۳	صحیح
٪ ۸۵/۷۱	۶	٪ ۴۰	اشتباه	٪ ۲۳/۲۱	۱۳	اشتباه
	۷		مجموع		۵۶	مجموع
نتایج طبقه‌بندی با الگوریتم ترکیبی SVM و LSVM						
اعتبارسنجی		آموزش		آزمون		Partition
٪ ۷۱/۴۳	۵	٪ ۷۰	۷	٪ ۸۵/۷۱	۴۸	صحیح
٪ ۲۸/۵۷	۲	٪ ۳۰	۳	٪ ۱۴/۲۹	۸	اشتباه
	۷		۱۰		۵۶	مجموع
نتایج طبقه‌بندی با الگوریتم ترکیبی SVM، LSVM، ANN						
اعتبارسنجی		آموزش		آزمون		Partition
٪ ۷۱/۴۳	۵	٪ ۷۰	۷	٪ ۸۹/۲۹	۵۰	صحیح
٪ ۲۸/۵۷	۲	٪ ۳۰	۳	٪ ۱۰/۷۱	۶	اشتباه
	۷		۱۰		۵۶	مجموع
نتایج طبقه‌بندی با الگوریتم ترکیبی SVM ، LSVM، ANN، KNN						

Partition	آزمون		آموزش		اعتبارسنجی	
صحیح	۵۱	٪ ۹۱/۰۷	۷	٪ ۷۰	۶	٪ ۸۵/۷۱
اشتباه	۵	٪ ۸/۹۳	۳	٪ ۳۰	۱	٪ ۱۴/۲۹
مجموع	۵۶		۱۰		۷	

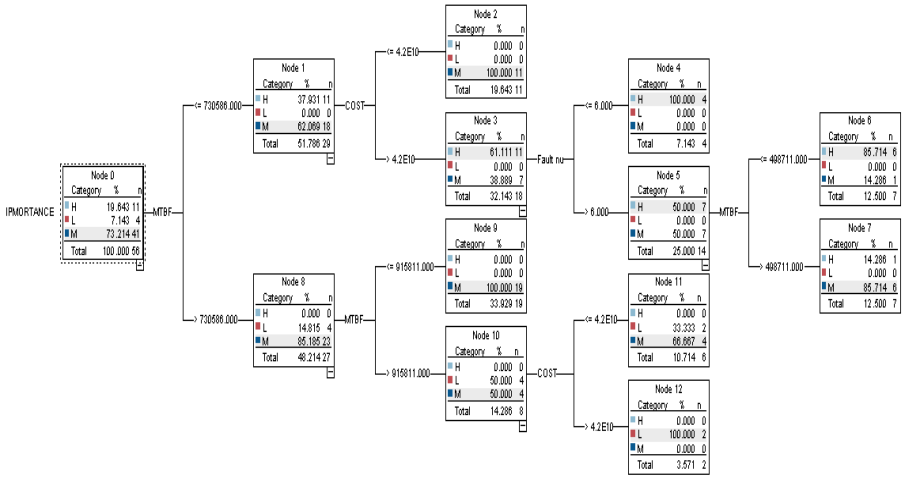
نتایج طبقه‌بندی با الگوریتم ترکیبی SVM، LSVM، ANN، KNN، C.5

Partition	آزمون		آموزش		اعتبارسنجی	
صحیح	۵۲	٪ ۹۲/۸۹	۵	٪ ۵۰	۴	٪ ۵۷/۱۴
اشتباه	۴	٪ ۷/۱۴	۵	٪ ۵۰	۳	٪ ۴۲/۸۶
مجموع	۵۶		۱۰		۷	

در تکنیک درخت C.5 ویژگی‌های «احتمال وقوع نقص»، «هزینه نگهداری و تعمیرات» و «تعداد خرابی» دارای بیشترین اهمیت است و مقادیر ۰/۶۴، ۰/۲۱ و ۰/۱۵ را به خود اختصاص داده‌اند. عمق درخت (تعداد سطوح) ایجادشده در مدل ۴ است و در هر برش ویژگی‌های بالا بر حسب اهمیت، مبنای برش بوده است. نمای کلی درخت و گره‌های ایجادشده با توجه به سطوح و عمق درخت در شکل ۲۰ و جزئیات آن که نتیجه پیاده‌سازی الگوریتم درخت تصمیم C.5 است، در شکل ۲۱، نشان داده شده است. در این شکل درصد فراوانی داده‌ها بر حسب ویژگی هدف درنظرگرفته‌شده (سطح اهمیت نقص)، مشخص شده است.

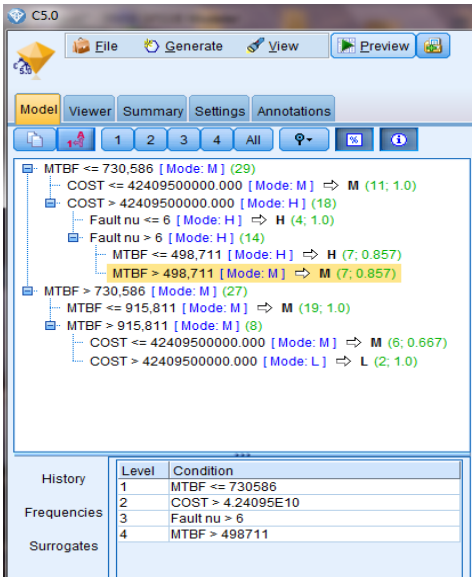


شکل ۲۰. نمای کلی درخت تصمیم



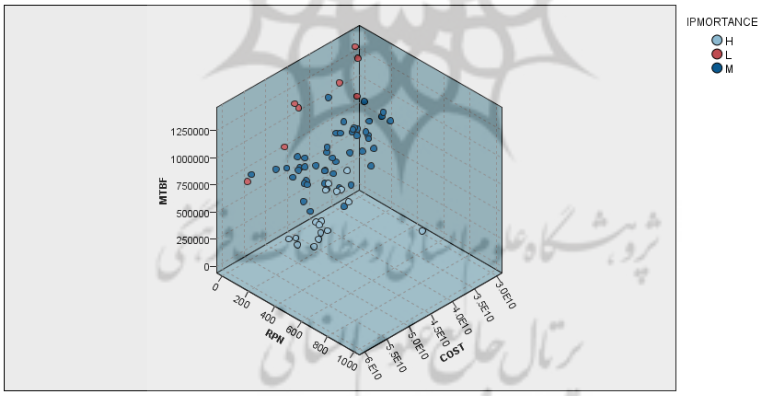
شکل ۲۱. ساختار درخت تصمیم و درصد فراوانی داده‌ها در هر گره

برای مثال، در برش اول درخت و در گره ۱ که تقسیم‌بندی آن بر اساس ویژگی MTBF با میزان بیشتر از ۷۳۰۵۸۶ و کمتر یا مساوی ۷۳۰۵۸۶ است، حدود ۳۸ درصد داده‌ها دارای اهمیت بالا، ۶۲ درصد دارای اهمیت متوسط و صفر درصد دارای اهمیت پایین هستند. مجموعه قوانینی که سبب تشکیل درخت تصمیم شده است بر اساس ویژگی‌هایی که از اهمیت بالایی برخوردار بوده‌اند، در شکل ۲۲، ارائه شده است. در هر شاخه درصد فراوانی و تعداد اعضای آن بر حسب اهمیت کلی نقص مشاهده می‌شود.

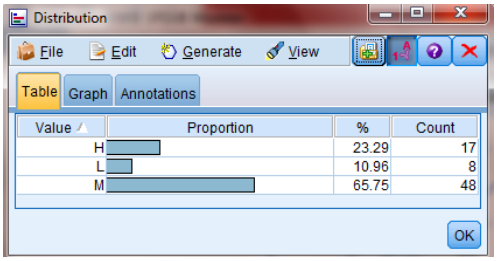


شکل ۲۲. مجموعه قوانین تشکیل‌دهنده درخت تصمیم

خروجی حاصل از طبقه‌بندی داده‌ها با استفاده از الگوریتم‌های استفاده‌شده در نمای سه‌بُعدی شکل ۲۳، نشان داده شده است و همان‌طور که مشاهده می‌شود، داده‌های با سطح اهمیت متوسط از تراکم بیشتری برخوردار هستند. بر این اساس تعداد اعضای هر سطح و درصد فراوانی آن در شکل ۲۴، نشان داده شده است.

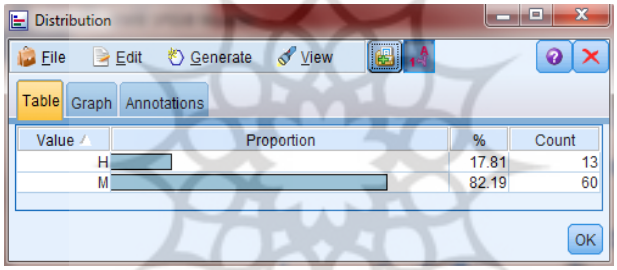


شکل ۲۳. نمای سه‌بُعدی طبقه‌بندی داده‌های پژوهش بر اساس ترکیب الگوریتم‌های طبقه‌بندی



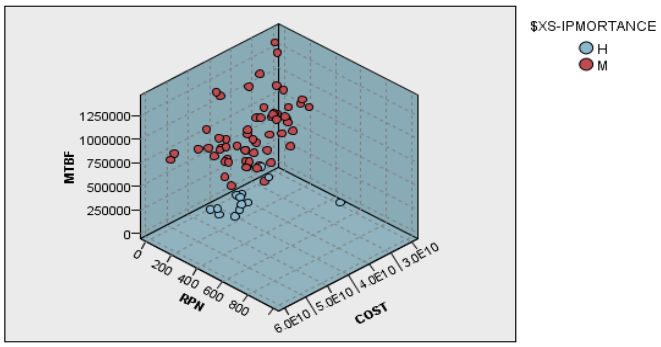
شکل ۲۴. فراوانی داده‌ها در هر طبقه بر اساس ترکیب الگوریتم‌های طبقه‌بندی

نتایج به‌دست‌آمده از الگوریتم‌های پیاده‌سازی‌شده با استفاده از گره Ensemble نرم‌افزار ادغام شده و خروجی مدل نشان‌دهنده افزایش صحت مدل است و این میزان برابر ۹۷/۵ درصد است. در این مدل بر اساس تحلیل‌های انجام‌شده، روش بالاترین درجه اطمینان برای ویژگی هدف (سطح کلی اهمیت نقص)، بالاترین میزان کارایی را داشته است. شکل ۲۴ و ۲۵، نتایج طبقه‌بندی داده‌های پژوهش را نشان می‌دهند که مبتنی بر آن، ۱۳ نقص دارای اهمیت بالا هستند و ۱۷/۸۱ درصد داده‌ها را شامل می‌شوند. ۶۰ نقص، معادل ۸۲/۱۹ درصد داده‌ها دارای اهمیت متوسط هستند.



شکل ۲۵. فراوانی بر اساس طبقه‌بندی نهایی داده‌ها

شکل ۲۶، نمای سه‌بعدی و نحوه پراکندگی داده‌ها در فضای پیش‌بینی بر حسب ورودی‌های و خروجی تعیین‌شده را نشان می‌دهد.



شکل ۲۶. نمای سه‌بُعدی طبقه‌بندی داده‌ها بر اساس طبقه‌بندی نهایی داده‌های پژوهش

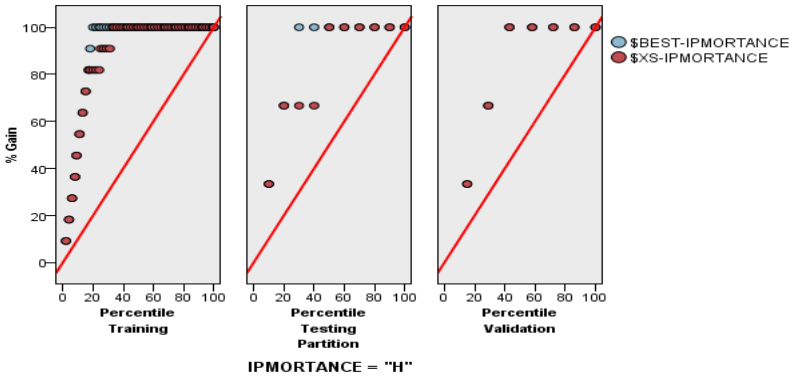
جدول ۱۴، نتایج تحلیل مدل‌ها پس از ادغام با استفاده از روش بالاترین درجه اطمینان در گره Ensemble برای داده‌های موردتوافق الگوریتم را نشان می‌دهد.

جدول ۱۴. نتایج تحلیل صحت و کارایی مدل نهایی طبقه‌بندی

داده‌های موردتوافق

Partition	آزمون		آموزش		اعتبارسنجی	
توافق	۴۱	٪ ۷۳/۲۱	۶	٪ ۶۰	۵	٪ ۲۸/۵۷
عدم‌توافق	۱۵	٪ ۲۶/۷۹	۴	٪ ۴۰	۲	٪ ۷۱/۴۳
مجموع	۵۶		۱۰		۷	
نتایج نهایی طبقه‌بندی با الگوریتم ترکیبی و Ensemble						
Partition	آزمون		آموزش		اعتبارسنجی	
صحیح	۴۰	٪ ۹۷/۵۶	۴	٪ ۶۶/۶۷	۱	٪ ۵۰
اشتباه	۱	٪ ۲/۴۴	۲	٪ ۳۳/۳۳	۱	٪ ۵۰
مجموع	۴۱		۶		۲	

شکل ۲۷، نشان‌دهنده روند یادگیری در داده‌های موردتوافق نهایی (۴۹ رکورد) است؛ به شکلی که یادگیری در ۹۷/۵۶ درصد از داده‌های آموزش (۴۱ رکورد) محقق شده است و این میزان در رابطه با داده‌های «اعتبارسنجی» و «آزمون» نیز قابل‌مشاهده است.



شکل ۲۷. روند یادگیری در داده‌های آموزش، آزمون و اعتبارسنجی

نقص‌های دارای سطح اهمیت بالا به همراه عنوان نقص و ارزش به‌دست‌آمده از مدل طبقه‌بندی در جدول ۱۵، نشان داده شده است.

جدول ۱۵. عنوان نقص‌ها/ ریسک‌های برگزیده

کُد	ارزش	عنوان نقص/ ریسک
E5F93	۰/۶۶۵	انتشار گازهای آلاینده و مواد سمی شیمیایی
C2F4	۰/۶۵۱	نشت میعانات گازی
A4F2	۰/۶۳۹	آسیب مکانیکی (برخورد ماشین‌آلات) با خط لوله
E5F140	۰/۵۸۲	نگهداری بیش از حد موردنیاز تجهیزات، قطعات، مواد اولیه قابل‌اشتعال و کپسول گازهای قابل‌انفجار و انتقال ناایمن
A4F7	۰/۵۷۷	ناکارآمدی تجهیزات کنترل فشار و نرخ جریان در شبکه انتقال
D3F4	۰/۵۶۹	خوردگی لوله
E5F39	۰/۵۶	کیفیت پایین و نقص شیرهای خودکار قطع جریان
E5F133	۰/۵۴۴	استفاده از مواد اولیه بی‌کیفیت در تجهیزات و قطعات
C3F2	۰/۵۳۹	ناکارآمدی سامانه حفاظت کاتدیک خطوط
D3F5	۰/۴۵۱	نقص در حفاری‌های ضروری
E5F91	۰/۴۲۵	نقص در عایق‌کاری شبکه انتقال
E5F151	۰/۴۳۱	نبود فشارسنج حساس در طول خط لوله
C2F8	۰/۳۴۱	عدم‌آموزش مناسب و توجیه‌نبودن پیمانکاران نسبت به موقعیت محیطی

بر اساس نتایج، ۱۳ ریسک فاکتور جزو نقص‌های بحرانی محسوب می‌شوند و به‌عنوان ریسک‌های دارای اولویت پیش‌بینی شده‌اند که در این میان «انتشار گازهای آلاینده و مواد سمی

شیمیایی» دارای بیشترین اولویت و «عدم‌آموزش مناسب و توجیه نبودن پیمانکاران نسبت به موقعیت محیطی خط لوله» دارای کمترین اولویت است. استفاده از روش‌های پیش‌بینی برای کشف ریسک‌های بحرانی و اولویت‌دار از ابزاری است که اگر به شکل دقیق و صحیح پیاده‌سازی شود، امر تشخیص ریسک‌ها را برای متخصصان تسهیل می‌کند و سبب استفاده از برنامه‌های نت صحیح، متناسب و مقرون‌به‌صرفه می‌شود.

۵. نتیجه‌گیری و پیشنهادها

امروزه پیاده‌سازی برنامه‌های صحیح مدیریت و ارزیابی ریسک و به‌تبع آن اجرای به‌موقع اقدامات اصلاحی کنترلی و فنی مهندسی، بخش مهمی از فعالیت‌های صنایع را شامل می‌شود که می‌تواند روند تولید را تحت تأثیر مستقیم قرار دهد؛ ازاین‌رو به‌کارگیری ابزارها، روش‌ها و رویکردهای نوین برای پایش برنامه تولید و پیشگیری از بروز رویدادهای ناخواسته بسیار مهم است. از این رهگذر، بهره‌مندی از ابزار داده‌کاوی برای تشخیص و پیش‌بینی ریسک‌ها و موقعیت‌های با ظرفیت ایجاد حوادث، موضوعیت می‌یابد. در این پژوهش تلاش شده است تا با بهره‌گیری از الگوریتم‌ها و تکنیک‌های تحلیل و پردازش داده بر اساس دو رویکرد داده‌کاوی توصیفی و داده‌کاوی پیش‌بینی، ریسک‌های بحرانی در خط لوله انتقال گاز، پیش‌بینی شده و تشخیص داده شود. بر این اساس از داده‌های تاریخی مربوط به پیاده‌سازی برنامه‌های نگهداری و تعمیرات و سوابق ارزیابی ریسک در بازه زمانی سه‌ساله در صنعت موردمطالعه استفاده شد. برای انجام عملیات داده‌کاوی نرم‌افزار SPSS Modeler 18.0 به‌کار رفت. در ابتدا فعالیت‌های مربوط به پیش‌پردازش و آماده‌سازی داده‌ها انجام شده و در ادامه برای خوشه‌بندی داده‌ها از الگوریتم‌های K-Means، Kohnen و Two-step و برای طبقه‌بندی از الگوریتم‌های شبکه عصبی، درخت C.5، نزدیک‌ترین همسایگی و بردار پشتیبان استفاده شد و نتایج خوشه‌بندی و طبقه‌بندی اعتبارسنجی شدند. درنهایت پس از اجرای مدل و تفسیر الگوریتم‌ها، نتایج تصویرسازی شد. نتایج پژوهش حاکی از آن است که ۱۳ ریسک فاکتور جزو نقص‌های بحرانی محسوب شده و به‌عنوان ریسک‌های دارای اولویت پیش‌بینی شده‌اند که در این میان «انتشار گازهای آلاینده و مواد سمی شیمیایی» دارای بیشترین اولویت و «عدم‌آموزش مناسب و توجیه‌نبودن پیمانکاران نسبت به موقعیت محیطی خط لوله» دارای کمترین اولویت است. با توجه به نتایج پژوهش و ماهیت نقص‌ها و رویدادهای بحرانی شناسایی‌شده، بروز حوادث و ایجاد شرایط خطر را می‌توان ناشی از تصادم و آسیب‌های مکانیکی دانست که بر این اساس ضرورت اقدامات کنترلی مهندسی و همچنین توجه به این موارد به‌منظور پیشگیری را در فاز طراحی ایجاب می‌کند. پیش‌بینی و شناسایی رویدادها و نواقص در سیستم‌های نت و درنظرگرفتن اقدامات کنترلی مهندسی متناسب با هر نقص قبل از وقوع خرابی‌ها، تأثیر بسزایی در کاهش

هزینه‌های نت دارد. با توجه به اینکه شناسایی ریسک‌های بحرانی، کاهش هزینه‌ها در انجام اقدامات اصلاحی و برنامه‌های نگهداری و تعمیرات را به‌دنبال خواهد داشت، دقت و صحت در استفاده از ابزار و روش‌های پیش‌بینی حائز اهمیت است. با درنظرداشتن این نکته که شبکه‌های انتقال گاز از صنایع حادثه‌خیز است، استفاده از روش‌های پیش‌بینی و رویکردهای پیش‌نگر توسط متخصصان مدیریت و ارزیابی ریسک، سبب می‌شود قبل از اینکه شرایط بالقوه خطر، بالفعل و تبدیل به حادثه شود و خسارت در پی داشته باشد، اقدامات کنترلی لازم صورت گیرد و این موضوع در مرحله طراحی سیستم‌ها باید موردتوجه قرار گیرد. آنچه در این پژوهش بر خلاف سایر پژوهش‌ها موردتوجه قرار گرفته، آموزش داده‌ها با یادگیری ماشین است؛ در این پژوهش الگوریتم ترکیبی پیش‌بینی برای بهینه‌سازی طبقه‌بندی داده‌ها به‌صورت تکاملی به‌کارگیری شده و در هر مرحله، هدف تقویت میزان صحت و اعتبار مدل طبقه‌بندی و افزایش میزان یادگیری داده‌ها است. درحالی‌که سایر پژوهش‌های مشابه بر رویکرد مقایسه‌ای در به‌کارگیری تکنیک‌ها متمرکز بوده‌اند و کارایی و صحت آن‌ها در مقایسه با یکدیگر موردقضاوت و ارزیابی قرار می‌گیرد؛ همچنین استفاده هم‌زمان از رویکردهای توصیفی و پیش‌بینی داده‌کاوی برای دستیابی به هدف پژوهش و تلفیق ویژگی‌ها و ابعاد دو مقوله «نگهداری تعمیرات» و «ارزیابی ریسک» برای جامعیت پایگاه داده و پیشگیری از خطا در انتخاب ویژگی‌های اصلی و کاهش ابعاد داده‌ها از موارد دیگری است که موردتوجه قرار گرفته است. به‌روزنبودن داده‌های مربوط به برنامه‌های نگهداری تعمیرات و سوابق ارزیابی ریسک‌های صورت‌گرفته و همچنین عدم‌امکان دسترسی به داده‌های جدید از محدودیت‌های اصلی پژوهش به‌شمار می‌رود. برای مطالعات آتی پیشنهاد می‌شود نقص‌ها و رویدادهای شناسایی‌شده در این پژوهش با استفاده از روش‌های بهینه‌سازی ترکیبی مبتنی بر رویداد و داده‌محور و بهره‌گیری از روش‌های قطعی و غیرقطعی بهینه‌سازی، مانند الگوریتم‌های متاهیوریستیک، برنامه‌ریزی عدد صحیح، برنامه‌ریزی آرمانی و نظریه بازی‌ها، بررسی و تحلیل شده و مدل‌سازی ریاضی و بهینه‌سازی‌ها در این زمینه با تمرکز بر شاخص‌های نت و مدیریت ریسک انجام شود. استخراج قوانین انجمنی با استفاده از داده‌کاوی در زمینه مدیریت ریسک می‌تواند از پژوهش‌های آتی در این زمینه باشد؛ همچنین پیشنهاد می‌شود روش‌های داده‌کاوی در چارچوب استانداردهای مدیریت فرایندهای پروژه با تمرکز بر فرایند مدیریت ریسک پروژه برای طبقه‌بندی ریسک‌ها یا زیرفرایندهای سیستم استفاده شود. برای توسعه مدل، می‌توان از این رویکرد در سایر پروژه‌ها نیز استفاده و نتایج را بررسی کرد؛ اگرچه تفاوت در ماهیت ریسک‌های شناسایی شده و ویژگی داده‌ها از محدودیت‌های توسعه مدل است.

منابع

1. Abdel-aziz, I. H., & Helal, M. (2012). Application of FMEA- FTA in Reliability-Centered Maintenance Planning. *15th International conferance on Applied Mechanics and Mechanicals Engineering.* Egypt.
2. Alborzi, M. (2013). *Neural Networks*. Publications of Sharif University of Technology, 5Ed, Tehran, Iran (In Persian).
3. Alikhanzadeh, A. (2012). *Data Mining.* Oloom Computer, 3Ed, Babol, Iran (In Persian).
4. Alizadeh, S., & Malek Mohammadi, S. (2013). *Data mining and knowledge discovery step by step with Clementine software*. Publications of Khwaja Nasiruddin Toosi University of Technology, 3Ed, Tehran, Iran (In Persian).
5. Ameri, H., Alizadeh, S., & Barzegari, A. (2012). Extracting knowledge from the data of diabetic patients using decision tree method C5. *Health Management*, 16(*53*), 58-72. (In Persian)
6. Ardeshir, A., Amiri, M., & Mohajer, M. (2012). Assessing safety risks in mass production projects using the combination of fuzzy AHP-DEA and fuzzy FTA methods, FMEA. *Bimonthly work health of Iran*, *10(6)*, 78-91. (In Persian)
7. Bashiri Nasab, M., Gholamreza, A., & Farzaneh, S. (1389). *Safety Management*. Fanavaran Publishing, 1Ed, Tehran, Iran (In Persian).
8. Chang, H. H., & Tsay, S. F. (2009). Integrating of SOM and K-mean in data mining clustering: An empirical study of CRM and profitability evaluation.
9. Chang, W., Meng, T., & Lim, C. (2015) .Clustering and visualization of failure modes using an evolving tree. *Expert Systems with Applications*, *42*, 7235–7244.
10. Chi, C. F., Sigmund, D., & Octavianus, M. (2020). Classification Scheme for Root Cause and Failure Modes and Effects Analysis (FMEA) of Passenger Vehicle Recalls. *Reliability Engineering and System Safety*, *200*, 906-929.
11. Dermohammadi, S., Alizadeh, S., Asghari, M., & Shami, M. (2013). Providing a predictive model for diagnosing infertility factors; Using data mining algorithms. *Health Management*, *17(57)*, 46-57. (In Persian)
12. Fahmi Hassan, A., Mughari, M., & Obadai, O. (1397). Prediction of blood donation using data mining based on decision tree algorithms, KNN, SVM, MLP. *Journal of engineering management and soft computing*, *4(1)*, 77-97. (In Persian)
13. Ghasemi, Sh., Yavari, K., Mahmoud-Vand, R., & Sahabi, B. (2013). Comparison of two different viewpoints in the application of FMEA method for risk assessment: a case study of Iranian gas refinery. *Quarterly journal of energy economy studies*, *10(42)*, 159-135 (In Persian).
14. Ghazanfari, M., Alizadeh, S., & Timurpur, B. (2015). *Data mining and knowledge discovery*. University of Science and Technology Publication. 5Ed, Tehran, Iran (In Persian).
15. Ghodoosi, M., Mirsaeedi, F., & Hasani, A. (2020). Presentation of Risks Analysis Model in Urban Projects Based on Data Mining Technique with Case Study. *The Journal of Industrial Management Perspective*, *10*, 137-159. (In Persian)
16. Habibi, E., & Alizadeh, M. (1386). *Functional safety and performance indicators in industry*. Fanavaran Publishing, 2Ed, Tehran, Iran. (In Persian)
17. Hamta, N., Ghobadi, Sh., & Boalhosni, P. (2017). Improving reliability in continuous manufacturing industries by applying data mining, FMEA and FTA

approaches. The first business management and optimization systems conference. Babol. Iran (In Persian).
18. Hosseini al-Madwari, M., Moghdasi, M. & Shafiizadeh Bafghi, M. (1390). Risk assessment by FMEA method and comparison of RPN before and after corrective measures in Bafq steel direct revitalization project. *The 7th national conference on occupational health and safety*. (In Persian)
19. Jafari Naimi, M. (1390). Clustering of digestive functional disorders. *Master's thesis*, Technical and Engineering Faculty, Isfahan University. (In Persian)
20. Mohaghar, A., & Khorasani, A. (2020). Designing the Model of Assessing the Risk management of Automaker Companies in Iran: Grounded Theory. *The Journal of Industrial Management Perspective, 10*, 137-159 (In Persian).
21. Moniri, M., Alem-Tabriz, A., & Ayough, A. (2022). Upstream Oil Process Plants Turnaround Projects Risk Evaluation Using a Hybrid Fuzzy MADM Method. *The Journal of Industrial Management Perspective, 12*, 135-173 (In Persian).
22. Morovati Sharif Abadi, A., Zanjirchi, M., & Abbasabadi, O. (2022). Processes Management of Maintenance using PCF and Data Mining. *The Journal of Industrial Management Perspective, 12*, 175-198. (In Persian)
23. Ramezani Beshli, P. (2011). Cluster analysis and analysis of foreshocks for earthquake prediction using data-mining techniques. *Master's Thesis*, Faculty of Earth Sciences, Shahrood University of Technology (In Persian).
24. Ravanbakhsh, S. (2017). Improving the efficiency of strategic equipment with the method of maintenance, damage analysis and simulation. *Journal of maritime transportation industry*, 4(3), 11-20. (In Persian).
25. Sabet Mutlaq, M., Ayazi, A., & Hosseini Dashiri, Jalaluddin. (2016). Presenting a hybrid approach to evaluate and rank failure modes using modified FMEA and fuzzy hierarchical analysis process. Case study: A company producing gears and industrial gearboxes active in Qom. *Scientific and Promotional Quarterly of Standard and Quality Management, 7(3), 19-30.* (In Persian).
26. Sheikhinejad, Z., Naderi Dehkordi, M., & Rastgari, H. (2013). A review of privacy-preserving methods in transactional database security. *The second computer science and education conference*. (In Persian)
27. Souza, R., & Alvares, A. (2007). FMEA and FTA Analysis for Application of the Reliability Centered Maintenance Methodology. *19th International congress of Mechanical Engineering*. Brasilia.
28. Steenwinckel, B., De Paepe, D., Heyvaert, P., & Moens, P. (2021). FLAGS: A methodology for adaptive anomaly detection and root cause analysis on sensor data streams by fusing expert knowledge withmachine learning. *Future Generation Computer Systems, 116*, 30–48.
29. Yang, C., Zou, Y., Lai, P., & Jiang, N. (2015). Data mining-based methods for fault isolation with validated FMEA model ranking. *Springer*.