پژوهشگاه نیرو

NRI-ER-1135-PPUSPN05-01-F

بررسی کاربرد تحلیل داده در مدیریت انرژی الکتریکی صنایع

مدیر پروژه: حمیدرضا آراسته

واحد مجری: گروه پژوهشی برنامه‌ریزی و بهره‌برداری سیستم‌های قدرت

DOI: 10.30503/nripress.2020.202


بسم الله الرحمن الرحيم

بررسی کاربرد تحلیل داده در مدیریت انرژی الکتریکی صنایع

مدیر پروژه: حمیدرضا آراسته

گروه پژوهشی پشتیبان: گروه پژوهشی برنامه‌ریزی و بهره‌برداری سیستم‌های قدرت

واحد مجری: گروه پژوهشی برنامه‌ریزی و بهره‌برداری سیستم‌های قدرت

انتشارات پژوهشگاه نیرو

آذر ۱۴۰۲

کد ثبت گزارش: NRI-ER-1135-PPUSPN05-01-F

DOI: 10.30503/nripress.2020.202

ایران. وزارت نیرو. پژوهشگاه نیرو. گروه پژوهشی برنامه‌ریزی و بهره‌برداری سیستم‌های قدرت

بررسی کاربرد تحلیل داده در مدیریت انرژی الکتریکی صنایع/ مدیر پروژه: حمیدرضا آراسته. کارفرما: پژوهشگاه نیرو، ۱۴۰۲.

۹۱ ص.: مصور، نمودار، جدول

۱. مسلمی، نیکی، مدیر واحد مجری. ۲. گروه پژوهشی برنامه‌ریزی و بهره‌برداری سیستم‌های قدرت. ۳. محور پژوهشی: مدیریت بار

انتشارات پژوهشگاه نیرو

طرح اکتساب دانش فنی طراحی، پیاده‌سازی و استقرار سیستم‌های نوین خدمات مشترکین

کلیه حقوق معنوی اثر متعلق به پدیدآورندگان، کلیه حقوق مادی متعلق به نام کارفرما و نشر مطالب در حدود متعارف، با ذکر منبع بلامانع است.

نشانی: تهران، شهرک غرب، انتهای بلوار شهید دادمان، پژوهشگاه نیرو، کدپستی: ۱۴۶۸۶۱۳۱۱۳

تلفن: ۹-۸۸۰۷۹۴۰۱

نمابر: ۸۸۰۷۸۲۹۶

وبگاه: www.nri.ac.ir

پست الکترونیک انتشارات: publications@nri.ac.ir

فرم تعهدنامه اصالت اثر

متعهد می‌شوم که مطالب مندرج در این گزارش حاصل کار پژوهشی اینجانب حمیدرضا آراسته، مدیر پروژه بررسی کاربرد تحلیل داده در مدیریت انرژی الکتریکی صنایع و همکاران به شرح زیر است و دستاوردهای پژوهشی دیگران که در این پژوهش از آن‌ها استفاده شده است مطابق مقررات، مورد ارجاع قرار گرفته و در فهرست منابع ذکر گردیده است.

■ گزارش حاصل انجام همه مراحل پروژه است.

□ گزارش حاصل انجام ..... مرحله از ..... مرحله پروژه است.

DOI:

فهرست مطالب

فهرست جدول‌ها..... ب
فهرست شکل‌ها..... ت
پیشگفتار..... ج
فصل اول: آشنایی با مفاهیم حوزه تحلیل داده..... ۱
۱-۱ مقدمه..... ۲
۱-۲ اهداف گزارش..... ۵
۱-۳ ساختار گزارش..... ۶
۱-۴ مروری بر مفاهیم..... ۶
۱-۵ روند داده‌کاوی..... ۹
۱-۶ روش‌های داده‌کاوی..... ۲۴
۱-۷ حفظ حریم خصوصی و ملاحظات اخلاقی در داده‌کاوی..... ۲۹
۱-۸ چالش‌ها و جهت‌گیری‌های آینده..... ۳۳
۱-۹ جمع‌بندی و نتیجه‌گیری..... ۳۶
فصل دوم) بررسی کاربرد تحلیل داده در مدیریت انرژی صنایع..... ۳۸
۲-۱ مقدمه..... ۳۹
۲-۲ تحلیل داده‌ها..... ۳۹
۲-۳ دسته‌بندی صنایع..... ۶۷
۲-۴ جمع‌بندی و نتیجه‌گیری..... ۷۳
جمع‌بندی و نتیجه‌گیری..... ۷۵
مراجع..... ۷۷

فهرست جدول‌ها

جدول ۱-۱: نمونه‌ای از مجموعه‌داده ۱۳
جدول ۱-۲: مجموعه‌داده‌های آموزش ۱۹
جدول ۱-۳: مجموعه‌داده‌های تست ۱۹
جدول ۲-۱: همبستگی ساعتی شرکت ۱..... ۵۳
جدول ۲-۲: میانگین ضرایب همبستگی به تفکیک هر ساعت برای شرکت ۱..... ۵۴
جدول ۲-۳: همبستگی ساعتی شرکت ۲..... ۵۶

فهرست شکل‌ها

شکل ۱-۱: نمای کلی چارچوب CRISP-DM ۱۰
شکل ۱-۲: روند داده‌کاوی ۱۱
شکل ۱-۳: گام‌های مدلسازی ۱۹
شکل ۱-۴: پراکندگی نمودار داده‌های آموزش و تست ۲۰
شکل ۲-۱: نمودار سالانه بار شرکت ۱ ..... ۴۰
شکل ۲-۲: نمودار سالانه بار شرکت ۲ ..... ۴۰
شکل ۲-۳: نمودار مصرف شرکت ۱ در روز یکشنبه، ۳۰ آبان ۱۴۰۰ ..... ۴۱
شکل ۲-۴: نمودار مصرف شرکت ۲ در روز یکشنبه، ۳۰ آبان ۱۴۰۰ ..... ۴۱
شکل ۲-۵: نمودار مصرف شرکت ۱ در روز یکشنبه، ۱۷ بهمن ۱۴۰۰ ..... ۴۲
شکل ۲-۶: نمودار مصرف شرکت ۲ در روز یکشنبه، ۱۷ بهمن ۱۴۰۰ ..... ۴۲
شکل ۲-۷: نمودار مصرف شرکت ۱ در روز یکشنبه، ۹ مرداد ۱۴۰۱ ..... ۴۳
شکل ۲-۸: نمودار مصرف شرکت ۲ در روز یکشنبه، ۹ مرداد ۱۴۰۱ ..... ۴۳
شکل ۲-۱۵: مقدار بار ساعتی روز دوشنبه برای شرکت ۱ ..... ۴۴
شکل ۲-۱۶: مقدار بار ساعتی روز جمعه برای شرکت ۱ ..... ۴۴
شکل ۲-۱۷: مقدار بار ساعتی روز دوشنبه برای شرکت ۲ ..... ۴۵
شکل ۲-۱۸: مقدار بار ساعتی روز جمعه برای شرکت ۲ ..... ۴۵
شکل ۲-۱۹: مقدار بار ساعتی روزهای مختلف هفته برای شرکت ۱ ..... ۴۶
شکل ۲-۲۰: مقدار بار ساعتی روزهای مختلف هفته برای شرکت ۲ ..... ۴۶
شکل ۲-۲۱: مقدار بار ساعت ۱۴ طی سال برای شرکت ۱ ..... ۴۷
شکل ۲-۲۲: مقدار بار ساعت ۱۴ طی سال برای شرکت ۲ ..... ۴۸
شکل ۲-۹: نمودار شاخص همبستگی ساعتی شرکت ۱ با استفاده از داده‌های کل سال (ساعات ۱، ۸، ۱۴، و ۲۰) ۴۹
شکل ۲-۱۰: نمودار شاخص همبستگی ساعتی شرکت ۲ با استفاده از داده‌های کل سال (ساعات ۱، ۸، ۱۴، و ۲۰) ۵۰
شکل ۲-۱۱: نمودار شاخص همبستگی ساعتی شرکت ۱ با استفاده از داده‌های تابستان ۱۴۰۱ (ساعات ۱، ۸، ۱۴، و ۲۰) ۵۰

شکل ۲-۱۲: نمودار شاخص همبستگی ساعتی شرکت ۲ با استفاده از داده‌های تابستان ۱۴۰۱ (ساعات ۱، ۸، ۱۴، و ۲۰)..... ۵۱
شکل ۲-۱۳: نمودار شاخص همبستگی روزانه شرکت ۱ با استفاده از داده‌های ۳۶۵ روز سال (برای روز ۱ نسبت به سایر روزها)..... ۵۱
شکل ۲-۱۴: نمودار شاخص همبستگی روزانه شرکت ۲ با استفاده از داده‌های ۳۶۵ روز سال (برای روز ۱ نسبت به سایر روزها)..... ۵۲
شکل ۲-۲۳: نمودار هیستوگرام ساعت ۱۴ شرکت ۱..... ۶۱
شکل ۲-۲۴: نمودار هیستوگرام ساعت ۱۴ شرکت ۲..... ۶۱
شکل ۲-۲۵: نمودار رگرسیون بار ساعت ۱ نسبت به ساعت ۱۵ برای شرکت ۱..... ۶۲
شکل ۲-۲۶: نمودار رگرسیون بار ساعت ۱ نسبت به ساعت ۱۵ برای شرکت ۲..... ۶۳
شکل ۲-۲۷: نمودار مقادیر بار ساعت ۱ نسبت به مقادیر بار ساعت ۱۵ برای شرکت ۱..... ۶۴
شکل ۲-۲۸: نمودار مقادیر بار ساعت ۱ نسبت به مقادیر بار ساعت ۱۵ برای شرکت ۲..... ۶۴
شکل ۲-۲۹: خوشه‌بندی بار برای شرکت ۱ بر اساس مقادیر بار در ساعات ۱ و ۱۵..... ۶۵
شکل ۲-۳۰: خوشه‌بندی بار برای شرکت ۲ بر اساس مقادیر بار در ساعات ۱ و ۱۵..... ۶۶

پیشگفتار

در این پروژه مفاهیم و روش‌های اصلی داده‌کاوی و کاربرد آن در مدیریت انرژی صنایع بررسی خواهند شد. داده‌کاوی، یک حوزه بین‌رشته‌ای است که بر استخراج الگوها و نکات قابل ارزش از مجموعه‌های داده بزرگ تمرکز دارد. در این گزارش، ابتدا در فصل اول مروری بر مفاهیم داده‌کاوی، اهمیت و جایگاه آن در دنیای امروزی و آینده، روند اجرای آن، روش‌های مختلف داده‌کاوی (از جمله دسته‌بندی، خوشه‌بندی، قواعد انجمن[۱] و تشخیص ناهنجاری[۲])، چالش‌های پیش‌رو و راه‌حل‌های آن (از قبیل مسائل حریم خصوصی و ملاحظات اخلاقی در داده‌کاوی)، و نیز نگاهی به چالش‌ها و جهت‌گیری‌های آینده داده‌کاوی ارائه شده است. در ادامه، در فصل دوم به بررسی کاربرد داده‌کاوی در مدیریت انرژی صنایع و تنظیم تعرفه‌ها پرداخته شده است. در این فصل، علاوه بر بیان مفاهیم و بررسی مفهومی موضوع، از داده‌های در دسترس دو نمونه از صنایع نیز استفاده شده و برخی تحلیل‌ها و نتایج حاصل نیز نشان داده شده‌اند. همچنین سعی شده است تا بر اساس تحلیل‌های انجام‌شده، نمونه‌هایی از مقایسه و دسته‌بندی مبتنی بر نتایج نیز ارائه شود. در پایان نیز جمع‌بندی و نتیجه‌گیری گزارش ارائه شده است.

---

[۱] Association

[۲] Anomaly

۱

آشنایی با مفاهیم حوزه تحلیل داده

۱-۱ مقدمه

تحلیل داده[۱] و داده‌کاوی[۲] دو حوزه مرتبط هستند که در آن‌ها از حجم عظیمی از داده‌ها جهت استخراج اطلاعات ارزشمند و الگوها به منظور تسهیل در تصمیم‌گیری، کشف دانش و پیش‌بینی استفاده می‌شود. این حوزه‌ها در دنیای کلان‌داده[۳] که حجم زیادی از داده‌های ساختاریافته و ساختارنیافته[۴] تولید و ذخیره می‌شوند، اهمیت به‌مراتب بیشتری پیدا کرده است.

تحلیل داده شامل مجموعه‌ای از تکنیک‌ها و فرآیندهایی است که برای بررسی[۵]، پاکسازی[۶]، تبدیل[۷] و مدل‌سازی داده‌ها[۸] به منظور کشف الگوها و روابط معنادار استفاده می‌شود. این فرآیند شامل مراحلی مانند تحلیل اکتشافی داده[۹]، تحلیل آماری[۱۰]، تجزیه‌وتحلیل تصویری داده[۱۱] و آزمون فرضیه[۱۲] است. هدف از تحلیل داده درک جامعی از داده‌ها و شناسایی روندها و همبستگی‌های مهم است.

از سوی دیگر، داده‌کاوی بر روی استفاده از الگوریتم‌ها و تکنیک‌های محاسباتی برای کشف الگوها[۱۳]، روابط[۱۴] و اطلاعات پنهان[۱۵] در داده‌ها تمرکز دارد. این حوزه شامل استفاده از روش‌ها و الگوریتم‌هایی مانند طبقه‌بندی[۱۶]، خوشه‌بندی[۱۷]، استخراج قواعد انجمن (AR)[۱۸] و شناسایی ناهنجاری[۱۹] است. هدف از داده‌کاوی، شناسایی الگوها و استخراج دانش مفید از داده‌ها است.

حوزه تحلیل داده و داده‌کاوی در طول سالیان گذشته با پیشرفت‌های قدرت محاسبی، قابلیت‌های ذخیره‌سازی، رشد روزافزون حجم داده‌ها و افزایش توانایی‌های الگوریتمی[۲۰] و روش‌های مورداستفاده بهبود یافته است. این پیشرفت‌ها موجب ایجاد فرصت‌های جدید برای کشف دانش و ارزش‌های پنهان در داده‌ها شده است. امروزه،

---

[۱] Data Analyses

[۲] Data Mining

[۳] big data

[۴] structured and unstructured data

[۵] examine

[۶] clean

[۷] transform

[۸] model data

[۹] exploratory data analysisg

[۱۰] statistical analysis

[۱۱] data visualization

[۱۲] hypothesis testing

[۱۳] pattern

[۱۴] relationships

[۱۵] insights

[۱۶] classification

[۱۷] clustering

[۱۸] Association Rule (AR) mining

[۱۹] anomaly detection

[۲۰] algorithmic techniques

سازمان‌ها در صنایع مختلف از تحلیل داده و داده‌کاوی برای کسب مزیت رقابتی[۱]، بهینه‌سازی عملیات[۲]، بهبود تجربه مشتری[۳] و تصمیم‌گیری مبتنی بر داده[۴] استفاده می‌کنند.

برای بهره‌برداری از اطلاعات تحلیل داده و داده‌کاوی، لازم است به مفاهیم و روش‌های پایه این حوزه‌ها پرداخت. این مساله شامل فهم مراحل فرآیند تحلیل داده و داده‌کاوی می‌شود که شامل تعیین مسئله[۵]، جمع‌آوری و ادغام داده‌ها[۶]، پیش‌پردازش داده‌ها[۷]، تحلیل اکتشافی داده[۸]، تحلیل آماری[۹] و تجزیه‌وتحلیل تصویری داده[۱۰] است.
در تحلیل داده، مباحثی مانند توزیع‌ها، استنتاج آماری، آزمون فرضیه، رگرسیون[۱۱] و تحلیل تفسیری داده مورد بررسی قرار می‌گیرد. از طرفی، داده‌کاوی شامل استفاده از الگوریتم‌ها و تکنیک‌هایی نظیر خوشه‌بندی، طبقه‌بندی، تجزیه‌وتحلیل انجمن، استخراج قواعد ارتباطی و مدل‌سازی پیشگویانه است. همان‌طور که بیان شد، هدف از داده‌کاوی کشف الگوها و اطلاعات مفید در داده‌ها است و مفاهیمی همچون جمع‌آوری داده‌ها، پیش‌پردازش داده‌ها، تحلیل اکتشافی داده و ارزیابی مدل‌ها مورد توجه قرار می‌گیرد. همچنین، ارزیابی کیفیت مدل‌ها و تفسیر نتایج نیز از اهمیت ویژه‌ای برخوردار است. برای انجام این فرآیندها، از ابزارها و زبان‌های برنامه‌نویسی متنوعی نظیر R، Python، SAS و MATLAB استفاده می‌شود.
در عصر اطلاعات، داده‌ها به عنوان یک دارایی ارزشمند در نظر گرفته می‌شوند. اما ارزش واقعی داده‌ها در قابلیت استخراج اطلاعات و دانش مفید از آنها است. همانطور که اشاره شد، در تحلیل داده جهت تجزیه‌وتحلیل داده‌ها و استخراج الگوها و اطلاعات مفید، از روش‌ها و تکنیک‌های متنوعی استفاده می‌شود. برخی از این روش‌ها عبارتند از [۱-۶][۱۲]:

۱. خوشه‌بندی: این روش به دسته‌بندی داده‌ها بر اساس شباهت‌ها و اختلافات آنها می‌پردازد. با استفاده از الگوریتم‌های خوشه‌بندی، داده‌ها به گروه‌های مشابه تقسیم می‌شوند و الگوها و روابط میان داده‌ها کشف می‌شوند.

---

[۱] competitive edge

[۲] optimize operations

[۳] improve customer experiences

[۴] make data-driven decisions

[۵] problem formulation

[۶] data collection and integration

[۷] data preprocessing

[۸] exploratory data analysis

[۹] statistical analysis

[۱۰] data visualization

[۱۱] Regression

[۱۲] توجه شود کسب دانش کافی درباره روش‌ها، الگوریتم‌ها، ارزیابی مدل‌ها و استقرار آنها نیز بسیار حائز اهمیت است. این روش‌های مطرح در حوزه تحلیل داده و داده‌کاوی تنها بخشی از موضوعات مورد بررسی است. در واقع، داده‌کاوی یک حوزه بسیار گسترده است که شامل بسیاری از روش‌ها، الگوریتم‌ها و مفاهیم است. به همین دلیل، برای بررسی کامل این حوزه، نیاز به مطالعه و پژوهش بیشتر است.

۲. طبقه‌بندی: در این روش، داده‌ها بر اساس ویژگی‌های خاص به دسته‌ها یا کلاس‌های مختلف تقسیم می‌شوند. با استفاده از الگوریتم‌های طبقه‌بندی، مدل‌هایی برای پیش‌بینی کلاس یا برچسب داده‌های جدید ایجاد می‌شود.
۳. استخراج قواعد ارتباطی: این روش برای کشف الگوها و ارتباط‌هایی بین آیتم‌های مختلف در داده‌ها استفاده می‌شود. با استخراج قواعد ارتباطی، می‌توان روابط متقابل بین آیتم‌ها را شناسایی کرده و از آنها برای انجام تحلیل‌های بیشتر و تصمیم‌گیری‌های بهتر استفاده کرد.
۴. مدل‌سازی پیشگویانه[۱]: در این روش، مدل‌های آماری و ریاضی برای پیش‌بینی یا تخمین مقادیر آینده بر اساس داده‌های حاضر ساخته می‌شوند. این مدل‌ها به عنوان ابزارهای قدرتمند در تصمیم‌گیری و تحلیل و پیش‌بینی در حوزه‌های مختلف از جمله بازاریابی، مالی، بهداشت و پزشکی، علوم اجتماعی و بسیاری از صنایع دیگر استفاده می‌شوند.
۵. تحلیل جریان داده[۲]: با پیشرفت فناوری، امروزه داده‌ها به صورت جریانی و در زمان واقعی تولید می‌شوند. تحلیل جریان داده به بررسی و تحلیل داده‌های جریانی با توجه به محدودیت‌های زمانی و حجمی آنها می‌پردازد.
۶. بهینه‌سازی[۳]: در حوزه داده‌کاوی، بهینه‌سازی برای بهبود کارایی و استفاده بهینه از منابع استفاده می‌شود. مسائل بهینه‌سازی با اهداف مختلفی مانند انتخاب ویژگی‌ها، تنظیم پارامترها و بهبود عملکرد الگوریتم‌ها در زمینه داده‌کاوی بررسی می‌شوند.
۷. تحلیل شبکه‌ها[۴]: با استفاده از روش‌های داده‌کاوی، می‌توان شبکه‌های پیچیده از جمله شبکه‌های اجتماعی، شبکه‌های مخابراتی و شبکه‌های عصبی را تحلیل کرده و ارتباطات، الگوها و خواص آنها را شناسایی کرد.
۸. تحلیل متن[۵]: تحلیل متن به استخراج اطلاعات، الگوها و دانش از متن‌ها با استفاده از روش‌های داده‌کاوی می‌پردازد.
۹. تحلیل کلان‌داده: با پیشرفت فناوری و حجم رو به افزایش داده‌ها، تحلیل کلان‌داده به یکی از چالش‌های اساسی در حوزه داده‌کاوی تبدیل شده است. این روش شامل استخراج، تحلیل و استفاده از داده‌های با حجم بسیار بزرگ، با توجه به تنوع آنها است.
۱۰. تحلیل عامل[۶]: این روش برای کاهش ابعاد و تحلیل ویژگی‌های یک مجموعه از متغیرها استفاده می‌شود. تحلیل عامل بر اساس ارتباطات وابسته میان متغیرها، عوامل عمده را شناسایی می‌کند و از آنها برای پیش‌بینی داده‌ها استفاده می‌کند.

---

[۱] Predictive Modeling
[۲] Stream Data Analysis
[۳] Optimization
[۴] Network Analysis
[۵] Text Mining
[۶] Factor Analysis

۱۱. تحلیل خوشه‌های مخفی[۱]: این روش بر اساس مشاهدات مشابه در داده‌ها، خوشه‌های مخفی را شناسایی می‌کند. این خوشه‌های مخفی بیانگر وجود گروه‌ها و الگوهای نهفته در داده‌ها هستند که با استفاده از روش‌های خوشه‌بندی معمولی قابل استخراج نیستند.
۱۲. تحلیل سری زمانی[۲]: در این روش، داده‌ها به صورت توالی‌های زمانی در نظر گرفته می‌شوند و الگوها و روندهای زمانی در آنها شناسایی می‌شوند. تحلیل روند زمانی می‌تواند برای پیش‌بینی، مدل‌سازی و تحلیل داده‌هایی مانند فروش، ترافیک، سرمایه‌گذاری و سایر حوزه‌ها مورد استفاده قرار گیرد.
۱۳. تحلیل الگوهای متوالی[۳]: این روش برای شناسایی الگوهای متوالی و تکراری در داده‌ها استفاده می‌شود. با استفاده از تحلیل الگوهای تکراری، می‌توان الگوهای متوالی و تکراری در داده‌ها را شناسایی کرده و اطلاعات مهم و مفیدی را استخراج کرد. این روش در حوزه‌هایی نظیر تجزیه‌وتحلیل سیستم‌های وب، تجزیه‌وتحلیل رفتار مشتریان و تحلیل داده‌های زمانی مورد استفاده قرار می‌گیرد.
۱۴. تحلیل پیشگویانه: تحلیل پیشگویانه به استفاده از روش‌های داده‌کاوی برای پیش‌بینی و تخمین وقوع رویدادها و رفتارها می‌پردازد. با تحلیل داده‌های موجود، الگوها و روندهایی شناسایی می‌شود که امکان پیش‌بینی رویدادها را فراهم می‌کند. این روش در حوزه‌هایی از قبیل پیش‌بینی رفتار مشتریان، پیش‌بینی بازدهی سرمایه‌گذاری و پیش‌بینی رفتار بازار کاربرد دارد.

این توضیحات مربوط به بخشی از پیشینه و مبانی حوزه داده‌کاوی است. همانطور که اشاره شد، داده‌کاوی یک حوزه بسیار گسترده است و در این گزارش به برخی از مفاهیم و روش‌های پایه در داده‌کاوی اشاره می‌شود. همچنین، هر روش و مفهوم مطرح شده نیز خود یک حوزه تحقیقاتی است. برای دستیابی به اطلاعات در حوزه تحلیل داده و داده‌کاوی، منابع متعددی وجود دارند که می‌توان از آنها استفاده کرد. برخی از منابعی که به طور عمومی در این زمینه مورد استفاده قرار می‌گیرند عبارتند از:

- کتاب "Data Mining: Concepts and Techniques" نوشته Han، Kamber و Pei؛
- کتاب "Data Mining: Practical Machine Learning Tools and Techniques" نوشته Witten، Frank، Hall و Pal؛
- کتاب "Introduction to Data Mining"" نوشته Tan، Steinbach و Kumar.

این کتاب‌ها شرح جامعی از مفاهیم، روش‌ها و الگوریتم‌های مختلف تحلیل داده و داده‌کاوی را ارائه می‌دهند. منابع دیگری که می‌توان به آنها مراجعه کرد شامل مقالات علمی دانشگاهی، انتشارات تحقیقاتی و منابع آنلاین سازمان‌ها، مؤسسات تحقیقاتی و صنایع معتبر هستند. برای مطالعه بیشتر و عمیق‌تر در هر موضوع، پیشنهاد می‌شود به مراجع [۷-۲۱] مراجعه شود.

۱-۲ اهداف گزارش

هدف از این گزارش بررسی مفاهیم تحلیل داده، آشنایی با روند و روش‌های انجام داده‌کاوی، مروری بر مسائل مرتبط با حوزه حریم خصوصی و ملاحظات اخلاقی در حوزه تحلیل داده، نگاهی به چالش‌ها و جهت‌گیری‌های آینده داده‌کاوی و همچنین بررسی کاربرد تحلیل داده در تنظیم تعرفه‌ها و نیز مدیریت انرژی صنایع است.

---

[۱] Latent Cluster Analysis

[۲] Time Series Analysis

[۳] Sequential Pattern Analysis

۱-۳ ساختار گزارش

در ادامه، ابتدا مروری بر مفاهیم تحلیل داده ارائه می‌شود. سپس، روند داده‌کاوی بررسی خواهد شد و در ادامه برخی روش‌های انجام داده‌کاوی نیز معرفی می‌شود. در بخش بعدی، مباحثی در ارتباط با حفظ حریم خصوصی و ملاحظات اخلاقی و چالش‌ها و راه‌حل‌های آن ارائه می‌شود. در پایان فصل اول، چالش‌ها و جهت‌گیری‌های آینده در مبحث داده‌کاوی مورد بررسی اجمالی قرار خواهد گرفت. سپس در فصل دوم، به جایگاه تحلیل داده در حوزه مدیریت انرژی صنایع و همچنین تنظیم تعرفه پرداخته خواهد شد. در این بخش، با استفاده از داده‌های در دسترس از داده‌های مربوط به دو نمونه از صنایع کشور، برخی تحلیل‌ها ارائه شده و مقایسه بین نتایج آنها نیز ارائه می‌گردد. در این بخش سعی بر آن خواهد بود تا بتوان چند نمونه دسته‌بندی نیز با توجه به نتایج تحلیل ارائه شود. در پایان جمع‌بندی و نتیجه‌گیری نهایی گزارش ارائه خواهد شد.

۱-۴ مروری بر مفاهیم

۱-۴-۱- تعریف داده‌کاوی و اهمیت آن

- ✓ تعریف

داده‌کاوی فرایند کشف الگوها، روابط و بینش از مجموعه‌های بزرگ داده است. این فرایند شامل استخراج اطلاعات و دانش مفید از داده‌هایی پیچیده و اغلب بدون ساختار (مانند پایگاه‌های داده[۱]، انبارهای داده[۲]، اسناد متنی[۳]، پیام‌های رسانه‌های اجتماعی[۴] و غیره) می‌شود. روش‌های داده‌کاوی از الگوریتم‌های آماری و محاسباتی مختلفی برای کشف الگوها، روندها و همبستگی‌های پنهان[۵] استفاده می‌کنند که می‌توانند برای اتخاذ تصمیمات درست، پیش‌بینی و شناسایی بینش‌های ارزشمند[۶] مورد استفاده قرار گیرند.

داده‌کاوی به طور معمول شامل چند مرحله است، از جمله آماده‌سازی داده[۷]، بررسی داده[۸]، ساخت مدل[۹] و اعتبارسنجی نتایج[۱۰]. این فرایند از روش‌های مربوط به حوزه‌هایی مانند یادگیری ماشین، آمار، تشخیص الگو و سیستم‌های پایگاه‌داده استفاده می‌کند. هدف داده‌کاوی استخراج اطلاعات معنادار و کاربردی از حجم بزرگی از

---

[۱] databases

[۲] data warehouses

[۳] text documents

[۴] social media feeds

[۵] hidden patterns, trends, and correlations

[۶] identifying valuable insights

[۷] data preparation

[۸] data exploration

[۹] model building

[۱۰] result validation

داده است که بتواند برای بهبود فرآیندهای تصمیم‌گیری، بهینه‌سازی استراتژی‌های تجاری، بهبود تجربه مشتری و کسب مزیت رقابتی در حوزه‌های مختلف مورد استفاده قرار گیرد.

- ✓ اهمیت

داده‌کاوی نقش بسیار مهمی در جهان پیشرفته مبتنی بر داده داشته و بهره‌های مهمی را به ارمغان می‌آورد. در ادامه تعدادی از دلایل مهم و برجسته اهمیت داده‌کاوی بررسی می‌شوند:

۱. کشف دانش[۱]: داده‌کاوی به سازمان‌ها کمک می‌کند تا الگوها، روندها و روابط پنهان در داده‌های خود را کشف کنند. با تجزیه‌وتحلیل مجموعه‌های بزرگ داده، شرکت‌ها می‌توانند بینش‌های ارزشمندی به دست آورند و دانشی استخراج کنند که با روش‌های تجزیه‌وتحلیل سنتی قابل درک نیست.

۲. تصمیم‌گیری و استراتژی[۲]: داده‌کاوی با فراهم کردن اطلاعات و بینش‌های مربوطه، تصمیم‌گیری آگاهانه را تسهیل می‌کند. با درک رفتار مشتریان و روندهای بازار و الگوهای تجاری، شرکت‌ها می‌توانند تصمیماتی بر پایه داده بگیرند تا عملکرد خود را بهینه کنند، استراتژی‌های خود را بهبود بخشند و فرصت‌های جدید را شناسایی کنند.

۳. تحلیل پیشگویانه: تکنیک‌های داده‌کاوی، مانند مدل‌سازی پیشگویانه، به شرکت‌ها اجازه می‌دهند بر اساس داده‌های پیشین، نتایج آینده را پیش‌بینی کنند. با بهره‌گیری از این پیش‌بینی‌ها، سازمان‌ها می‌توانند روندهای بازار، ترجیحات (اولویت‌ها) مشتریان[۳]، نوسانات تقاضا و خطرات محتمل[۴] را پیش‌بینی کنند. این امر آنان را توانمند می‌سازد تا با بررسی پیش‌بینی‌ها، اقدامات پیشگیرانه مناسبی اتخاذ نموده و استراتژی‌های خود را متناسب با آن تنظیم کنند.

۳. مدیریت ارتباط با مشتری[۵]: داده‌کاوی سبب بهبود مدیریت ارتباط با مشتری می‌شود؛ زیرا سازمان‌ها را قادر می‌سازد تا ترجیحات، رفتار و نیازهای مشتریان را درک کنند. با تجزیه‌وتحلیل داده‌های مشتریان، شرکت‌ها می‌توانند خود را متناسب با پویش‌های بازاریابی ارتقا دهند[۶]، رضایت مشتری را بهبود بخشند و استراتژی‌های هدفمندی برای جذب و حفظ مشتریان توسعه دهند.

۵. کشف و پیشگیری از تقلب[۷]: روش‌های داده‌کاوی به طور گسترده در شناسایی و جلوگیری از فعالیت‌های متقلبانه استفاده می‌شوند. سازمان‌ها با تجزیه‌وتحلیل الگوها و شناسایی ناهنجاری‌ها و موارد پرت در داده‌ها[۸] می‌توانند معاملات، فعالیت‌ها یا رفتارهای متقلبانه بالقوه را شناسایی کنند. این مهم به کمینه کردن ضررهای مالی، حفاظت از منافع مشتری و حفظ یکپارچگی عملیات تجاری کمک می‌کند. این مسئله به ویژه در صنایع مالی، بیمه و امنیت سایبری[۹] بسیار اهمیت دارد.

---

۱ Knowledge Discovery

۲ Decision Making and Strategy

۳ customer preferences

۴ potential risks

۵ Customer Relationship Management

۶ can personalize marketing campaigns

۷ Fraud Detection and Prevention

۸ outliers in data

۹ cybersecurity

۶. بهبود عملیات و کارآیی[۱]: داده‌کاوی به بهینه‌سازی فرآیندهای کسب و کار و بهبود کارایی عملیات کمک می‌کند. با تجزیه‌وتحلیل داده‌های عملیاتی، سازمان‌ها می‌توانند گلوگاه‌ها[۲]، ناکارآمدی‌ها و حوزه‌های لازم جهت بهبود را شناسایی کنند. این امر منجر به تسهیل گردش کار، کاهش هزینه‌ها، افزایش بهره‌وری و بهبود کلی عملکرد می‌شود.

۷. اطلاعات بازار و مزیت رقابتی[۳]: داده‌کاوی اطلاعات بازاری مفیدی را برای سازمان‌ها تأمین می‌کند و آنها را قادر می‌سازد تا در رقابت با سایر رقبا پیشتاز باشند. با تجزیه‌وتحلیل روندهای بازار، رفتار مصرف‌کنندگان، استراتژی‌های رقبا و الگوهای جدید، شرکت‌ها می‌توانند نقاط ضعف بازار[۴] را شناسایی کرده، محصولات نوآورانه توسعه دهند، استراتژی‌های بازاریابی را بهبود بخشند و به سطح رقابتی مناسبی دست پیدا کنند.

۸. توسعه محصول و نوآوری[۵]: داده‌کاوی نقشی حیاتی در توسعه محصول و نوآوری دارد. با تجزیه‌وتحلیل بازخورد مشتری، روند بازار و تجزیه‌وتحلیل رقبا، سازمان‌ها می‌توانند حوزه‌های ارتقای محصول، فرصت‌های جدید و ویژگی‌های نوآورانه را شناسایی کنند. داده‌کاوی به کسب و کارها کمک می‌کند تا محصولات ارائه شده خود را با نیازها و ترجیحات مشتری هماهنگ کنند و منجر به افزایش رضایت مشتری و مزیت رقابتی خود در بازار شوند.

۹. بهینه‌سازی منابع[۶]: داده‌کاوی سبب می‌شود که سازمان‌ها بتوانند منابع خود را به طور موثر بهینه کنند. با تجزیه‌وتحلیل داده‌های مربوط به تخصیص منابع، استفاده از الگوها و کارا شدن عملیات، کسب و کارها می‌توانند حوزه‌های تلفات و ناکارآمدی و راه‌های کاهش هزینه را شناسایی کنند. این بهینه‌سازی منجر به بهبود مدیریت منابع، افزایش بهره‌وری و صرفه‌جویی در هزینه می‌شود.

۱۰. مراقبت‌های بهداشتی و تحقیقات پزشکی[۷]: در صنعت مراقبت‌های بهداشتی، داده‌کاوی نقش مهمی در مراقبت از بیمار، تشخیص بیماری و تحقیقات پزشکی ایفا می‌کند. با تجزیه‌وتحلیل داده‌های مربوط به مراقبت‌های بهداشتی، می‌توان الگوها و روندها را برای بهبود تشخیص[۸] و اثربخشی درمان[۹] شناسایی کرد. داده‌کاوی همچنین با شناسایی همبستگی‌ها، عوامل ریسک[۱۰] و درمان‌ها یا مداخلات بالقوه[۱۱] به تحقیقات پزشکی کمک می‌کند.

---

۱ Improved Operations and Efficiency

۲ bottlenecks

۳ Market Intelligence and Competitive Advantage

۴ market gaps

۵ Product Development and Innovation

۶ Resource Optimization

۷ Healthcare and Medical Research

۸ diagnosis

۹ treatment effectiveness

۱۰ risk factors

۱۱ potential treatments or interventions

۱۱. شخصی‌سازی و بازاریابی هدفمند[۱]: داده‌کاوی سازمان‌ها را قادر می‌سازد تا اقدامات بازاریابی خود را شخصی‌سازی کرده و کمپین‌های هدفمند ارائه دهند[۲]. با تجزیه‌وتحلیل داده‌های مشتری (از قبیل شناخت جمعیت[۳]، رفتار و ترجیحات)، کسب‌وکارها می‌توانند پیام‌های بازاریابی، پیشنهادات و توصیه‌های خود را متناسب با مشتریان ارائه کنند. این مساله منجر به بازاریابی موثرتر و تعامل بیشتر و بهتر با مشتری می‌شود.
۱۲. مدیریت زنجیره تامین[۴]: داده‌کاوی نقش مهمی در بهینه‌سازی مدیریت زنجیره تامین ایفا می‌کند. با تجزیه‌وتحلیل داده‌های زنجیره تامین، سازمان‌ها می‌توانند زمینه‌های ناکارآمدی، تنگناها[۵] و تاخیرها را شناسایی کنند. این مساله به آنها اجازه می‌دهد تا فرآیندها را ساده کنند، میزان دارایی را بهینه کنند[۶]، سطح تدارکات را بهبود بخشند[۷] و عملکرد کلی زنجیره تامین را بهبود دهند.
۱۳. مدیریت و ارزیابی ریسک[۸]: داده‌کاوی به شناسایی و مدیریت ریسک‌های مختلف مرتبط با عملیات تجاری کمک می‌کند. با تجزیه‌وتحلیل داده‌های پیشین و روندهای بازار و عوامل خارجی، سازمان‌ها می‌توانند ریسک‌هایی مانند ریسک‌های مالی، ریسک‌های عملیاتی و ریسک‌های بازار را ارزیابی و کاهش دهند. این مهم به کسب و کارها اجازه می‌دهد تا تصمیمات آگاهانه بگیرند، استراتژی‌های کاهش ریسک را اجرا کنند و از منافع خود محافظت کنند.

به طور کلی، داده‌کاوی سازمان‌ها را توانمند می‌کند تا از قدرت داده‌های خود استفاده کرده و بینش‌های قابل‌توجهی را استخراج کنند و بتوانند تصمیم‌گیری‌هایی بر پایه این داده‌ها انجام دهند. این امر به آنها امکان می‌دهد تا فرصت‌های پنهان را کشف کنند، عملکردهای خود را بهینه کنند، تجربه مشتریان را بهبود بخشند، ریسک‌ها را مدیریت کنند و در دنیای امروزی مبتنی بر داده قادر به رقابت با رقبای دیگر باشند.

۱-۵ روند داده‌کاوی[۹]

کشف روابط و الگوهای موجود در داده‌ها با استفاده از مجموعه‌ای از اقدامات حاصل می‌شود که به عنوان روند داده‌کاوی شناخته می‌شود. یک روند استاندارد برای داده‌کاوی شامل این مراحل است [۲۲]: اول؛ شناخت مساله[۱۰]، دوم؛ تهیه نمونه داده[۱۱]، سوم؛ توسعه مدل، چهارم؛ اعمال مدل بر روی مجموعه‌داده‌ها جهت بررسی اینکه عملکرد مدل به چه صورت است، و پنجم؛ اجرا[۱۲].

---

۱ Personalization and Targeted Marketing

۲ deliver targeted campaigns

۳ demographics

۴ Supply Chain Management

۵ bottlenecks

۶ optimize inventory levels

۷ improve logistics

۸ Risk Assessment and Management

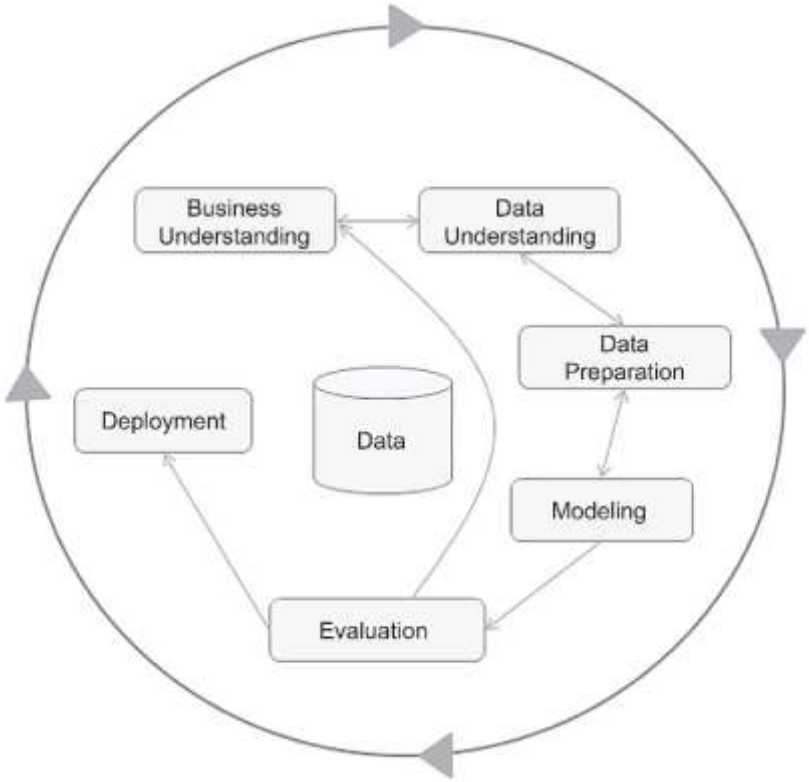
۹ Data Mining Process

۱۰ understanding the problem

۱۱ preparing the data samples

۱۲ production deployment

با گسترش کاربرد داده‌کاوی، چارچوب‌های مختلفی برای روند داده‌کاوی توسط محققین دانشگاهی و تجاری ارائه شده است. یکی از چارچوب‌های محبوب برای انجام روند داده‌کاوی CRISP-DM[۱] است که توسط ائتلافی[۲] از چندین شرکت فعال در حوزه داده‌کاوی ارائه شده است [۲۳]. چارچوب CRISP-DM پرکاربردترین چارچوب در زمینه روند انجام داده‌کاوی است. در شکل (۱-۱)، نمایی کلی از این چارچوب نشان داده شده است [۲۲].



شکل ۱-۱: نمای کلی چارچوب CRISP-DM [۲۲]

برخی دیگر از چارچوب‌های داده‌کاوی عبارت هستند از: SEMMA[۳] که توسط موسسه SAS توسعه داده شده است [۲۴]، DMAIC[۴] که در Six Sigma practice استفاده شده است [۲۵] و چارچوب انتخاب، پردازش، تبدیل، داده‌کاوی، تفسیر و ارزیابی[۵] که در تحلیل KDD[۶] استفاده شده است [۲۶]. هر کدام از چارچوب‌های داده‌کاوی اجرای مجموعه‌ای از وظایف را جهت دستیابی به خروجی بهینه پیشنهاد می‌دهند. روند استخراج اطلاعات یک مسیر مبتنی بر تکرار است. گام‌های موجود در داده‌کاوی غیرخطی بوده و شامل حلقه‌های فراوانی جهت بازگشت به عقب، حرکت به پیش و حرکت بین گام‌های مختلف و گاهی اوقات بازگشت به گام اول جهت بازتعریف مساله داده‌کاوی است.

---

[۱] Cross Industry Standard Process for Data Mining

[۲] consortium

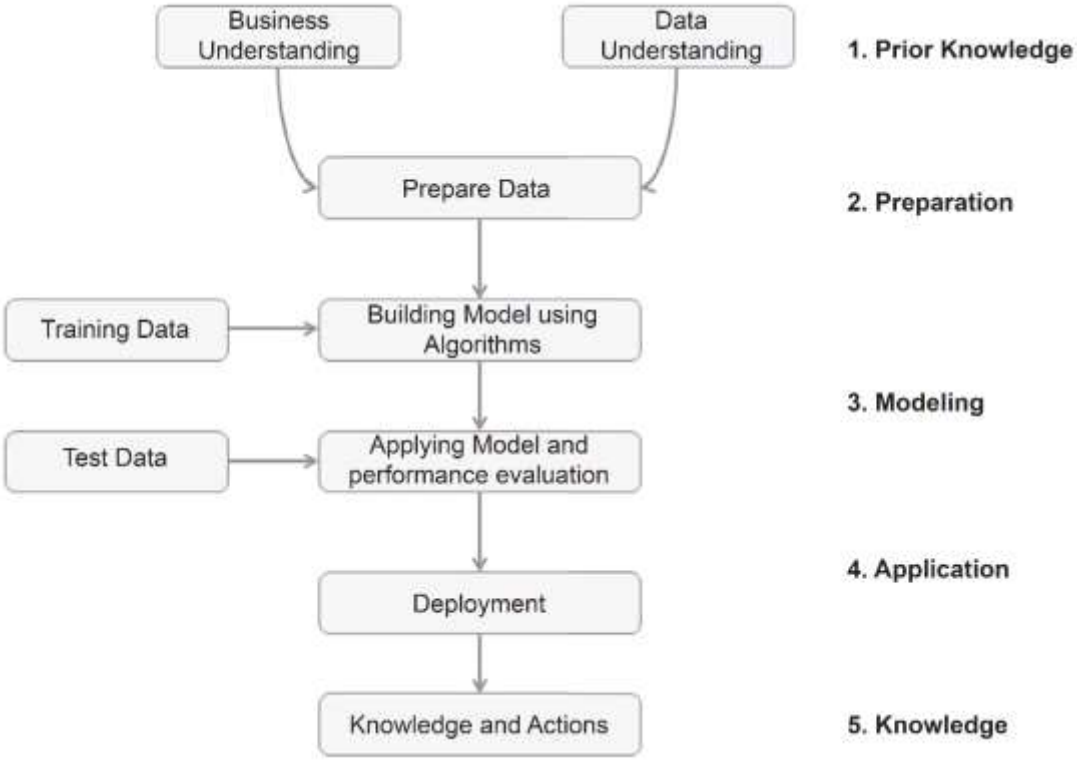
[۳] Sample, Explore, Modify, Model, and Assess (SEMMA)

[۴] Define, Measure, Analyze, Improve and Control (DMAIC)

[۵] Selection, Preprocessing, Transformation, Data Mining, Interpretation, and Evaluation framework

[۶] knowledge discovery in databases (KDD)

در شکل (۱-۲) گام‌های عمومی جهت انجام داده‌کاوی نشان داده شده است که شامل تجارت[۱]، الگوریتم و ابزار داده‌کاوی است [۲۲]. هدف اساسی هر فرآیندی در داده‌کاوی پرداختن به سوال اصلی تحلیل است. مساله موجود می‌تواند مواردی از قبیل تقسیم‌بندی مشتریان، پیش‌بینی آب و هوا یا بررسی الگوی داده‌ها باشد.



شکل ۱-۲: روند داده‌کاوی [۲۲]

داده‌کاوی، به ویژه در زمینه داده‌های بزرگ، در چند سال اخیر اهمیت زیادی پیدا کرده است. شاید موردبحث‌ترین بخش داده‌کاوی، مرحله سوم یعنی مدل‌سازی باشد که شامل ساخت مدلی است که می‌تواند از مجموعه‌ای از داده‌های نمونه مشتق شود و برای پیش‌بینی‌ها (مدل‌سازی پیشگویانه) یا توصیف الگوی اساسی در داده‌ها (مدل‌سازی توصیفی یا توضیحی[۲]) استفاده شود [۲۲].

متخصصان باتجربه داده‌کاوی می‌توانند این واقعیت را تأیید کنند که وقت گیرترین بخش فرآیند داده‌کاوی، بخش ساخت مدل نیست؛ بلکه آماده‌سازی داده‌ها و حصول درک تجاری است. ابزارهای داده‌کاوی زیادی، هم به صورت دسترسی آزاد[۳] و هم به صورت تجاری[۴]، در بازار موجود هستند که می‌توانند ساخت مدل را به صورت خودکار انجام دهند. پرکاربردترین ابزارها عبارتند از RapidMiner، R، Weka، SAS، SPSS، Oracle Data Miner، Salford و Statistica [۲۷].

[۱] business

[۲] descriptive or explanatory modeling

[۳] open source

[۴] commercial

پرسیدن سوالات تجاری مناسب، به دست آوردن درک عمیق تجاری، منبع‌یابی[۱] و آماده‌سازی داده‌ها برای انجام داده‌کاوی، کاهش ملاحظات پیاده‌سازی، و از همه مفیدتر کسب دانش از فرآیند داده‌کاوی، برای موفقیت در انجام داده‌کاوی بسیار مهم است. جهت بررسی روند انجام داده‌کاوی می‌توان از مرحله ابتدایی یعنی درک موضوع و چارچوب‌بندی کردن داده‌کاوی شروع کرد که ادامه بیان شده است [۲۲].

۱-۵-۱- دانش پیشین[۲]

دانش پیشین به اطلاعاتی اطلاق می‌شود که از قبل در مورد یک موضوع وجود دارد. هدف داده‌کاوی به صورت مجزا از دانش پیشین ظاهر نمی‌شود و همواره مبتنی بر موضوع و اطلاعات زمینه‌ای موجود توسعه می‌یابد. این مرحله کمک می‌کند تا مشخص شود مشکل موجود چیست، چه اهمیتی در کسب و کار دارد و به چه داده‌هایی برای حل مشکل نیاز است [۲۲]. در ادامه، حوزه‌های مختلف مربوط به دانش پیشین بررسی می‌شوند.

۱-۵-۱-۱- هدف

فرآیند داده‌کاوی با یک نیاز جهت تحلیل، یک سوال یا یک هدف تجاری شروع می‌شود. این احتمالاً مهم‌ترین مرحله در فرآیند داده‌کاوی است [۲۸]. با اینکه فرآیند داده‌کاوی یک فرآیند متوالی است و بازگشت به مراحل قبلی و تجدیدنظر در مفروضات، رویکردها و تاکتیک‌ها معمول است، بدون یک بیان کاملاً مشخص از مشکل، دستیابی به مجموعه‌داده مناسب و انتخاب الگوریتم داده‌کاوی مناسب غیرممکن است [۲۲].

به عنوان مثال، یک کسب و کار وام را در نظر بگیرید که نرخ بهره وام به طیفی از متغیرها مانند نرخ فعلی وجوه فدرال[۳]، امتیاز اعتبار وام‌گیرنده و مبلغ سپرده اولیه (پیش‌پرداخت) بستگی دارد. هدف تجاری این مورد می‌تواند این باشد که اگر نرخ بهره وام‌گیرندگان گذشته با طیفی از امتیازات اعتباری معلوم باشد، آیا می‌توان نرخ بهره را برای وام‌گیرنده جدید پیش‌بینی کرد [۲۲].

۱-۵-۱-۲- موضوع[۴]

فرآیند داده‌کاوی الگوهای پنهان در مجموعه‌داده را با افشای روابط بین ویژگی‌ها آشکار می‌کند. اما مسئله این است که الگوهای زیادی را آشکار می‌کند. سیگنال‌های نادرست یک مشکل عمده در این فرآیند هستند. این بر عهده متخصص داده‌کاوی است که الگوها را فیلتر کند و الگوهایی را بپذیرد که برای پاسخ به سؤال معتبر و مرتبط هستند. از این رو، دانستن موضوع، زمینه و فرآیند کسب و کار ضروری است [۲۲].

به عنوان نمونه، مثال پیشین را در نظر بگیرید. کسب و کار وام[۵] یکی از قدیمی‌ترین، رایج‌ترین و پیچیده‌ترین مشاغل است. اگر هدف داده‌کاوی پیش‌بینی نرخ بهره باشد، مهم است که مشخص باشد که کسب‌وکار وام‌دهی چگونه کار می‌کند، چرا پیش‌بینی اهمیت دارد، پس از اطلاع از نرخ بهره پیش‌بینی‌شده چه کاری باید انجام شود، چه داده‌هایی را می‌توان از وام‌گیرندگان جمع‌آوری کرد، چه نقاط داده‌ای را نمی‌توان به دلیل مقررات موجود جمع‌آوری کرد، چه عوامل خارجی دیگری می‌تواند بر نرخ بهره تأثیر بگذارد، چگونه بایستی اعتبار نتیجه حاصل را تأیید کرد و غیره. درک مدل‌های فعلی و شیوه‌های کسب‌وکار، شالوده و دانش شناخته شده را ایجاد می‌کند.

---

[۱] sourcing

[۲] Prior Knowledge

[۳] federal funds rate

[۴] Subject Area

[۵] ding business

تجزیه‌وتحلیل و استخراج داده‌ها دانش جدیدی را فراهم می‌کند که می‌تواند مبتنی بر دانش موجود ساخته شود [۲۹].

۱-۵-۱-۳- داده‌ها

مشابه دانش پیشین در رابطه با موضوع، دانش پیشین در رابطه با داده‌ها نیز وجود دارد. داده‌ها معمولاً به عنوان بخشی از فرآیندهای تجاری در یک شرکت جمع‌آوری می‌شوند. درک نحوه جمع‌آوری، ذخیره، تبدیل، گزارش و استفاده از داده‌ها برای فرآیند داده‌کاوی ضروری است. در این مرحله طیف وسیعی از عوامل را باید در نظر گرفت: کیفیت داده‌ها، کمیت داده‌ها، در دسترس بودن داده‌ها، اینکه اگر داده‌ها در دسترس نباشند چه اتفاقی می‌افتد، آیا کمبود داده کارشناس را مجبور به تغییر سؤال تجاری خواهد کرد، و غیره. هدف از این کار، ارائه یک مجموعه‌داده است که استخراج و تحلیل آن به سؤالات تجاری موجود پاسخ خواهد داد [۲۲]. برخی از اصطلاحات مورد استفاده در فرآیند داده‌کاوی در رابطه با توصیف داده‌ها به صورت زیر است [۲۲]:

- مجموعه‌داده (مجموعه نمونه)[۱] مجموعه‌ای از داده‌ها با ساختار تعریف‌شده است. جدول (۱-۱) یک مجموعه‌داده را نشان می‌دهد که ساختاری کاملاً مشخص با ۱۰ سطر و ۳ ستون به همراه عناوین ستون‌ها را دارد.
- نقاط داده (رکورد یا شی یا نمونه)[۲] یک نمونه واحد در مجموعه‌داده است. هر ردیف در جدول (۱-۱) یک نقطه داده است. هر نمونه شامل همان ساختار مجموعه‌داده است.
- یک ویژگی (ویژگی یا ورودی یا بعد یا متغیر یا پیش‌بینی‌کننده)[۳] یک ویژگی واحد از مجموعه‌داده است. هر ستون در جدول (۱-۱) یک ویژگی است. ویژگی‌ها می‌توانند عدد، دسته‌بندی، تاریخ (زمان)، متن یا انواع داده‌های پولی باشند. در جدول (۱-۱)، امتیاز اعتباری و نرخ بهره ویژگی‌های عددی هستند.
- برچسب (برچسب کلاس یا خروجی یا پیش‌بینی یا هدف یا پاسخ)[۴] ویژگی خاصی است که باید بر اساس همه ویژگی‌های ورودی پیش‌بینی شود. در جدول (۱-۱)، نرخ بهره متغیر خروجی است.
- شناسه‌ها[۵] ویژگی‌های خاصی هستند که برای مکان‌یابی[۶] یا ارائه مفاد برای هر داده مجزا[۷] استفاده می‌شوند. به عنوان مثال، ویژگی‌های رایج مانند نام، شماره حساب و شناسه کارمند ویژگی‌های شناسه هستند. شناسه‌ها اغلب به عنوان کلیدهای جستجو برای ترکیب چند مجموعه‌داده استفاده می‌شوند. آنها هیچ اطلاعاتی ندارند که برای ساخت مدل‌های داده‌کاوی مناسب باشد و بنابراین باید برای مرحله مدل‌سازی واقعی حذف شوند. به عنوان مثال، در جدول (۱-۱) ID شناسه است.

جدول ۱-۱: نمونه‌ای از مجموعه‌داده [۲۲]

---

[۱] data set (example set)

[۲] data point (record or data object or example)

[۳] An attribute (feature or input or dimension or variable or predictor)

[۴] label (class label or output or prediction or target or response)

[۵] Identifiers

[۶] locating

[۷] providing context to individual records

Borrower ID	Credit Score	Interest Rate
01	500	7.31%
02	600	6.70%
03	700	5.95%
04	700	6.40%
05	800	5.40%
06	800	5.70%
07	750	5.90%
08	550	7.00%
09	650	6.50%
10	825	5.70%

۱-۵-۱-۴- علیت و همبستگی[۱]

فرض شود در مثال پیشین هدف پیش‌بینی معکوس شده است؛ یعنی آیا بر اساس داده‌های جدول (۱-۱) می‌توان امتیاز اعتباری وام‌گیرنده را بر اساس نرخ بهره پیش‌بینی کرد؟ پاسخ مثبت است؛ اما این کار منطقی نیست. امتیاز اعتباری بر نرخ بهره وام تأثیر می‌گذارد. پیش‌بینی امتیاز اعتباری بر اساس نرخ بهره این رابطه علی را معکوس می‌کند. این سوال همچنین یکی از جنبه‌های کلیدی ساخت مدل را آشکار می‌کند. همبستگی بین ویژگی‌های ورودی و خروجی، علیت را تضمین نمی‌کند. از این رو، بسیار مهم است که پرسش داده‌کاوی را به درستی و با استفاده از دانش موجود چارچوب‌بندی کرد [۲۲].

۱-۵-۲- آماده‌سازی داده‌ها[۲]

آماده‌سازی مجموعه‌داده‌ها برای تطابق با یک وظیفه مشخص زمان‌برترین بخش فرآیند داده‌کاوی است. به ندرت داده‌ها به شکل موردنیاز در دسترس هستند. بسیاری از الگوریتم‌های داده‌کاوی نیاز به یک قالب جدولی برای ساختار داده‌ها دارند؛ به طوری که ردیف‌ها مربوط به رکوردها و ستون‌ها مربوط به ویژگی‌ها باشند. اگر داده‌ها در هر فرمت دیگری باشند، باید داده‌ها را به شکل موردنظر تبدیل کرد. همچنین، سوال دیگر این است که برای تغییر شکل داده‌ها به ساختار موردنیاز، اگر داده‌های نادرست یا مقادیر از دست رفته وجود داشته باشند، چه اقدامی باید صورت پذیرد. در ادامه برخی از اقدامات لازم در مرحله آماده‌سازی داده‌ها به همراه چالش‌های رایج و راه‌حل آنها بیان شده است [۲۲].

۱-۵-۲-۱- کاوش داده‌ها[۳]

آماده‌سازی داده‌ها با کاوش عمیق داده‌ها و به دست آوردن درک بیشتر از مجموعه‌داده‌ها آغاز می‌شود. کاوش داده‌ها، که به عنوان تجزیه‌وتحلیل داده‌های اکتشافی[۴] نیز شناخته می‌شود، مجموعه‌ای از ابزارهای ساده را برای دستیابی به درک اولیه از داده‌ها فراهم می‌کند. کاوش اساسی می‌تواند ساختار داده‌ها، نحوه توزیع مقادیر داده‌ها[۵]

[۱] Causation vs. Correlation

[۲] Data Preparation

[۳] Data Exploration

[۴] Exploratory Data Analysis (EDA)

[۵] distribution of the values

و وجود مقادیر نامتعارف[۱] را نشان دهد و روابط موجود درون مجموعه‌داده را نمایان سازد[۲]. آمار توصیفی[۳] مانند میانگین[۴]، میانه[۵]، انحراف از معیار[۶] و بازه هر ویژگی[۷] بیانی ساده از ویژگی‌های کلیدی داده‌ها را ارائه می‌دهد. از سوی دیگر، با استفاده از یک نمودار می‌توان درکی سریع از تمام داده‌ها را فراهم کرد [۲۲].

۱-۵-۲-۲- کیفیت داده[۸]

هر جایی که داده‌ها جمع‌آوری، پردازش و ذخیره می‌شوند، یکی از نگرانی‌هایی که وجود دارد مربوط به کیفیت داده‌ها است. به عنوان مثال، در مجموعه‌داده‌ای که پیشتر استفاده شد (جدول ۱-۱)، چگونه می‌توان از دقت داده‌های امتیاز اعتباری و نرخ بهره مطمئن بود. خطاهای ممکن در داده‌ها در میزان گویایی[۹] مدل اثر می‌گذارد. سازمان‌ها از روش‌های پاکسازی و تبدیل داده‌ها[۱۰] برای بهبود و مدیریت کیفیت داده‌ها استفاده می‌کنند و آن‌ها را در محل ذخیره اطلاعات کل شرکت به نام انبار داده[۱۱] ذخیره می‌کنند. داده‌هایی که منبع آن از انباره داده‌هایی باشد که به خوبی نگهداری شده‌اند دارای کیفیت بالاتری هستند، زیرا کنترل های مناسبی برای اطمینان از دقت داده‌های جدید و موجود وجود دارد. روش‌های پاکسازی داده‌ها شامل مواردی از قبیل حذف رکوردهای تکراری، جدا کردن رکوردهای پرت (که خارج از بازه‌ها هستند)[۱۲]، استانداردسازی مقادیر مشخصه[۱۳] و جایگزینی مقادیر از دست رفته[۱۴] است. صرف‌نظر از این، پیش از ایجاد مدل‌ها و برای اطمینان از درجه خاصی از کیفیت داده‌ها، ضروری است علاوه بر استفاده از دانش قبلی ، داده‌ها با استفاده از روش‌های اکتشاف داده[۱۵] بررسی شوند [۲۲].

۱-۵-۲-۳- مقادیر از دست رفته[۱۶]

یکی از رایج ترین مشکلات کیفیت داده‌ها این است که برخی رکوردها دارای مقادیر مشخصه از دست رفته هستند. چندین روش برای مقابله با این مشکل وجود دارد، اما هر روش دارای مزایا و معایبی است. اولین قدم در مدیریت مقادیر از دست رفته، درک دلیل نبود مقادیر است. ردیابی ریشه یک داده از منبع داده[۱۷] می‌تواند منجر به شناسایی مسائل سیستمایک[۱۸] در جمع‌آوری داده‌ها یا خطا در تبدیل داده‌ها شود، یا اینکه ممکن است پدیده‌ای وجود داشته

---

۱ presence of extreme values

۲ highlights

۳ Descriptive statistics

۴ mean

۵ median

۶ standard deviation

۷ range for each attribute

۸ Data Quality

۹ representativeness

۱۰ data cleansing and transformation

۱۱ companywide repositories called Data Warehouses

۱۲ quarantining outlier records that exceed the bounds

۱۳ standardization of attribute values

۱۴ substitution of missing values

۱۵ data exploration techniques

۱۶ Missing Values

۱۷ data lineage of the data source

۱۸ systemic issues

باشد که کاربر هنوز متوجه آن نشده باشد. غالبا دانستن منبع یک مقدار از دست رفته می‌تواند به انتخاب روش مناسب جهت کاهش این موارد کمک کند. می‌توان مقادیر از دست رفته را با طیفی از داده‌های مصنوعی[۱] به گونه‌ای جایگزین کرد که بتوان این مشکل را با تأثیراتی قابل چشم‌پوشی بر مراحل بعدی داده‌کاوی مدیریت کرد. به عنوان مثال، بسته به ویژگی‌های یک مساله، برخی مقادیر از دست رفته را می‌توان با مقداری که از مجموعه‌داده‌ها قابل محاسه است (مانند میانگین، حداقل یا حداکثر) جایگزین کرد. این روش در صورتی مفید است که مقادیر از دست رفته کاملاً تصادفی رخ دهند و فراوانی وقوع آنها بسیار نادر باشد. در غیر این صورت، توزیع مشخصه‌ای که داده‌های از دست رفته دارد، مخدوش می‌شود[۲]. از طرف دیگر، برای ساخت مدلی گویا، می‌توان تمام رکوردهای داده با مقدار از دست رفته یا رکوردهایی با کیفیت داده ضعیف را نادیده گرفت. این روش اندازه مجموعه‌داده‌ها را کاهش می‌دهد. برخی از الگوریتم‌های داده‌کاوی در مدیریت رکوردهای با مقادیر از دست رفته خوب عمل می‌کنند؛ در حالی که در برخی دیگر لازم است که در مرحله آماده‌سازی داده‌ها و ساخت و اعمال مدل، این مساله حل شده باشد. به عنوان مثال، الگوریتم k-نزدیکترین همسایه[۳] برای انجام دسته‌بندی اغلب در برابر داده‌های از دست رفته مقاوم است؛ در حالی که مدل‌های شبکه عصبی برای انجام دسته‌بندی تحت این شرایط به خوبی عمل نمی‌کنند و بنابراین مرحله آماده‌سازی داده‌ها برای توسعه مدل‌های شبکه عصبی ضروری است [۲۲].

۱-۵-۲-۴- انواع داده‌ها و تبدیل آنها[۴]

ویژگی‌های موجود در یک مجموعه‌داده می‌توانند انواع مختلفی داشته باشند، از قبیل اعداد پیوسته (مانند نرخ بهره)، اعداد صحیح (مانند امتیاز اعتباری) و انواع قابل طبقه‌بندی (مانند ضعیف، خوب و عالی). الگوریتم‌های مختلف داده‌کاوی محدودیت‌های متفاوتی را در مورد نوع داده‌ای که به عنوان ورودی می‌پذیرند اعمال می‌کنند. اگر مدل موردنظر یک مدل رگرسیون خطی ساده باشد، ویژگی‌های ورودی باید به صورت عددی باشند، بنابراین، اگر داده‌های موجود به صورت دسته‌بندی‌شده هستند، باید به ویژگی عددی پیوسته تبدیل شوند. روش‌های مختلفی برای تبدیل انواع دسته‌بندی به ویژگی‌های عددی وجود دارد. به عنوان مثال، می‌توان یک عدد خاص را به هر دسته نسبت داد (مانند ضعیف = ۴۰۰، خوب = ۶۰۰ و عالی = ۷۰۰). به طور مشابه، مقادیر عددی را می‌توان با استفاده از روشی به نام binning به انواع داده‌های طبقه‌بندی تبدیل کرد (به عنوان مثال کم = [۴۰۰-۵۰۰]) [۲۲].

۱-۵-۲-۵- تغییر[۵]

در برخی از الگوریتم‌های داده‌کاوی مانند k-NN، انتظار می‌رود که ورودی‌ها عددی و عادی باشند، زیرا این الگوریتم ها اعداد مختلف را مقایسه و فاصله بین نقاط داده را محاسبه می‌کنند. لازم است اطمینان داشت که یک عدد خاص به دلایلی از قبیل داشتن مقدار خیلی بزرگ در نتیجه نهایی غالب نباشد. به عنوان نمونه، در مثالی که پیش‌تر بیان شد، درآمد و امتیاز اعتباری از لحاظ عددی ممکن است تفاوت زیادی داشته باشند (مثلا یکی در حدود هزار یا میلیون و دیگری در حدود صد باشد). یک راه‌حل این است که اعداد مختلف با توجه به محدوده خود نرمالیزه شوند و به مقیاسی بین صفر و یک تبدیل شوند. به این ترتیب، می‌توان یک مقایسه یکسان بین دو ویژگی مختلف با واحدهای مختلف انجام داد. با این حال، وجود نقاط پرت ممکن است نتایج نرمالیزه کردن را منحرف کند. در برخی

---

[۱] artificial data

[۲] will be distorted

[۳] k-nearest neighbor (k-NN)

[۴] Data Types and Conversion

[۵] Transformation

از مسائل داده‌کاوی، لازم است تعداد ویژگی‌ها کاهش یابد. با استفاده از برخی روش‌های آماری مانند تجزیه‌وتحلیل مؤلفه های اصلی[۱] می‌توان تعداد ویژگی‌ها را به چند ویژگی کلیدی یا اصلی کاهش داد [۲۲]. وجود چندین ویژگی که همبستگی بالایی دارند ممکن است برای برخی روش‌ها نامطلوب باشد. به عنوان مثال، درآمد سالانه و مالیات پرداختی همبستگی بالایی دارند و از این رو ممکن است لازم باشد یکی از ویژگی‌ها را حذف کرد [۲۲].

۱-۵-۲-۶- موارد پرت[۲]

طبق تعریف، نقاط پرت ناهنجاری‌هایی در مجموعه‌داده‌ها هستند. موارد پرت ممکن است به طور مشروع (مانند درآمد میلیاردی) یا به اشتباه (مانند ثبت قد انسان برابر با ۱,۷۳ سانتی‌متر) رخ دهد. لازم است موارد پرت به درستی شناسایی و اصلاح شوند. هدف از ساختن یک مدل، تعمیم یک الگو یا ارتباط در داده‌ها است و وجود نقاط پرت کارایی مدل را تحت تاثیر قرار خواهد داد. لازم به ذکر است ممکن است در برخی از کاربردهای داده‌کاوی، شناسایی نقاط پرت خود به عنوان هدف اصلی مطرح باشد (مانند تشخیص تقلب[۳] و تشخیص نفوذ[۴]) [۲۲].

۱-۵-۲-۷- انتخاب مشخصه (ویژگی)[۵]

بسیاری از مسائل داده‌کاوی شامل یک مجموعه‌داده با صدها یا هزاران ویژگی[۶] است. در کاربردهای متن‌کاوی[۷]، هر کلمه متمایز یک ویژگی در مجموعه‌داده در نظر گرفته می‌شود. بنابراین، مجموعه‌داده مورد استفاده در این مورد شامل هزاران ویژگی است. همه ویژگی‌ها در پیش‌بینی مقدار هدف مورد نظر به یک اندازه مهم یا مفید نیستند. برخی از ویژگی‌ها ممکن است به شدت با یکدیگر مرتبط باشند (مانند درآمد سالانه و مالیات پرداختی). وجود تعداد زیادی از ویژگی‌ها در مجموعه‌داده‌ها به طور قابل‌توجهی پیچیدگی یک مدل را افزایش می‌دهد و ممکن است عملکرد مدل را به دلیل مشکل ابعادی مساله تضعیف کند. هرچند وجود اطلاعات دقیق‌تر در داده‌کاوی مطلوب است (زیرا کشف اجزای مختلف یک الگو در داده‌ها یکی از جذابیت‌های استفاده از داده‌کاوی است)؛ اما با افزایش تعداد داده‌ها، داده‌ها در فضایی با ابعاد بزرگ پراکنده می‌شوند و این شرایط قابلیت‌اطمینان مدل‌ها[۸] را (به ویژه در مواردی مانند خوشه‌بندی و دسته‌بندی) کاهش می‌دهد [۳۰]. کاهش تعداد مشخصه‌ها بدون کاهش قابل‌توجه در عملکرد مدل، انتخاب مشخصه (ویژگی) نامیده می‌شود. کاهش تعداد ویژگی‌ها در مجموعه‌داده منجر به مدلی ساده‌تر شده و به بیان بهتر مدل کمک می‌کند [۲۲].

۱-۵-۲-۸- نمونه‌برداری از داده‌ها[۹]

نمونه‌برداری فرآیند انتخاب یک زیرمجموعه از داده‌ها به عنوان نماینده مجموعه‌داده‌های اصلی جهت استفاده در تجزیه‌وتحلیل داده‌ها یا مدل‌سازی است. داده‌های نمونه نماینده مجموعه‌داده‌های اصلی با ویژگی‌های مشابه (مانند

---

[۱] principal component analysis (PCA)

[۲] Outliers

[۳] fraud detection

[۴] intrusion detection

[۵] Feature Selection

[۶] attributes

[۷] text mining

[۸] reliability of the models

[۹] Data Sampling

میانگین مشابه با مجموعه‌داده اصلی) هستند. با استفاده از نمونه‌برداری می‌توان تعداد داده‌های موردنیاز برای تجزیه‌وتحلیل و مدل‌سازی را کاهش داد. در بیشتر موارد، کار با نمونه داده‌های انتخاب شده جهت حصول نتیجه مورد نظر کافی است. نمونه‌برداری روند مدل‌سازی را سرعت می‌بخشد. لازم به ذکر است، از نظر تئوری خطای ایجاد شده در نتیجه استفاده از داده‌های نمونه بر دقت مدل تأثیر می‌گذارد، اما مزایای آن بسیار بیشتر از معایب است [۲۲].

۱-۵-۳- مدل‌سازی

مدل نمایشی انتزاعی از داده‌ها[۱] و روابط آن در یک مجموعه‌داده معین است. یک عبارت ساده مانند "نرخ بهره وام مسکن با افزایش امتیاز اعتباری کاهش می‌یابد" یک مدل است. اگرچه این عبارت اطلاعات کمّی ارائه نمی‌دهد، ولی اطلاعاتی را برای درک رابطه بین امتیاز اعتباری و نرخ بهره فراهم می‌کند. امروزه از چندین الگوریتم داده‌کاوی استفاده می‌شود که از آمار، یادگیری ماشین، تشخیص الگو[۲] و دانش علوم کامپیوتر مشتق شده‌اند. خوشبختانه، ابزارهای تحلیلی پیشگویانه و ابزارهای داده‌کاوی زیادی به صورت دسترسی آزاد و تجاری‌شده وجود دارند که این الگوریتم‌ها را پیاده‌سازی می‌کنند. تنها چیزی که یک متخصص داده‌کاوی باید به آن توجه داشته باشد این است که یک دید کلی از الگوریتم داشته باشد، بداند که یک الگوریتم به چه صورت کار می‌کند و تعیین کند که چه پارامترهایی باید بر اساس درک ما از کسب و کار و داده‌ها مشخص شوند. مدل‌های داده‌کاوی را می‌توان به این صورت طبقه‌بندی کرد: دسته‌بندی، رگرسیون، تجزیه‌وتحلیل انجمن، خوشه‌بندی و تشخیص داده‌های پرت یا ناهنجاری‌ها. هر دسته دارای چند ده الگوریتم مختلف است که هر کدام یک رویکرد کمی متفاوت برای حل مشکل موجود ارائه می‌دهند. دسته‌بندی و رگرسیون روش‌های پیشگویانه هستند؛ زیرا آنها یک خروجی را بر اساس یک یا چند متغیر ورودی پیش‌بینی می‌کنند. الگوریتم‌های پیشگویانه برای یادگیری[۳] مدل به مجموعه‌ای از داده‌های پیشین نیاز دارند. شکل (۱-۳) گام‌های لازم جهت مدل‌سازی داده‌کاوی پیشگویانه را نشان می‌دهد. تجزیه‌وتحلیل انجمن و خوشه‌بندی، روش‌های داده‌کاوی توصیفی[۴] هستند که در آن هیچ متغیر هدفی برای پیش‌بینی وجود ندارد. به هر حال، هر دو مدل پیش‌بینی و توصیفی دارای یک مرحله ارزیابی هستند. تشخیص ناهنجاری نیز چنانچه داده‌های لازم در دسترس باشد می‌تواند پیشگویانه باشد [۲۲].

۱-۵-۳-۱- مجموعه‌داده‌های آموزشی[۵] و آزمایشی[۶]

برای توسعه یک مدل پایدار، باید از یک مجموعه‌داده از قبل آماده استفاده کرد که در آن همه ویژگی‌ها، از جمله ویژگی کلاس هدف[۷] شناخته شده هستند. این مجموعه‌داده، مجموعه‌داده آموزشی نامیده می‌شود و برای ایجاد

---

[۱] abstract representation of the data

[۲] pattern recognition

[۳] learn

[۴] descriptive data mining

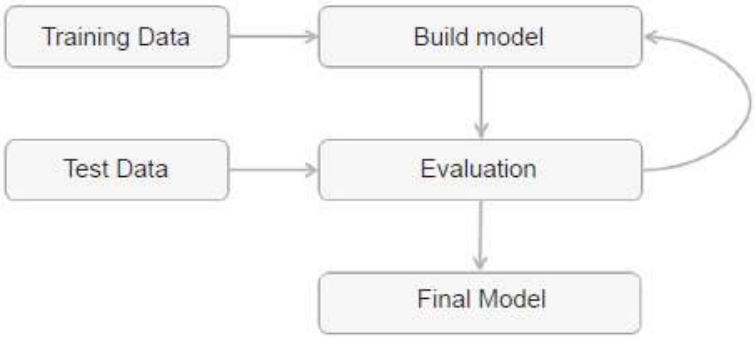
[۵] Training

[۶] Testing

[۷] target class attribute

یک مدل استفاده می‌شود. همچنین باید اعتبار مدل ایجاد شده را با مجموعه‌داده‌های شناخته‌شده دیگری به نام مجموعه‌داده آزمایشی یا مجموعه‌داده اعتبارسنجی[۱] بررسی کرد [۲۲].

برای تسهیل این فرآیند، مجموعه‌داده‌ها را می‌توان به یک مجموعه‌داده آموزشی و یک مجموعه‌داده تست تقسیم کرد. یک قانون سرانگشتی استاندارد[۲] این است که دو سوم داده‌ها به آموزش و یک سوم به مجموعه‌داده‌های تست تخصیص یابند. جدول‌های (۱-۲) و (۱-۳) تقسیم تصادفی داده‌های آموزش و تست را بر اساس مجموعه‌داده نمونه نشان داده شده در جدول (۱-۱) نشان می‌دهد. شکل (۱-۴) نمودار پراکندگی کل مجموعه‌داده نمونه را نشان می‌دهد که در آن مجموعه‌داده‌های آموزشی و تست مشخص شده‌اند [۲۲].



شکل ۱-۳: گام‌های مدلسازی

جدول ۱-۲: مجموعه‌داده‌های آموزش [۲۲]

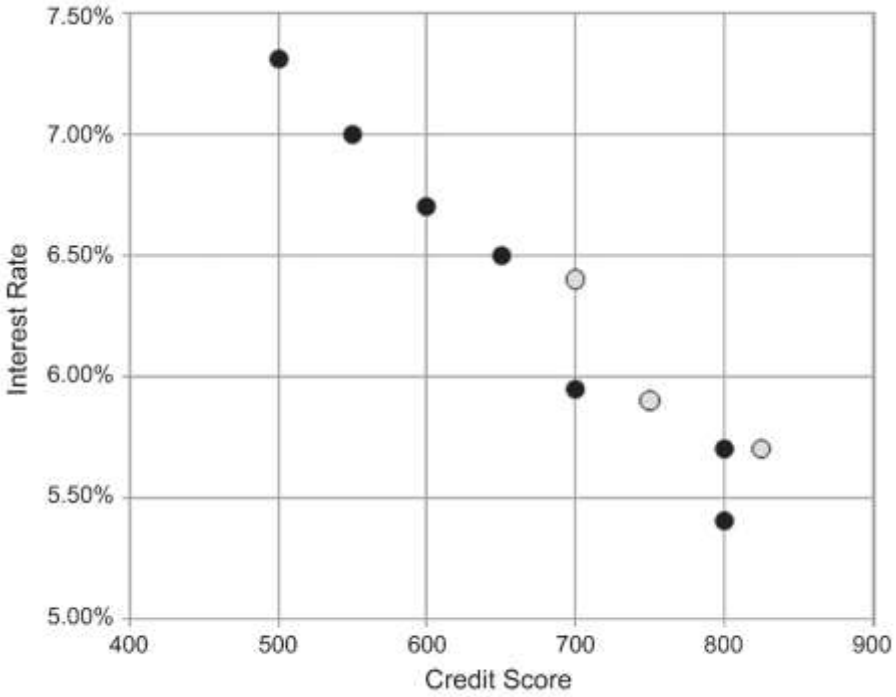
Borrower	Credit Score (X)	Interest Rate (Y)
01	500	7.31%
02	600	6.70%
03	700	5.95%
05	800	5.40%
06	800	5.70%
08	550	7.00%
09	650	6.50%

جدول ۱-۳: مجموعه‌داده‌های تست [۲۲]

Borrower	Credit Score (X)	Interest Rate (Y)
04	700	6.40%
07	750	5.90%
10	825	5.70%

[۱] validation data set

[۲] standard rule of thumb



شکل ۱-۴: پراکندگی نمودار داده‌های آموزش و تست [۲۲]

۱-۵-۳-۲- الگوریتم یا روش مدلسازی[۱]

مساله موجود و نیز داده‌های در دسترس تعیین می‌کنند که از چه نوع داده‌کاوی (از قبیل تجزیه‌وتحلیل انجمن، طبقه‌بندی یا رگرسیون) باید استفاده شود. متخصص داده‌کاوی الگوریتم داده‌کاوی مناسب را از بین انواع موجود انتخاب می‌کند. به عنوان مثال، هر یک از این الگوریتم‌ها را می‌توان انتخاب کرد: درخت‌های تصمیم‌گیری[۲]، القاء قوانین[۳]، شبکه‌های عصبی[۴]، مدل‌های بیزیان[۵]، k-NN و غیره. همچنین، استفاده از چندین دسته و الگوریتم داده‌کاوی برای حل یک سوال تجاری غیر معمول نیست [۲۲].

۱-۵-۳-۳- ارزیابی مدل[۶]

نتایج حاصل از تخمین به کمک یک مدل ممکن است دقیقاً مشابه مقادیر موجود در داده‌های آموزشی نباشد. نیازی نیست که مدل همان مقادیری را که در داده‌های آموزشی وجود دارد، به خاطر سپرده و خروجی دهد. پدیده‌ای که در آن مدل مورد استفاده داده‌های آموزشی را به خاطر می سپارد، بیش برازش[۷] نامیده می‌شود. یک مدل بیش از حد برازش‌شده فقط سوابق آموزشی را به خاطر می‌سپارد و در داده‌های واقعی عملکرد ضعیفی دارد. مدل مورداستفاده بایستی رابطه بین ورودی و خروجی را تعمیم داده یا یاد بگیرد. با استفاده از مدل می‌توان مقدار واقعی خروجی را با مقدار پیش‌بینی‌شده آن مقایسه کرد و به این ترتیب خطای پیش‌بینی را محاسبه کرد. تا زمانی

---

[۱] Algorithm or Modeling Technique

[۲] decision trees

[۳] rule induction

[۴] neural networks

[۵] Bayesian models

[۶] Evaluation of the Model

[۷] overfitting

که خطا قابل قبول باشد، می‌توان از این مدل استفاده کرد. نرخ خطا را می‌توان برای مقایسه یک مدل با مدل‌های دیگر استفاده کرد [۲۲].

۱-۵-۳-۴- مدلسازی گروهی[۱]

مدل‌سازی گروهی فرآیندی است که در آن یا با استفاده از الگوریتم‌های مدل‌سازی مختلف یا با استفاده از مجموعه‌داده‌های آموزشی متفاوت چندین مدل مختلف برای پیش‌بینی یک نتیجه ایجاد می‌شود. سپس، مدل گروهی حاصل از مدل‌های پایه جمع‌آوری شده و پیش‌بینی نهایی برای داده‌های مجهول[۲] انجام می‌شود. هدف از استفاده از مدل‌های گروهی کاهش خطای تعمیم[۳] پیش‌بینی است. تا زمانی که مدل‌های پایه متنوع[۴] و مستقل[۵] باشند، خطای پیش‌بینی مدل با استفاده از رویکرد گروهی کاهش می‌یابد. این رویکرد به دنبال خرد جمعی[۶] در پیش‌بینی است. با وجود اینکه مدل گروهی دارای چندین مدل پایه در داخل خود است، خودش به عنوان یک مدل عمل می‌کند. اکثر روش‌های عملی داده‌کاوی از مدل‌سازی گروهی استفاده می‌کنند [۲۲].

با توجه به موارد بیان‌شده، در پایان مرحله مدل‌سازی، (۱) سؤال تجاری تجزیه‌وتحلیل شده است، (۲) منبع داده‌های مورداستفاده شناسایی شده است، (۳) یک روش داده‌کاوی برای پاسخ به سؤال انتخاب شده است، (۴) یک الگوریتم داده‌کاوی انتخاب شده و داده‌ها به طور متناسب با الگوریتم آماده شده است، (۵) داده‌ها به مجموعه‌داده‌های آموزشی و تست تقسیم شده است، (۶) یک مدل تعمیم یافته[۷] از مجموعه‌داده‌های آموزشی ایجاد شده است و (۷) مدل با استفاده از مجموعه‌داده‌های تست صحت‌سنجی شده است. اکنون می‌توان از این مدل برای پیش‌بینی متغیر موردنظر استفاده کرد [۲۲].

۱-۵-۴- کاربرد[۸]

کاربرد یا پیاده‌سازی[۹] مرحله‌ای است که در آن مدل به اصلاح آماده تولید[۱۰] یا "زنده[۱۱]" می‌شود. مرحله پیاده‌سازی شامل چندین عمل کلیدی است که عبارت هستند از: ارزیابی آمادگی مدل[۱۲]، یکپارچه‌سازی فنی[۱۳]، زمان پاسخ[۱۴]، نگهداری مدل[۱۵]، و ادغام[۱۶] [۲۲].

---

[۱] Ensemble Modeling

[۲] unseen data

[۳] generalization error

[۴] diverse

[۵] independent

[۶] wisdom of crowds

[۷] generalized model

[۸] Application

[۹] Deployment

[۱۰] production ready

[۱۱] live

[۱۲] assessing model readiness

[۱۳] technical integration

[۱۴] response time

[۱۵] model maintenance

[۱۶] assimilation

۱-۵-۴-۱- آمادگی[۱]

بخش آمادگی در پیاده‌سازی، ویژگی‌های کیفی اساسی[۲] موردنیاز برای دستیابی به هدف را تعیین می‌کند. به عنوان نمونه، دو مورد استفاده متمایز را در مثال پیشین در نظر بگیرید: تعیین اینکه آیا یک مصرف‌کننده واجد شرایط وام هست یا خیر و تعیین گروه‌بندی مشتریان برای یک شرکت. فرآیند تایید اعتبار مصرف‌کننده اقدامی است که باید به صورت آنی انجام شود. در واقع، تصمیمات و شرایط اعتبار باید به محض ارائه اطلاعات مشتریان، در همان لحظه ارائه شود. ارائه یک تصمیم سریع و در عین حال ارائه نتایج دقیق به عنوان یک مزیت رقابتی برای این مورد تلقی می‌شود. در این مورد، مدل تصمیم‌گیری نیاز به جمع‌آوری داده‌های ورودی لازم و تصمیم‌گیری در مورد تأیید و شرایط وام در عرض چند ثانیه دارد. بنابراین، در این مورد ویژگی کیفی حیاتی در پیاده‌سازی مدل، انجام بسیار سریع کار است [۲۲].

تقسیم بندی مشتریان بر اساس رابطه آنها با شرکت فرآیندی است که لازم است طی آن سیگنال‌های تعاملات از از بخش‌های مختلف یک شرکت بررسی شود. بر اساس الگوها، مشتریان مشابه در یک گروه قرار می‌گیرند و استراتژی‌های لازم در راستای تعامل بهتر با مشتری طراحی می‌شوند. برای این کاربرد، پردازش دسته‌ای[۳] نیاز است که در آن داده‌ها از بخش‌ها و منابع مختلف جمع‌آوری شده و سوابق کلی مشتریان نیز تقسیم‌بندی می‌شوند. ویژگی کیفی مهم برای این کاربرد، توانایی یافتن الگوهای منحصر به فرد در بین مشتریان است (نه زمان پاسخگویی مدل). انتخاب‌هایی که باید در مراحل آماده‌سازی داده‌ها و مدل‌سازی انجام شود مبتنی بر کاربرد تجاری موردنظر مشخص می‌شود [۲۲].

۱-۵-۴-۲- یکپارچه‌سازی فنی

به احتمال زیاد نوعی ابزار داده‌کاوی (مانند R، RapidMiner، SAS، SPSS و غیره) برای ایجاد مدل‌های داده‌کاوی استفاده می‌شود. استفاده از ابزارهای داده‌کاوی به دلیل عدم نیاز به نوشتن کد جهت پیاده‌سازی یک الگوریتم باعث صرفه‌جویی در زمان می‌شوند. این امر باعث می‌شود که کارشناس بتواند روی داده‌ها، منطق تجاری و استخراج الگو از داده‌ها تمرکز کند. مدل‌های ایجاد شده توسط ابزارهای داده‌کاوی را می‌توان با استفاده از زبان نشانه‌گذاری مدل پیشگویانه[۴] [۳۲] یا با فراخوانی ابزارهای داده‌کاوی[۵] به برنامه‌های کاربردی منتقل کرد. PMML یک قالب قابل‌حمل و سازگار[۶] از مدل را فراهم می‌کند که می‌تواند توسط اکثر ابزارهای تحلیل پیشگویانه و داده‌کاوی استفاده شود. در نتیجه مدل انعطاف‌پذیر نیز خواهد بود تا مثلا بتوان مدل را با استفاده از یک ابزار (به عنوان مثال RapidMiner) توسعه داد و آن را در ابزار دیگری (مانند SAS) پیاده‌سازی کرد. استانداردهای PMML توسط یک اتحادیه پیشرو در صنعت توسعه و نگهداری می‌شود. مدل‌هایی مانند رگرسیون ساده و درخت‌های تصمیم‌گیری را می‌توان مستقیماً در برنامه‌های تجاری و سیستم‌های هوشمند گنجاند. از آنجایی که این مدل‌ها با معادلات ساده و قوانین if-then نشان داده می‌شوند، می‌توان آنها را به راحتی به اکثر زبان‌های برنامه‌نویسی منتقل کرد [۲۲].

۱-۵-۴-۳- زمان پاسخ

---

[۱] Production Readiness

[۲] critical qualities

[۳] batch processing

[۴] Predictive Model Markup Language (PMML)

[۵] invoking data mining tools

[۶] Portable and consistent format

ایجاد برخی از الگوریتم‌های داده‌کاوی مانند k-NN ساده است؛ اما در پیش‌بینی متغیرهای هدف بسیار کند هستند. ساختن الگوریتم‌هایی مانند درخت تصمیم‌گیری زمان‌بر است، اما می‌توان آنها را به قوانین ساده‌ای تقلیل داد که تقریباً برای تمامی کاربردها قابل انجام است. لازم است بین دقت عملکرد و زمان انجام آن سبک و سنگین[۱] شود و در صورت نیاز، اگر مثلا زمان پاسخ قابل قبول نباشد، باید مرحله مدل‌سازی بازبینی شود. کیفیت پیش‌بینی، دسترسی به داده‌های ورودی و زمان پاسخگویی، مهم‌ترین فاکتورهای کیفیت در کاربردهای تجاری هستند [۲۲].

۱-۵-۴-۴- مدلسازی مجدد[۲] و نگهداری مدل

معیار اصلی برای تداوم مدل[۳]، گویا بودن مجموعه‌داده‌ای است که پردازش می‌شود. کاملاً طبیعی است که شرایطی که در آن مدل ساخته شده است، پس از ارسال آن جهت پیاده‌سازی تغییر کند. به عنوان مثال، رابطه بین امتیاز اعتباری و نرخ بهره بر اساس شرایط اقتصاد کلان غالباً تغییر می‌کند. از این رو مدل باید به طور مکرر به‌روز شود. به طور معمول، اعتبار مدل را می‌توان با استفاده از مجموعه‌داده‌های تست جدید و محاسبه میزان خطا آزمایش کرد. اگر میزان خطا از یک آستانه خاص فراتر رود، می‌توان مدل را بازسازی کرد و پیاده‌سازی را دوباره انجام داد. ایجاد یک برنامه نگهداری، بخش کلیدی پیاده‌سازی است [۲۲].

۱-۵-۴-۵- ادغام

غالبا چالش مطرح در کاربردهای داده‌کاوی توصیفی، ادغام دانش کسب‌شده از داده‌کاوی به سازمان یا یک کاربرد خاص است. به عنوان مثال، ممکن است هدف یافتن خوشه‌هایی منطقی در پایگاه‌داده مشتریان باشد تا بتوان برای هر خوشه مشتری راهکار مجزایی ارائه داد. سپس، مرحله بعدی ممکن است دسته‌بندی مشتریان جدید باشد تا آنها را در یکی از خوشه‌های شناخته شده قرار داد. به عنوان مثال، تجزیه‌وتحلیل انجمن می‌تواند راه‌حلی را ارائه دهد که هدف آن تعیین این است که کدام دو محصول بیشتر با هم خریداری می‌شوند. چالش پیش‌روی متخصص داده‌کاوی بیان این یافته‌ها، تعیین ارتباط با سؤال اصلی کسب و کار، کمی‌سازی ریسک‌های مدل و بیان تأثیر تجاری موردانتظار است. جامعه کاربران، تلفیقی از دیدگاه‌های مختلف و مجموعه‌ای از مهارت‌ها است. همه در مورد فرآیند داده‌کاوی و آنچه که می‌تواند یا نمی‌تواند انجام دهد آگاه نیستند. برخی از جنبه‌های این چالش را می‌توان با تمرکز بر نتیجه نهایی و دانستن اطلاعات به جای فرآیند فنی استخراج اطلاعات، برطرف کرد [۲۲].

۱-۵-۵- دانش[۴]

فرآیند داده‌کاوی چارچوبی را برای استخراج اطلاعات از داده‌ها فراهم می‌کند. با ظهور فضای ذخیره‌سازی عظیم، افزایش جمع‌آوری داده‌ها و الگوهای محاسباتی پیشرفته، داده‌های در اختیار همواره در حال افزایش هستند. برای استخراج دانش از این دارایی‌ها بایستی از رویکردهای پیشرفته مانند الگوریتم‌های داده‌کاوی استفاده کرد.
اگرچه بسیاری از این الگوریتم‌ها می‌توانند استخراج دانشی ارزشمند را ارائه دهند، اما این به متخصصان تحلیل‌گر بستگی دارد که الگوریتم‌های مناسب را به طور ماهرانه‌ای اعمال نموده و یک مساله تجاری را به یک مساله داده

---

[۱] trade-offs

[۲] Remodeling

[۳] ongoing relevance of the model

[۴] Knowledge

تبدیل کنند. داده‌کاوی، مانند هر فناوری دیگری، گزینه‌هایی را از نظر الگوریتم‌ها و پارامترهای آنها فراهم می‌کند. هنر اصلی استفاده از این گزینه‌ها برای استخراج اطلاعات درست است که با تمرین قابل توسعه است.
فرآیند داده‌کاوی با دانش قبلی شروع می‌شود، سپس در طول فرآیند آن بینشی افزایشی در مورد کسب و کار حاصل می‌شود و در نهایت به دانش پسین ختم می‌شود. مانند هر تحلیل کمی، فرآیند داده‌کاوی می‌تواند الگوهای جعلی نامربوط را نشان دهد. لازم به ذکر است همه الگوهای کشف شده به دانش منجر نمی‌شوند. باز هم این مساله بر عهده متخصص داده‌کاوی است که الگوهای نامربوط را باطل کند و اطلاعات معنادار را شناسایی کند.
تأثیر اطلاعات به دست آمده از طریق داده‌کاوی را می‌توان با استفاده از یک کاربرد سنجید. این مساله تفاوت بین داشتن اطلاعات از طریق فرآیند داده‌کاوی و بینش حاصل از تجزیه‌وتحلیل داده‌های اولیه است. در واقع، فرآیند داده‌کاوی چارچوبی برای فراخوانی سؤالات درست است [۲۳] که از طریق رویکردهای صحیح، متخصصین را برای حل یک مشکل تجاری راهنمایی می‌کند. داده‌کاوی مجموعه‌ای از قوانین سفت و سخت نیست، بلکه مجموعه‌ای از مراحل تکرار شونده و متمایز است که برای کشف دانش استفاده می‌شود.

۱-۶ روش‌های داده‌کاوی

همانطور که بیان شد، منظور از داده‌کاوی فرآیند یافتن الگوهای مفید بالقوه با استفاده از مجموعه‌داده‌های عظیم است. در طول این فرآیند راهکارهای متنوعی از جمله یادگیری ماشین، آمار و هوش مصنوعی برای استخراج اطلاعات در مورد احتمال رویدادهای آینده استفاده می‌شود. جنبه‌های متنوع داده‌کاوی شامل مواردی از قبیل طبقه‌بندی داده‌ها، یکپارچه‌سازی داده‌ها[۱]، تبدیل داده‌ها[۲]، گسسته‌سازی داده‌ها[۳] و ارزیابی الگو[۴] است. روش‌های مختلف داده‌کاوی برای کشف روابط پنهان و نامعلوم بین داده‌ها استفاده می‌شود و برای مواردی از قبیل بازاریابی، فروش، کشف تقلب[۵]، اکتشافات علمی، توسعه محصول، مراقبت های بهداشتی[۶] و آموزش استفاده می‌شود. در ادامه برخی از روش‌های داده‌کاوی معرفی می‌شوند [۶۲].

۱-۶-۱- یادگیری قانون انجمن[۷]

در علم داده، روش AR برای کشف همبستگی بین پایگاه‌های داده و مجموعه‌داده‌های به ظاهر مستقل و مشاهده الگوهای متداول استفاده می‌شود. یک قانون انجمن دارای دو بخش است: بخش مقدم[۸] (اگر[۹]) و بخش نتیجه[۱۰] (پس[۱۱]). بخش مقدم آن چیزی است که در داده‌ها یافت می‌شود و نتیجه بخشی است که در ترکیب با بخش مقدم

---

۱ data integration
۲ data transformation
۳ data discretization
۴ pattern evaluation
۵ fraud detection
۶ healthcare
۷ AR Learning
۸ antecedent
۹ if
۱۰ consequent
۱۱ then

یافت می‌شود. دو الگوی اصلی که در AR استفاده می‌شود پشتیبانی[۱] و اطمینان[۲] هستند (که معیارهایی هستند که توسط کاربر تعریف می‌شوند) [۶۳]. پشتیبانی میزان دفعات ظاهر شدن یک مجموعه آیتم[۳] در مجموعه‌داده است (که در آن برای یک قانون معین، مجموعه آیتم‌ها فهرستی از همه موارد موجود در بخش‌های مقدم و نتیجه آن است). به عبارت دیگر، پشتیبانی نشان‌دهنده فراوانی قانون در معاملات است. مقدار بالا به این معنی است که این قانون شامل ارزش زیادی از پایگاه‌داده است. اطمینان معیاری است برای احتمال وقوع یک نتیجه[۴] [۶۲].
متداول‌ترین الگوریتم‌های مورد استفاده در روش‌های قانون انجمن عبارتند از «الگوریتم Apriori»، «الگوریتم Eclat» و «الگوریتم رشد الگوی مکرر[۵]». استخراج قانون انجمن برای مواردی از قبیل داده‌های غیرعددی و طبقه‌بندی شده در تجزیه‌وتحلیل سبد بازار، تشخیص پزشکی، تجزیه‌وتحلیل داده‌های سرشماری[۶] و مدیریت ارتباط با مشتری[۷] مناسب است [۶۲].

۱-۶-۲- طبقه‌بندی

طبقه‌بندی، پیش‌بینی یک متغیر تصادفی گسسته (Y) با استفاده از متغیر تصادفی دیگر (X) است و گاهی اوقات به آن تمایز[۸]، طبقه‌بندی الگو یا تشخیص الگو[۹] می‌گویند. طبقه‌بندی روشی است که داده‌ها را به تعداد معینی از کلاس‌ها دسته‌بندی کرده و برچسبی به هر کلاس تخصیص می‌دهد. ایده اصلی الگوریتم‌های طبقه‌بندی، پیش‌بینی کلاس هدف با تجزیه‌وتحلیل مجموعه‌داده‌های آموزشی، یعنی طبقه‌بندی داده‌ها در تعداد معینی از کلاس‌ها است. در واقع از مجموعه‌داده‌های آموزشی برای بدست آوردن شرایط مرزی بهتر[۱۰] و اختصاص داده‌های جدید به دسته‌ها یا کلاس‌های از پیش تعیین‌شده استفاده می‌شود [۶۴]، [۶۵].

---

[۱] support

[۲] confidence

[۳] itemset

[۴] Confidence is the measure of the likeliness of occurrence of consequence on the cart given that the cart already has the antecedents.

[۵] Frequent Pattern (FP) Growth Algorithm

[۶] Census data analysis

[۷] Customer Relationship Management (CRM)

[۸] discrimination

[۹] pattern recognition

[۱۰] better boundary conditions

به طور کلی، انواع الگوریتم‌های طبقه‌بندی عبارت هستند از: طبقه‌بندی‌کننده‌های خطی[۱]، ماشین‌های بردار پشتیبانی[۲]، طبقه‌بندی‌کننده‌های درجه دوم[۳]، تخمین کرنل[۴]، القای درخت تصمیم[۵]، شبکه‌های عصبی[۶]، آموزش کوانتیزاسیون برداری[۷] و روش درخت تصمیم تقویت‌شده[۸] [۶۲].

۱-۶-۳- تجزیه‌وتحلیل رگرسیون

تحلیل رگرسیون یک روش یادگیری آماری[۹] شناخته شده است که برای تخمین رابطه بین یک متغیر وابسته با یک یا چند متغیر مستقل استفاده می‌شود که در آن متغیر مستقل به عنوان ورودی فرضی استفاده شده و با هدف بررسی اثر آن بر یک متغیر وابسته تغییر می‌کند. به عبارت دیگر، تحلیل رگرسیون یک فرآیند داده‌کاوی است که به درک همبستگی و استقلال متغیرها کمک می‌کند تا مشخص شود کدام عوامل بیشتر اهمیت دارند و کدام عوامل را می‌توان نادیده گرفت و در نهایت، چگونه این عوامل بر یکدیگر تأثیر می‌گذارند [۶۲].
بسته به عواملی مانند نوع متغیر هدف، شکل خط رگرسیون و تعداد متغیرهای مستقل، انواع مختلفی از روش‌های تجزیه‌وتحلیل رگرسیون توسعه داده شده است. برخی از انواع روش‌های رگرسیون عبارت هستند از: رگرسیون خطی[۱۰]، رگرسیون لجستیک[۱۱]، رگرسیون لاسو[۱۲] و رگرسیون چندجمله‌ای[۱۳] [۶۲].

۱-۶-۴- تشخیص داده‌های پرت[۱۴]

نقاط پرت داده‌هایی هستند که از الگوی کلی یک نمونه متفاوت هستند و عمدتاً نشان‌دهنده تغییرپذیری در اندازه‌گیری[۱۵]، خطاهای آزمایشی[۱۶] یا داده‌های جدید[۱۷] است. نقاط پرت می‌توانند تک متغیره (در فضای تک‌بعدی) یا چند متغیره (در فضای n بعدی) باشند [۶۲].

---

۱ Linear Classifiers
۲ Support Vector Machines (SVM)
۳ Quadratic classifiers
۴ Kernel Estimation
۵ Decision Trees Induction
۶ Neural Networks
۷ Learning Vector Quantization
۸ Boosted Decision Trees Method
۹ statistical learning technique
۱۰ Linear Regression
۱۱ Logistic Regression
۱۲ Lasso Regression
۱۳ Polynomial Regression
۱۴ Outlier Detection
۱۵ variability in a measurement
۱۶ experimental errors
۱۷ novelty

نقاط پرت را می‌توان بسته به انواع مختلف داده بدست آورد، مانند نقاطی که دور از سایر نقاط هستند، نقاطی که می‌توانند به دلیل نویز در داده‌ها باشند (به‌عنوان مثال، سیگنال نویز در پس‌زمینه هنگام تشخیص گفتار)، یا موارد پرتی مانند سیگنالی که ممکن است نشان‌دهنده کشف پدیده‌های جدید باشد [۶۲].

رایج‌ترین علل وجود داده‌های پرت در مجموعه‌داده‌ها می‌تواند به دلیل خطای ورود داده‌ها (به دلیل اشتباه انسانی یا خطای اندازه‌گیری به دلیل دقت ابزار)، خطاهای آزمایشی، پردازش داده‌ها و خطاهای نمونه‌برداری باشد. بنابراین، کیفیت داده‌ها به اندازه کیفیت مدل پیش‌بینی یا طبقه‌بندی مهم است. به همین دلیل است که تشخیص نقاط پرت از اهمیت زیادی برخوردار است [۶۲].

برخی از رایج‌ترین روش‌ها برای تشخیص نقاط پرت عبارتند از [۶۲]: Z-score یا تجزیه‌وتحلیل ارزش افراطی (پارامتریک)[۱]، مدل‌سازی احتمالی و آماری (پارامتریک)[۲]، مدل‌های رگرسیون خطی مانند تجزیه‌وتحلیل مؤلفه اصلی[۳] و حداقل میانه مربع‌ها[۴] [۶۶]، مدل‌های مبتنی بر مجاورت (ناپارامتریک)[۵] و مدل‌های تئوری اطلاعات[۶].

۱-۶-۵- روش‌های مدل‌سازی پیشگویانه

منظور از تجزیه‌وتحلیل پیشگویانه استفاده از روش‌های آماری و مبتنی بر یادگیری ماشین برای شناسایی احتمال نتایج آینده[۷] بر اساس داده‌های تاریخی با هدف ساده‌سازی تصمیم‌گیری[۸] و تولید بینش‌های جدید است. تجزیه‌وتحلیل پیشگویانه برای پیش‌بینی رفتار و روندها، درک بهتر مشتریان و بهبود تصمیم‌گیری و عملکرد تجاری استفاده می‌شود. برخی از حوزه‌های کاربردی تجزیه‌وتحلیل پیشگویانه عبارت هستند از: کشف تقلب و امنیت، بازاریابی، بهره‌برداری و شناسایی ریسک. مدل‌های طبقه‌بندی و خوشه‌بندی پرکاربردترین مدل‌ها در تجزیه‌وتحلیل پیشگویانه هستند. با استفاده از مدل خوشه‌بندی، می‌توان مشتریان را بر اساس ویژگی‌های مشترک به گروه‌های مشابه تقسیم کرد و برای هر گروه استراتژی‌هایی طراحی کرد. علاوه بر این، داده‌های پرت و مدل‌های سری زمانی نیز به‌عنوان روش‌های پیشگویانه استفاده می‌شوند که طی آنها داده‌های ورودی غیرعادی در یک مجموعه‌داده شناسایی می‌شوند یا با استفاده از زمان به عنوان پارامتر ورودی، دنباله‌ای از نقاط داده شناسایی می‌شوند [۶۲].

به طور کلی، الگوریتم‌های پیشگویانه رایج را می‌توان به دو گروه تقسیم کرد: یادگیری ماشینی و یادگیری عمیق[۹]. یادگیری ماشینی شامل داده‌های ساختاری است و هر دو نوع خطی و غیرخطی را شامل می‌شود و فرآیند یادگیری آن سریع است. یادگیری عمیق زیرمجموعه‌ای از یادگیری ماشینی است که برای سر و کار داشتن با صدا، ویدئو، متن و تصاویر[۱۰] محبوب‌تر است. در مدل‌سازی پیشگویانه یادگیری ماشین، الگوریتم‌های مختلفی وجود دارد که

---

۱ Z-score or Extreme Value Analysis (parametric)

۲ Probabilistic and Statistical Modeling (parametric)

۳ Principal Component Analysis (PCA)

۴ Least Median of Squares (LMS)

۵ the Proximity Based Models (non-parametric)

۶ Information Theory Models

۷ likelihood of future outcomes

۸ streamline decision making

۹ Deep Learning

۱۰ audio, video, text, and images

می‌توان از آنها استفاده کرد که رایج‌ترین آنها روش جنگل تصادفی[۱]، مدل خطی تعمیم‌یافته برای دو مقدار[۲]، مدل تقویت شده گرادیان[۳]، K-Means و الگوریتم Prophet هستند [۶۲].

۱-۶-۶- الگوهای متوالی

مشابه روش استخراج قواعد انجمنی، با استفاده از الگوهای متوالی، می‌توان الگوها و قوانین جالب و مفید آماری را در جدولی در مقیاس بزرگ کشف کرد که شامل دنباله‌ای از معاملات است[۴] [۶۷]. الگوی متوالی یک زیردنباله مکرر[۵] است که در یک دنباله یا مجموعه‌ای از دنباله‌ها وجود دارد.

الگوریتم الگوهای متوالی روشی است که برای کشف الگوهای جالب در داده‌های متوالی مانند داده‌های سری زمانی یا یک توالی از رویدادها استفاده می‌شود. این تکنیک بر تشخیص الگوها و ترندهای تکراری که در طول زمان یا در یک توالی اتفاق می‌افتند تمرکز دارد. روش الگوهای متوالی شامل تحلیل مجموعه‌داده‌ای است که شامل یک سری توالی از آیتم‌ها یا رویدادها است، که هر توالی نمایانگر یک موجودیت یا فرآیند خاص است. به عنوان مثال، می‌تواند یک توالی از تراکنش‌های مشتریان، کلیک‌های وب یا توالی‌های DNA باشد [۶۲].

هدف از استخراج الگوهای متوالی کشف زیرتوالی‌هایی است که به طور مکرر در مجموعه داده تکرار می‌شوند. این زیرتوالی‌ها می‌توانند اندازه‌ها و طول‌های مختلفی داشته باشند و ممکن است بر روی توالی‌های مختلفی انتشار یابند. الگوهای کشف شده می‌توانند بینش‌های ارزشمندی درباره رفتار، وابستگی‌ها و روندهای موجود در داده ارائه دهند [۶۲].

فرآیند استخراج الگوهای متوالی معمولاً شامل مراحل زیر است [۶۲]:

۱. آماده‌سازی داده: در این مرحله داده‌های ورودی به صورت چند توالی سازماندهی می‌شوند؛ به طوری که هر توالی شامل یک سری از آیتم‌ها یا رویدادها باشد. داده‌ها ممکن است شامل زمان یا ویژگی‌هایی باشند که با هر آیتم یا رویداد مرتبط هستند.

۲. محاسبه پشتیبان[۶]: پشتیبان الگو به تکرار الگو در مجموعه داده اشاره دارد. برای شناسایی الگوهای متداول، پشتیبان هر الگوی احتمالی محاسبه می‌شود. یک آستانه پشتیبان حداقل[۷] تعیین می‌شود و الگوهایی که این آستانه را برآورده یا تجاوز می‌کنند، به عنوان الگوهای قابل‌توجه در نظر گرفته می‌شوند.

۳. تولید کاندید[۸]: پس از آنکه الگوریتم الگوهای متوالی، الگوهای کاندید را با بررسی آیتم‌ها یا رویدادها تولید کرد، با استفاده از الگوهای متداول کشف‌شده در تکرارهای قبلی، الگوهای کاندید بلندتری را تولید می‌کند.

---

[۱] Random Forest

[۲] Generalized Linear Model (GLM) for two Values

[۳] Gradient Boosted Model (GBM)

[۴] sequences of transactions

[۵] frequent subsequence

[۶] Support Calculation

[۷] minimum support threshold

[۸] Candidate Generation

۴. ارزیابی الگو[۱]: الگوهای کاندید تولید شده با توجه به مجموعه داده بررسی می‌شوند تا پشتیبان آنها محاسبه شود. الگوهایی که آستانه پشتیبان را برآورده یا تجاوز می‌کنند، به عنوان الگوهای متداول شناخته شده و به عنوان الگوهای معتبر ذخیره می‌شوند.
۵. انتخاب الگو[۲]: الگوهای متداول کشف شده ممکن است بی‌شمار بوده و برای تحلیل و شناخت عملی نباشند. بنابراین، معیارها یا محدودیت‌های تعریف شده توسط کاربر می‌توانند برای انتخاب الگوهای مناسب استفاده شوند.
۶. تفسیر الگو[۳]: مرحله نهایی تفسیر و تحلیل الگوهای متوالی کشف شده است. این مرحله ممکن است شامل درک روابط بین الگوهای مختلف، شناسایی ناهنجاری‌ها یا موارد پرت یا پیش‌بینی بر اساس الگوها باشد.
برخی از الگوریتم‌هایی که می‌توانند برای استخراج الگوهای متوالی استفاده شوند عبارت هستند از: الگوریتم‌های Apriori-like، الگوریتم‌های مبتنی بر BFS[۴]، الگوریتم‌های مبتنی بر DFS[۵]، الگوریتم‌های مبتنی بر الگوی ترتیبی بسته[۶] و الگوریتم‌های مبتنی بر افزایش[۷] [۶۸]، [۶۹]. هر الگوریتم ممکن است ویژگی‌ها، کارایی و تناسب متفاوتی برای انواع مختلف داده‌ها داشته باشد.
به طور کلی، تکنیک الگوهای متوالی ابزار قدرتمندی برای کشف دانش و بینش‌های پنهان در داده‌های متوالی است که برنامه‌های کاربردی را در حوزه‌های مختلف مانند تجزیه‌وتحلیل سبد بازار، تجزیه‌وتحلیل رفتار مشتری، وب‌کاوی، بیوانفورماتیک[۸] و غیره فعال می‌کند [۶۲].

۱-۷ حفظ حریم خصوصی و ملاحظات اخلاقی در داده‌کاوی[۹]

۱-۷-۱- مقدمه

فناوری‌های محاسباتی ابعاد وسیعی از زندگی انسان‌ها را فراگرفته و حجم زیادی از داده‌ها را تولید می‌کنند که برای تجزیه‌وتحلیل در دسترس هستند. با این حال، این نگرانی وجود دارد که اگر داده‌های جمع‌آوری‌شده قبل از انتشار برای بررسی به‌طور مناسب پاکسازی[۱۰] نشوند، ممکن است داده‌های حساس[۱۱] عمومی‌شوند [۳۳]. در این راستا، روش‌های حفظ حریم خصوصی[۱۲] متعددی در مطالعات مختلف ارائه شده است.
جمع‌آوری داده‌ها در مقیاس بزرگ تنها در صورتی می‌تواند معنا داشته باشد که قابل استفاده باشند و بتوان از آنها در تصمیم‌گیری استفاده کرد [۳۴]. اینجاست که داده‌کاوی با بررسی روابط نامعلوم در داده‌ها و ارائه بینش مفید به صاحبان داده کمک می‌کند. علاوه بر این، چنین قابلیت‌هایی ممکن است برای تجزیه‌وتحلیل بیشتر نیاز به

---

۱ Pattern Evaluation
۲ Pattern Selection
۳ Pattern Interpretation
۴ BFS (Breadth First Search)-based algorithms
۵ DFS (Depth First Search)-based algorithms
۶ closed sequential pattern-based algorithms
۷ incremental-based algorithms
۸ bioinformatics
۹ Privacy and Ethical Considerations in Data Mining
۱۰ sanitized
۱۱ potentially sensitive data
۱۲ privacy-preserving methods

اشتراک‌گذاری با طرف‌های خارجی داشته باشد. در این فرآیند، انواع مختلفی از اطلاعات ممکن است فاش شوند که این امر می‌تواند منجر به نقض حریم خصوصی شود. توانایی به اشتراک‌گذاری اطلاعات در حین جلوگیری از افشای اطلاعات یکی از مهم ترین چالش‌های فنی، قانونی، اخلاقی و اجتماعی است. در واقع، سازمان‌های دولتی و تجاری مختلف، مقادیر زیادی از داده‌های کاربر، از جمله اطلاعات اعتباری فردی، سلامت، وضعیت مالی و ترجیحات شخصی[۱] را جمع‌آوری می‌کنند. شبکه‌های اجتماعی، بانک‌داری و سیستم‌های مراقبت بهداشتی نمونه‌هایی از سیستم‌هایی هستند که چنین اطلاعات خصوصی را مدیریت می‌کنند [۳۵] و اغلب به دلیل استفاده غیرمستقیم از اطلاعات خصوصی، حریم خصوصی را نادیده می‌گیرند. سیستم‌های اطلاعاتی دیگری نیز وجود دارند که از حجم عظیمی از اطلاعات خصوصی حساس برای مدل‌سازی و پیش‌بینی پدیده‌های مرتبط با انسان (مانند جنایات [۳۶]، اپیدمی‌ها [۳۷] و چالش‌های بزرگ در فیزیک اجتماعی[۲] [۳۸]) استفاده می‌کنند. از این رو، حفظ حریم خصوصی می‌تواند به یک مشکل بسیار پیچیده تبدیل شود و نیاز به راه‌حل‌های قوی دارد [۳۹] و [۴۰].

داده‌کاوی با حفظ حریم خصوصی[۳] امکان استفاده از روش‌های داده‌کاوی بدون افشای اطلاعات خصوصی را ارائه می‌دهد. رویکردهای PPDM شامل مواردی از قبیل اغتشاش داده‌ها[۴] (اصلاح داده‌ها[۵]) [۴۱، ۴۲] و رمزگذاری[۶] [۴۳] است. روش‌های رمزنگاری[۷] برای ایمن‌سازی داده‌ها مشهور هستند. در مطالعات پیشین مثال‌های زیادی وجود دارد که در PPDM به طور موثر از روش‌های رمزنگاری استفاده شده است [۴۴].

تعریف اثربخشی روش‌های حفظ حریم خصوصی با استفاده از مدل حریم خصوصی[۸] و شناسایی محدودیت‌های حفاظت و افشای اطلاعات خصوصی ضروری است [۳۵]. K−anonymity، l−diversity، (α,k)−anonymity، و t−closeness برخی از مدل‌های حریم خصوصی هستند که آسیب‌پذیری در برابر حملات مختلف را نشان می‌دهند [۳۵].

DP[۹] یک مدل حریم خصوصی است که با به حداقل رساندن احتمال شناسایی داده‌های شخصی، حداکثر حریم خصوصی را ارائه می‌دهد [۴۵]. DP محلی[۱۰]، که با اغتشاش ورودی[۱۱] [۴۵] به دست می‌آید، با تصادفی کردن نمونه‌های شخصی یک پایگاه‌داده[۱۲] [۴۶] به تحلیلگران اجازه انتشار کامل یا جزئی داده‌ها[۱۳] را می‌دهد [۴۷]. DP

---

۱ personal preferences

۲ social physics

۳ Privacy-preserving data mining (PPDM)

۴ data perturbation

۵ data modification

۶ encryption

۷ Cryptographic methods

۸ privacy model

۹ Differential Privacy (DP)

۱۰ Local differential privacy (LDP)

۱۱ input perturbation

۱۲ by randomizing the individual instances of a database

۱۳ full or partial data release

جامع[۱] که به آن مدل متصدی قابل اعتماد[۲] نیز می‌گویند، به تحلیل‌گران اجازه می‌دهد فقط از متصدی[۳] بخواهند تا درخواست‌هایی[۴] را در پایگاه‌داده اجرا کنند. متصدی نویزی که با دقت کالیبره شده است را در نتایج درخواست‌ها اعمال می‌کند تا DP را فراهم کند [۵۴، ۴۷]. با این حال، GDP و LDP برای مجموعه‌داده‌های کوچک مناسب نیستند [۳۳].

اگرچه DP به طور کامل مورد مطالعه قرار گرفته است، تنها چند راه‌حل مناسب برای انتشار کامل/جزئی داده با استفاده از LDP وجود دارد. پایه محکم و از لحاظ نظری جذاب DP برای حفاظت از حریم خصوصی، عملی بودن اجرای راه‌حل‌های کارآمد برای داده‌های بزرگ را محدود کرده است. علاوه بر این، الگوریتم‌های LDP موجود شامل مقدار قابل‌توجهی از اضافه کردن نویز (به عنوان مثال تصادفی‌سازی) است که منجر به کاربرد کم داده می‌شود. بر این اساس، مطلوبیت و حریم خصوصی اغلب به عنوان عوامل متضاد ظاهر می‌شوند و بهبود حریم خصوصی معمولاً مستلزم کاهش مطلوبیت است [۳۳].

در ادامه برخی از راهکارهای حفظ حریم خصوصی به طور خلاصه معرفی می‌شوند.

۱-۷-۲- نمونه‌هایی از راهکارهای حفظ حریم خصوصی

حفاظت از حریم خصوصی افراد با گسترش فناوری‌های مشترکین با قابلیت اتصال به اینترنت[۵] به یک وظیفه چالش‌برانگیز تبدیل شده است. مطالعات پیشین رویکردهای متفاوتی را برای یافتن راه‌حل این چالش نشان می‌دهد. برخی از رویکردها بر افزایش آگاهی تمرکز می‌کنند [۴۸] و برخی دیگر سعی می‌کنند از روش‌های مختلفی برای حفظ حریم خصوصی فردی استفاده کنند [۴۹]. حجم عظیم داده‌ها چالش‌های زیادی را برای حفظ حریم خصوصی ایجاد می‌کند [۵۰]. اگرچه نگرانی‌های امنیتی و حفظ حریم خصوصی داده‌های بزرگ کاملاً جدید نیستند، اما به دلیل ویژگی‌های محیط‌ها و پویایی دستگاه‌های مورد استفاده همچنان نیاز به توجه و تحقیق دارند [۵۱]. پیشرفت‌های این محیط‌ها و تنوع دستگاه‌ها همواره باعث افزایش پیچیدگی می‌شود و امنیت و حفظ حریم خصوصی را پیچیده می‌کند. برای مقابله با چالش‌ها و پیچیدگی‌های متنوع، سه رویکرد فن‌آوری متفاوت را می‌توان مشاهده کرد: کنترل افشا[۶]، PPDM و فن‌آوری‌های افزایش حریم خصوصی[۷] [۵۲]. رمزگذاری مبتنی بر ویژگی[۸]، کنترل دسترسی از طریق احراز هویت[۹]، کنترل دسترسی موقت و مبتنی بر مکان[۱۰] و استفاده از پروتکل‌های مبتنی بر محدودیت[۱۱] مکانیزم‌هایی هستند که برای بهبود حریم خصوصی سیستم‌ها در محیط‌های پویا مورد استفاده قرار

---

۱ Global differential privacy (GDP)

۲ trusted curator model

۳ curator

۴ queries

۵ proliferation of Internetenabled consumer technologies

۶ disclosure control

۷ privacy-enhancing technologies

۸ Attribute-based encryption

۹ controlling access via authentication

۱۰ temporal and location-based access control

۱۱ employing constraint-based protocols

می‌گیرند [۳۵]. در میان رویکردهای مختلف PPDM، اغلب اغتشاش داده[۱] به دلیل سادگی و کارایی آن ترجیح داده می‌شود [۵۳]. هر دو حالت اغتشاش ورودی و خروجی اغلب مورد استفاده قرار می‌گیرند. اغتشاش خروجی مبتنی بر اضافه کردن نویز[۲] و پنهان کردن قوانین[۳] است؛ در حالی که اغتشاش ورودی یا با اضافه کردن [۵۴] یا با ضرب نویز[۴] [۳۵] انجام می‌شود.

اغتشاش ورودی را می‌توان به اغتشاش تک‌بعدی[۵] و چندبعدی[۶] تقسیم کرد [۵۵]. اغتشاش افزایشی[۷] [۵۴]، پاسخ تصادفی[۸] [۴۵]، تعویض[۹] [۵۶] و ریزتجمیع[۱۰] [۵۷] نمونه‌هایی از اغتشاش ورودی تک‌بعدی هستند؛ در حالی که تراکم[۱۱] [۵۸]، چرخش تصادفی[۱۲] [۵۹]، اغتشاش هندسی[۱۳] [۴۲]، طرح ریزی تصادفی[۱۴] [۶۰] و اغتشاش ترکیبی[۱۵] ترکیبی[۱۵] نمونه‌هایی از نوع چندبعدی هستند [۵۳].

در اغتشاش افزایشی، نویز تصادفی به داده‌های اصلی اضافه می‌شود به گونه‌ای که ویژگی‌های آماری اساسی[۱۶] حفظ شوند. یک مشکل مهم در این رویکرد، کاربرد کم داده‌های حاصل است [۶۱]. علاوه بر این، روش‌های موثر بازسازی نویز که در پاسخ ایجاد شده‌اند، می‌توانند به طور قابل‌توجهی سطح حریم خصوصی را کاهش دهند [۵۵]. روش‌های تصادفی‌سازی[۱۷] مانند پاسخ تصادفی یکی دیگر از رویکردها [۴۵] است (به عنوان مثال، تصادفی‌سازی پاسخ‌های مصاحبه‌شوندگان به منظور حفظ حریم خصوصی پاسخ‌دهندگان). به دلیل تصادفی‌سازی بالای داده‌های ورودی، روش‌های تصادفی‌سازی مانند پاسخ تصادفی اغلب حریم خصوصی بالایی را ارائه می‌کنند در حالی که سودمندی از نظر تخمین آمار[۱۸] یا انجام تحلیل‌ها[۱۹] می‌تواند کم باشد [۴۵]. ریزتجمیع بر اساس قوانین محرمانه است که اجازه انتشار مجموعه‌داده‌های خرد[۲۰] را می‌دهد. ریزتجمیع مجموعه‌داده را به خوشه‌هایی از k عنصر تقسیم می‌کند و مقادیر هر خوشه را با مرکز خوشه جایگزین می‌کند [۳۳].

---

۱ data perturbation

۲ noise addition

۳ rule hiding

۴ noise addition or multiplication

۵ unidimensional perturbation

۶ multidimensional perturbation

۷ Additive perturbation

۸ randomized response

۹ swapping

۱۰ microaggregation

۱۱ condensation

۱۲ random rotation

۱۳ geometric perturbation

۱۴ random projection

۱۵ hybrid perturbation

۱۶ underlying statistical properties

۱۷ Randomization techniques

۱۸ estimating statistics

۱۹ conducting analyses

۲۰ micro datasets

هنگامی که داده‌های منتشرشده شامل اطلاعاتی در مورد روش حفاظت و پارامترهای آن باشد، ریزتجمیع به یک متغیر واحد (ریزتجمیع تک متغیره[۱]) در برابر حملات شفافیت[۲] آسیب‌پذیر است [۵۷]. ریزتجمیع چند متغیره[۳] نیز پیشنهاد شده است، اما پیچیده است و ثابت شده است که NP-hard است [۵۷].

در تراکم، مجموعه‌داده ورودی به چندین گروه با اندازه از پیش تعریف‌شده تقسیم می‌شود، به‌گونه‌ای که تفاوت بین رکوردها در یک گروه خاص حداقل است و سطح مشخصی از اطلاعات آماری در مورد رکوردهای مختلف در هر گروه حفظ می‌شود. سپس داده‌های پاکسازی‌شده با استفاده از توزیع تصادفی یکنواخت[۴] بر اساس بردارهای ویژه[۵] که با استفاده از تجزیه ویژه ماتریس‌های کوواریانس مشخصه هر گروه همگن[۶] ایجاد شده‌اند، تولید می‌شوند [۵۸]. تراکم یک نقص قابل‌توجه دارد، زیرا ممکن است کیفیت داده‌ها را به میزان قابل‌توجهی کاهش دهد [۳۳].

توضیحات بیشتر و جزئی‌تر در مورد انواع مختلف روش‌های موجود جهت حفظ محرمانگی خارج از هدف اصلی این گزارش است و در اینجا صرفا چند مورد به عنوان نمونه معرفی شد. جهت آشنایی بیشتر با انواع روش‌ها در این حوزه می‌توان به مراجع موجود در این حوزه از قبیل [۳۳] مراجعه کرد.

۱-۸ چالش‌ها و جهت‌گیری‌های آینده

۱-۸-۱- داده‌کاوی و کلان‌داده

در این راستا، چندین جنبه کلیدی شایسته توجه است:

- چالش‌ها و فرصت‌ها در ارتباط با کلان‌داده: حجم، سرعت و تنوع روزافزون داده‌های تولید شده در دنیای امروز، گنجینه‌ای از اطلاعات را فراهم می‌کند. با این حال، استفاده از این داده‌ها برای کسب بینش معنادار یک چالش بزرگ است. داده‌کاوی در زمینه کلان‌داده باید بتواند با مسائل مربوط به کیفیت داده‌ها، حریم خصوصی و امنیت، و نیز استفاده از روش‌های پیشرفته جهت مدیریت حجم عظیمی از داده‌ها، دست و پنجه نرم کند. علاوه بر این، تنوع منابع داده، از جمله داده‌های ساختاریافته، بدون ساختار و نیمه‌ساختار یافته[۷]، نیازمند رویکردهای نوآورانه برای یکپارچه‌سازی و تجزیه‌وتحلیل است.
- مقیاس‌پذیری، ذخیره‌سازی و ملاحظات محاسباتی[۸]: مقیاس بزرگ داده‌های در دسترس مستلزم آن است که روش‌های داده‌کاوی بسیار مقیاس‌پذیر باشند. الگوریتم‌های سنتی ممکن است برای

---

[۱] univariate microaggregation

[۲] transparency attacks

[۳] Multivariate microaggregation

[۴] uniform random distribution

[۵] eigenvectors

[۶] eigendecomposition of the characteristic covariance matrices of each homogeneous group

[۷] structured, unstructured, and semi-structured data

[۸] Scalability, storage, and computational considerations

مواجهه با مجموعه‌داده‌هایی در حدود ترابایت[1] یا حتی پتابایت[2]، مشکل داشته باشند. بنابراین، استفاده از راه‌حل‌های مقیاس‌پذیر داده‌کاوی ضروری است که می‌تواند مواردی از قبیل محاسبات توزیع‌شده، پردازش موازی و روش‌های ذخیره‌سازی و بازیابی داده را در بر داشته باشد. این راه‌حل‌ها باید از نظر محاسباتی کارا باشند، زیرا پردازش کلان‌داده‌ها می‌تواند زمان‌بر و سنگین باشد.

در آینده، حوزه داده‌کاوی به تکامل خود ادامه خواهد داد تا نیازهای کلان‌داده‌ها را برآورده کند. محققان و متخصصان باید روش‌ها و ابزارهای جدیدی را برای استخراج بینش معنادار از کلان‌داده‌ها توسعه دهند. علاوه بر این، پرداختن به نگرانی‌های اخلاقی و حفظ حریم خصوصی مرتبط با داده‌کاوی، یک چالش مداوم خواهد بود که نیازمند تعادل بین نوآوری و مدیریت مسئولانه داده[3] است.

۱-۸-۲- داده‌کاوی بلادرنگ[4]

داده‌کاوی بلادرنگ حوزه نوظهوری در زمینه تجزیه‌وتحلیل داده است که بر پردازش و استخراج بینش از داده‌های درحین‌تولید تمرکز دارد. این بخش به اهمیت داده‌کاوی بلادرنگ و روش‌های به‌کاررفته برای تصمیم‌گیری به موقع می‌پردازد. بنابراین، لازم است موارد زیر مدنظر قرار داشته باشند:

- داده‌کاوی بلادرنگ و تحلیل جریان داده: داده‌کاوی بلادرنگ به دلیل گسترش برنامه‌ها و سیستم‌هایی که نیاز به پاسخ فوری به رویدادهای داده را دارند، به طور فزاینده‌ای حیاتی است. بر خلاف داده‌کاوی سنتی که اغلب با مجموعه‌داده‌های ایستا یا تاریخی سروکار دارد، داده‌کاوی بلادرنگ داده‌ها را هنگام تولید پردازش و تجزیه‌وتحلیل می‌کند. این رویکرد به کسب و کارها و سازمان‌ها این امکان را می‌دهد که به سرعت به شرایط در حال تغییر واکنش نشان دهند، ناهنجاری‌ها را شناسایی کنند و الگوها را در زمان واقعی شناسایی کنند. تجزیه‌وتحلیل جریان داده یک جزء کلیدی داده‌کاوی بلادرنگ است که امکان پردازش مداوم داده‌ها و تصمیم‌گیری سریع را فراهم می‌کند.
- روش‌های تجزیه‌وتحلیل داده‌ها در زمان واقعی و تصمیم‌گیری به موقع: روش‌های داده‌کاوی بلادرنگ شامل پردازش جریان داده، تشخیص رویداد و پردازش رویدادهای پیچیده است. سیستم‌های پردازش جریانی می‌توانند جریان‌های داده‌ای با سرعت بالا[5]، مانند داده‌های حاصل از حسگرها، رسانه‌های اجتماعی یا تراکنش‌های مالی را مدیریت کنند. الگوریتم‌های تشخیص رویداد، رویدادها یا ناهنجاری‌های خاص در جریان‌های داده را شناسایی و پاسخ‌های مناسب را می‌دهند. پردازش رویداد پیچیده[6] داده‌ها را از چندین منبع ترکیب می‌کند تا الگوهای پیچیده را شناسایی کرده و در زمان واقعی تصمیم‌گیری کند. یادگیری ماشین و مدل‌سازی پیش‌گویانه نیز برای پیش‌بینی رویدادهای آینده بر اساس داده‌های دریافتی استفاده می‌شوند.

---

[1] terabytes

[2] petabytes

[3] responsible data handling

[4] Real-Time Data Mining

[5] high-velocity data streams

[6] Complex event processing (CEP)

داده‌کاوی بلادرنگ در حوزه‌های مختلفی از جمله مالی (برای تشخیص تقلب)، مراقبت‌های بهداشتی (برای نظارت بر بیمار) و امنیت سایبری (برای تشخیص تهدید) استفاده می‌شود. با ادامه رشد حجم و پیچیدگی جریان‌های داده، روش‌های داده‌کاوی بلادرنگ نقشی اساسی در تصمیم‌گیری به موقع و آگاهانه ایفا خواهند کرد.

۱-۸-۳- رویکردهای میان‌رشته‌ای[۱]

رویکردهای میان‌رشته‌ای در داده‌کاوی بر ماهیت مشارکتی این رشته تأکید می‌کند و دامنه آن را به حوزه‌های مختلفی مانند علوم اجتماعی[۲]، زیست‌شناسی[۳] و اعصاب[۴] گسترش می‌دهد. این بخش اهمیت فزاینده همکاری میان‌رشته‌ای و استفاده از دانش حوزه را برای بهبود نتایج داده‌کاوی بررسی می‌کند. بنابراین، لازم است موارد زیر مدنظر قرار گیرد:

- بررسی ماهیت میان‌رشته‌ای داده‌کاوی: داده‌کاوی، به دلیل ماهیت خود، از مرزهای رشته‌ای فراتر می‌رود. این مساله شامل استخراج بینش و دانش از داده‌ها است که می‌تواند در طیف گسترده‌ای از زمینه‌ها اعمال شود. از پیش‌بینی شیوع بیماری در اپیدمیولوژی[۵] گرفته تا تجزیه‌وتحلیل رفتار مصرف‌کننده در بازاریابی، داده‌کاوی ابزارهای همه‌کاره را برای کاوش بینش‌های مبتنی بر داده ارائه می‌دهد. جنبه میان‌رشته‌ای داده‌کاوی شامل تطبیق روش‌های آن برای مطابقت با چالش‌های منحصر به فرد ارائه شده در حوزه‌های مختلف است.
- همکاری بین دانش داده‌کاوی و سایر زمینه ها: همکاری بین کارشناسان داده‌کاوی و متخصصان در سایر حوزه‌ها به پیشرفت‌های قابل‌توجهی منجر شده است. به عنوان مثال، روش‌های داده‌کاوی در ژنومیک[۶] برای شناسایی نشانگرهای ژنتیکی[۷] مرتبط با بیماری‌ها استفاده می‌شود. در علوم اجتماعی، داده‌کاوی به تجزیه‌وتحلیل احساسات و درک رفتار انسان از طریق تجزیه‌وتحلیل داده‌های رسانه‌های اجتماعی کمک می‌کند. ارتباط داده‌کاوی با علوم اعصاب منجر به پیشرفت‌هایی در رابط‌های مغز و کامپیوتر و تجزیه‌وتحلیل داده‌های عصبی شده است.
- استفاده از دانش حوزه‌ای[۸] برای بهبود نتایج داده‌کاوی: دانش حوزه‌ای در پروژه‌های داده‌کاوی بسیار ارزشمند است و زمینه و بینش‌هایی را ارائه می‌کند که ممکن است روش‌های مبتنی بر داده به تنهایی نتوانند آنها را درک کنند. با همکاری با کارشناسان حوزه، داده‌کاوی می‌تواند مشکلات را بهتر تعریف کند، ویژگی‌های مرتبط را انتخاب کند و نتایج را تأیید کند. علاوه بر این، دانش حوزه‌ای می‌تواند در انتخاب مراحل پیش‌پردازش راهگشا بوده و اطمینان حاصل کند که فرآیند داده‌کاوی با اهداف خاص یک حوزه معین هماهنگ است.

---

۱ Interdisciplinary Approaches

۲ neuroscience

۳ biology

۴ neuroscience

۵ epidemiology

۶ genomics

۷ genetic markers

۸ Leveraging Domain Knowledge

ماهیت میان‌رشته‌ای داده‌کاوی نه تنها کاربردهای آن را گسترش می‌دهد، بلکه با استخراج دانش از حوزه‌های مختلف و شناسایی چالش‌های منحصربه‌فرد آنها، دانش آن زمینه را نیز غنی می‌کند. همکاری بین متخصصان داده‌کاوی و کارشناسان سایر زمینه‌ها، کیفیت نتایج داده‌کاوی را افزایش می‌دهد و راه‌حل‌ها و اکتشافات نوآورانه را تقویت می‌کند.

۱-۸-۴- داده‌کاوی قابل‌توضیح و قابل‌اعتماد[۱]

منظور از این بخش اهمیت قابل‌توضیح و قابل‌اعتماد بودن در مدل‌های داده‌کاوی است. توانایی درک و اعتماد به تصمیمات اتخاذشده توسط مدل‌های داده‌کاوی به دلیل استفاده از آنها در طیف گسترده‌ای از برنامه‌ها بسیار مهم است. در این راستا لازم است موارد زیر توجه شوند:

- اهمیت توضیح‌پذیری و تفسیرپذیری در مدل‌های داده‌کاوی: مدل‌های داده‌کاوی می‌توانند بسیار پیچیده باشند و درک اینکه چرا یک تصمیم یا پیش‌بینی خاص انجام شده است را برای کاربران چالش‌برانگیز می‌کند. این عدم شفافیت می‌تواند مانعی برای پذیرش داده‌کاوی در حوزه‌های مختلف، به‌ویژه مواردی که مقررات سخت‌گیرانه دارند یا با جان و ایمنی انسان‌ها در ارتباط است، باشد. بنابراین، اولویت دادن به توسعه مدل‌هایی که قابل توضیح و تفسیر هستند بسیار مهم است. این بدان معناست که کاربران می‌توانند بینش‌هایی به دست آورند که چگونه مدل به یک نتیجه‌گیری خاص رسیده است، چه ویژگی‌هایی تأثیرگذار بوده‌اند و اینکه آیا پیش‌بینی‌های مدل با دانش حوزه انسانی همخوانی دارد یا خیر.
- روش‌هایی برای شفاف‌تر کردن و قابل اعتماد کردن مدل‌های داده‌کاوی: دستیابی به شفافیت و قابل‌اعتماد بودن در مدل‌های داده‌کاوی، مستلزم به‌کارگیری روش‌های مختلف است. برخی از این روش‌ها شامل ساده‌سازی معماری‌های مدل[۲]، کاهش ابعاد[۳]، ارائه امتیاز برای اهمیت یک مشخصه[۴] و تولید توضیحات قابل درک برای تصمیم‌گیری‌ها است. تجسم‌ها[۵]، مدل‌های مبتنی بر قانون، و نمایش‌های ساختاریافته از فرآیند تصمیم‌گیری نیز در افزایش شفافیت مؤثر هستند.

اطمینان از شفافیت و قابل‌اعتماد بودن مدل‌های داده‌کاوی نه تنها برای پذیرش کاربر، بلکه برای ملاحظات قانونی و اخلاقی نیز ضروری است. این مساله ذینفعان را قادر می‌سازد تا تصمیمات اتخاذشده توسط این مدل‌ها را درک و تأیید کنند و بنابراین، امکان استفاده مسئولانه و اخلاقی داده‌کاوی در حوزه‌های مختلف را فراهم می‌کند.

۱-۹ جمع‌بندی و نتیجه‌گیری

هدف از این فصل کسب دانش کافی در رابطه با جایگاه و کاربردهای تحلیل داده، معرفی جامع روال انجام یک داده‌کاوی موفق، آشنایی با برخی روش‌های انجام داده‌کاوی، بیان برخی چالش‌های موجود در رابطه با حفظ حریم خصوصی و امنیت، و نیز بررسی اجمالی چالش‌ها و جهت‌گیری‌های آینده داده‌کاوی بوده است. بنابراین، در این فصل، ابتدا مقدمه‌ای از مبحث تحلیل داده بیان شد و در ادامه به اهمیت و جایگاه حوزه داده‌کاوی در دنیای امروز

[۱] Explainable and Trustworthy Data Mining

[۲] simplifying model architectures

[۳] reducing dimensionality

[۴] providing clear feature importance scores

[۵] Visualizations

اشاره شد. سپس، مروری جامع بر مفاهیم حوزه تحلیل داده انجام و روال انجام کار نیز بیان شد. همچنین مروری نیز بر انواع روش‌های داده‌کاوی صورت پذیرفت و برخی الگوریتم‌ها و ابزارهای موجود جهت انجام داده‌کاوی نیز معرفی شد. در ادامه موضوع حفظ حریم خصوصی و برخی راهکارهای موجود در این حوزه نیز معرفی شدند. سپس، چالش‌ها و جهت‌گیری‌های آینده داده‌کاوی نیز مورد بررسی قرار گرفت. نتایج حاصل از این مطالعه نشان می‌دهد که می‌توان از پروسه‌های استانداردی جهت پیش‌برد اهداف تحلیل داده استفاده کرد و در این راستا ابزار و الگوریتم‌های متنوعی نیز توسعه داده شده است و همچنان می‌توان در این راستا مطالعات بیشتری انجام داد تا بتوان روش‌های بهتری ارائه داد. همچنین، برخی نگرانی‌ها و چالش‌ها از قبیل مسائل مربوط به حفظ حریم خصوصی و ملاحظات اخلاقی، یا نحوه برخورد با کلان‌داده و توسعه روش‌های مناسب جهت تحلیل بلادرنگ آنها نیز وجود دارد. اگرچه جهت حل این قبیل چالش‌ها تاکنون راهکارهای گوناگونی در منابع مختلف ارائه شده است؛ ولی با توجه به گسترش روزافزون نوع و حجم داده‌های در دسترس و حساس بودن بسیاری از این داده‌ها، همچنان نیاز به توجه و مطالعه بیشتر در این زمینه وجود دارد.

۲

بررسی کاربرد تحلیل داده در مدیریت انرژی صنایع

۲-۱ مقدمه

هدف از این فصل بررسی بار مربوط به برخی از صنایع کشور و انجام چند نمونه تحلیل اولیه بر روی بار این مشترکین با هدف اصلاح تعرفه یا الگوی مصرف است.

۲-۲ تحلیل داده‌ها

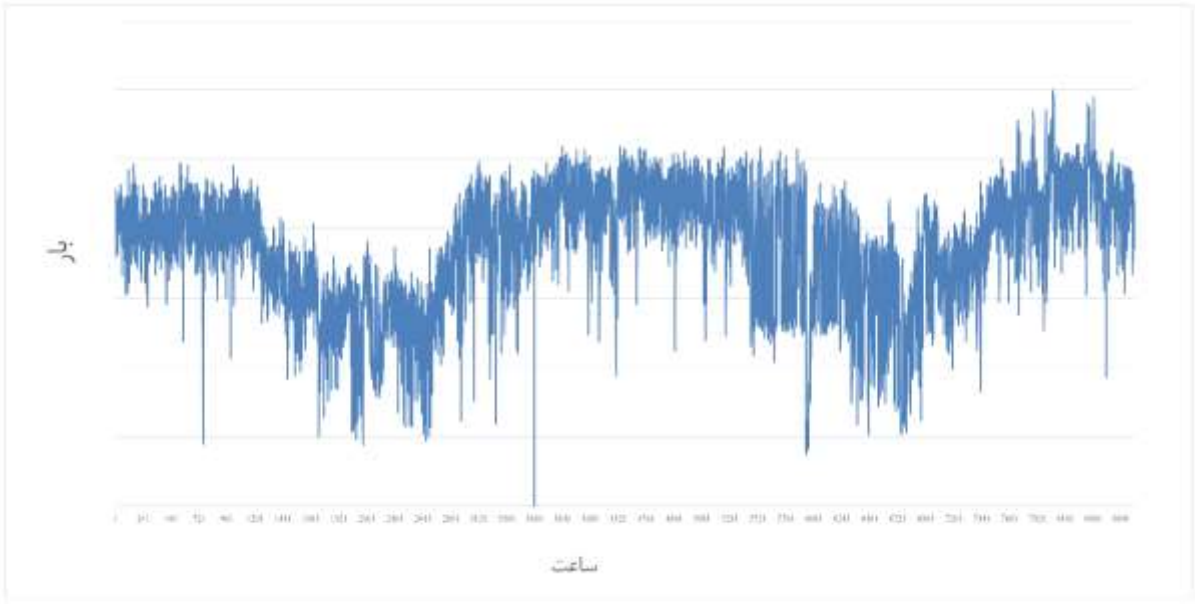
در این بخش از اطلاعات بار ساعتی ۷۳ نمونه از صنایع بزرگ کشور طی یک سال از اول آبان سال ۱۴۰۰ تا ۳۰ مهر ۱۴۰۱ استفاده شده است. در این بخش برخی از تحلیل‌های انجام شده معرفی خواهند شد.

لازم به ذکر است جهت انجام هر کدام از تحلیل‌ها لازم بوده داده‌های دسترس در قالب مناسب و قابل استفاده تهیه شوند و سپس تحلیل‌های لازم روی آن‌ها صورت گیرد. در این پروژه با توجه به حجم زیاد داده‌ها، از نرم‌افزار پایتون جهت تهیه داده‌ها در قالب مناسب و سپس انجام تحلیل‌های لازم استفاده شده است.

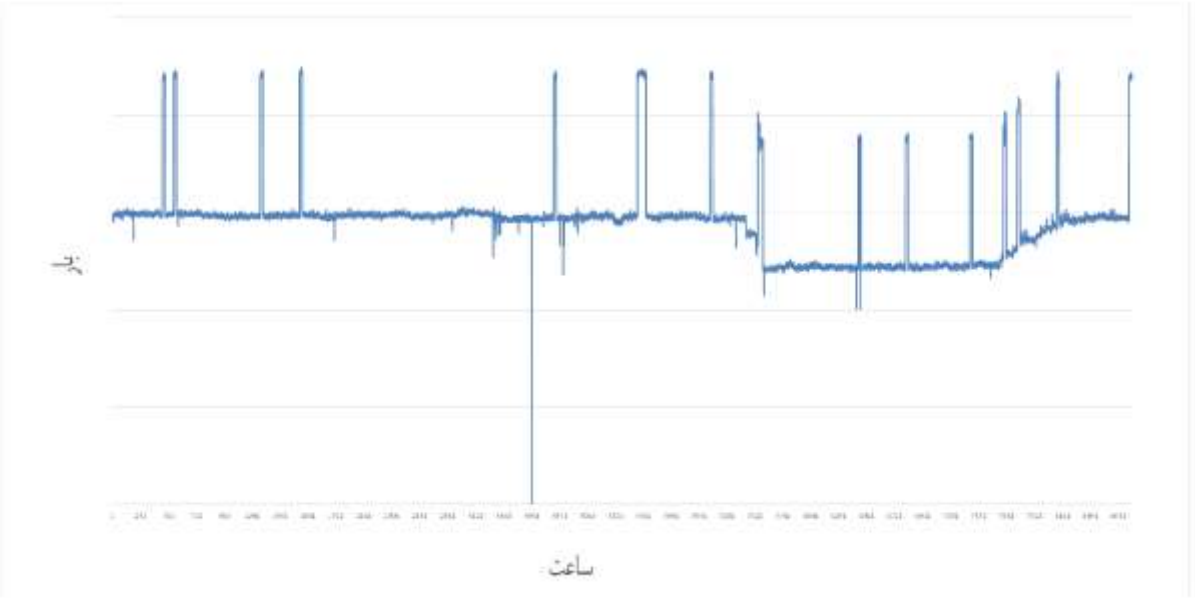
در ادامه هر کدام از تحلیل‌ها با جزئیات بیشتری بیان شده‌اند و برخی از خروجی‌ها نیز نشان داده شده‌اند. لازم به ذکر است کل دوره تحت بررسی در تمامی موارد از اول آبان سال ۱۴۰۰ تا سی مهر سال ۱۴۰۱ است. ذکر این نکته ضروری است که برای بررسی دقیق رفتار مشترکین صنعتی نیاز است هر گونه عاملی که ممکن است در بازه‌هایی از زمان بر روی مقدار بار واقعی شرکت اثرگذار هستند شناسایی شده و در برآورد بار واقعی شرکت لحاظ شوند. به عنوان مثال، لازم است مقدار مشارکت صنایع در برنامه‌های مدیریت بار در تابستان سال ۱۴۰۱ در دسترس بوده و برآورد صحیحی از رفتار واقعی هر شرکت داشت تا بتوان مقدار بار دستکاری نشده هر شرکت را محاسبه کرد که البته در حال حاضر این امکان وجود نداشته است.

همچنین، باید توجه شود در تمامی مواردی که در ادامه بیان خواهد شد، جهت حفظ محرمانگی، مواردی از قبیل نام دقیق و مقدار بار صنایع ذکر نخواهد شد. در ادامه برخی از خروجی‌ها و تحلیل‌های حاصل ارائه می‌شود.

- ✓ نمودار سالانه بار شرکت‌ها جهت برآورد یک دید کلی از الگوی مصرف آنها: در این بخش نمودار سالانه بار چند نمونه از شرکت‌ها نشان داده شده است (شکل‌های ۲-۱ و ۲-۲).



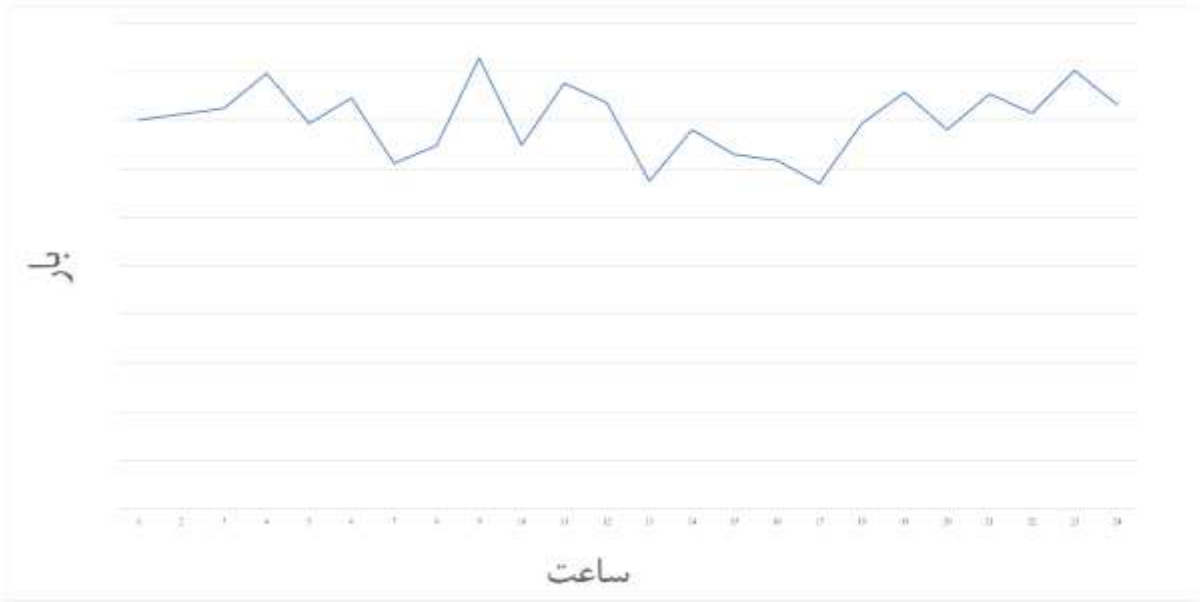
شکل ۲-۱: نمودار سالانه بار شرکت ۱



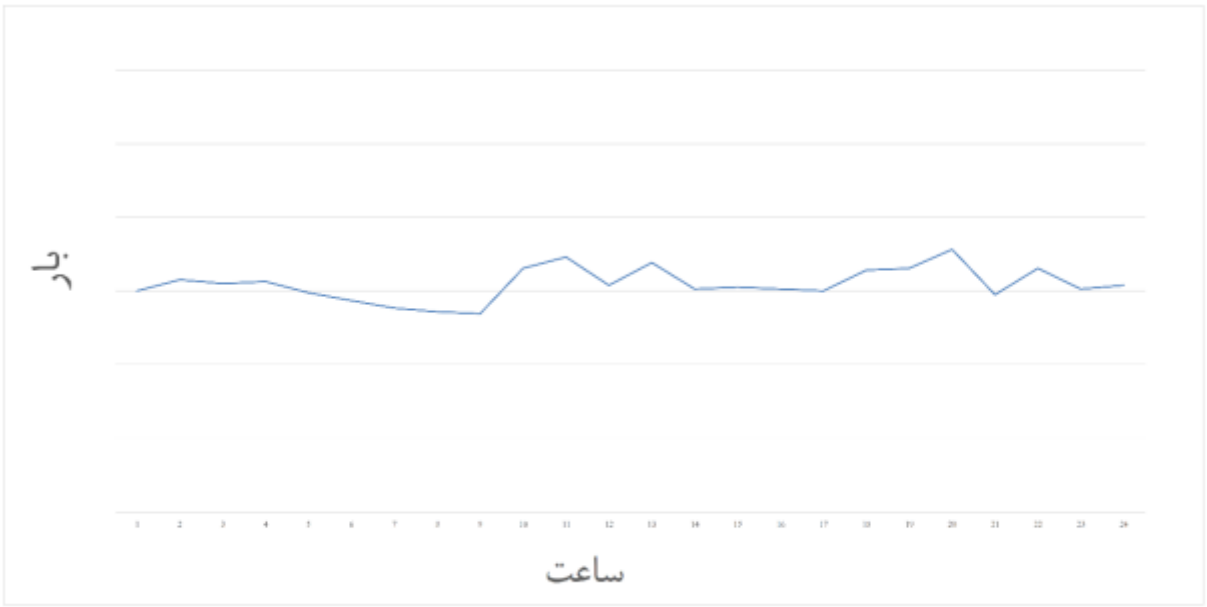
شکل ۲-۲: نمودار سالانه بار شرکت ۲

همانطور که در شکل‌های ۲-۱ و ۲-۲ مشاهده می‌شود، شرکت ۱ رفتاری کاملا متغیر طی ساعات مختلف در طول سال دارد؛ در حالی که شرکت ۲ باری تقریبا ثابت را طی این بازه سالانه نشان داده است.

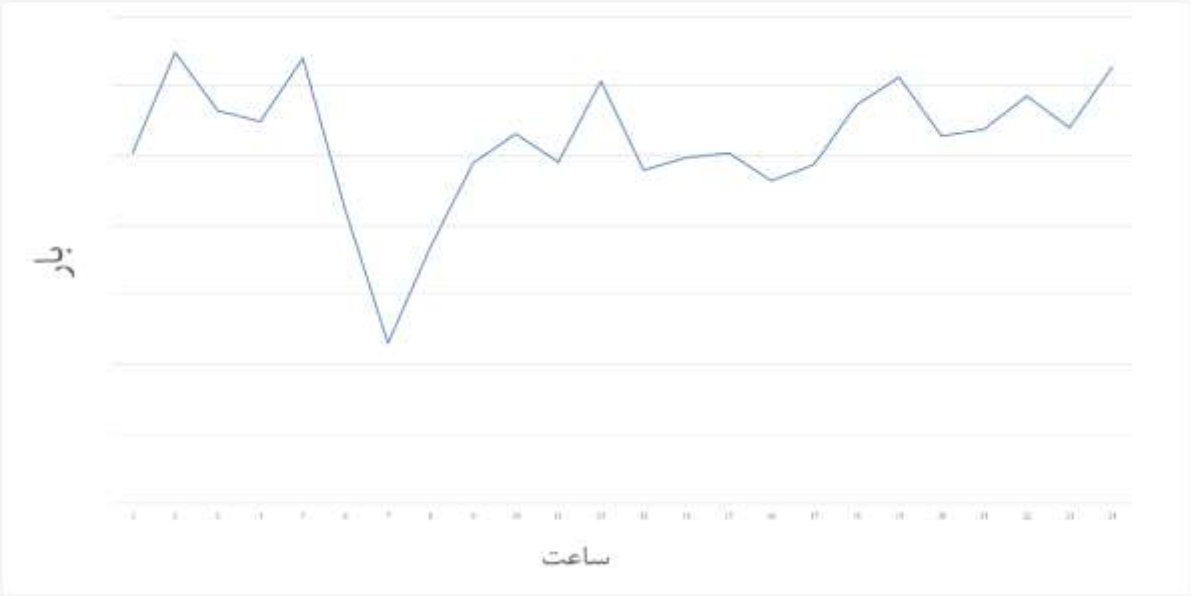
- ✓ نمودار مصرف روزانه برای چند روز خاص جهت بررسی الگو و تغییرات روزانه مصرف: در این بخش نمودار مصرف روزانه برای چند روز خاص نشان داده شده است (شکل‌های ۲-۳ تا ۲-۸).



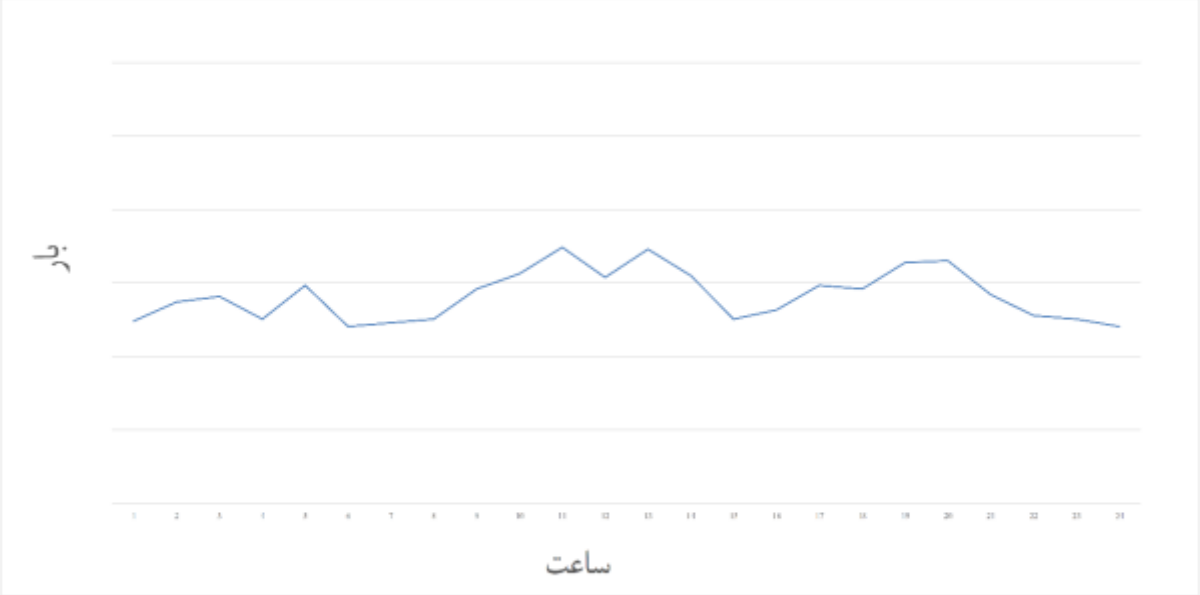
شکل ۲-۳: نمودار مصرف شرکت ۱ در روز یکشنبه، ۳۰ آبان ۱۴۰۰



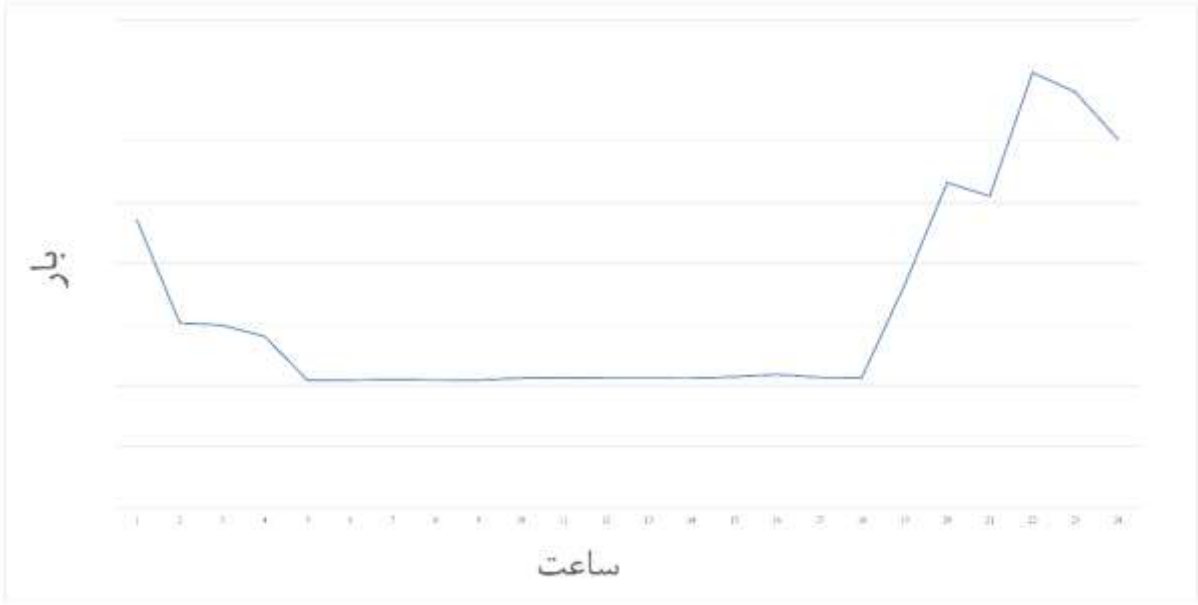
شکل ۲-۴: نمودار مصرف شرکت ۲ در روز یکشنبه، ۳۰ آبان ۱۴۰۰



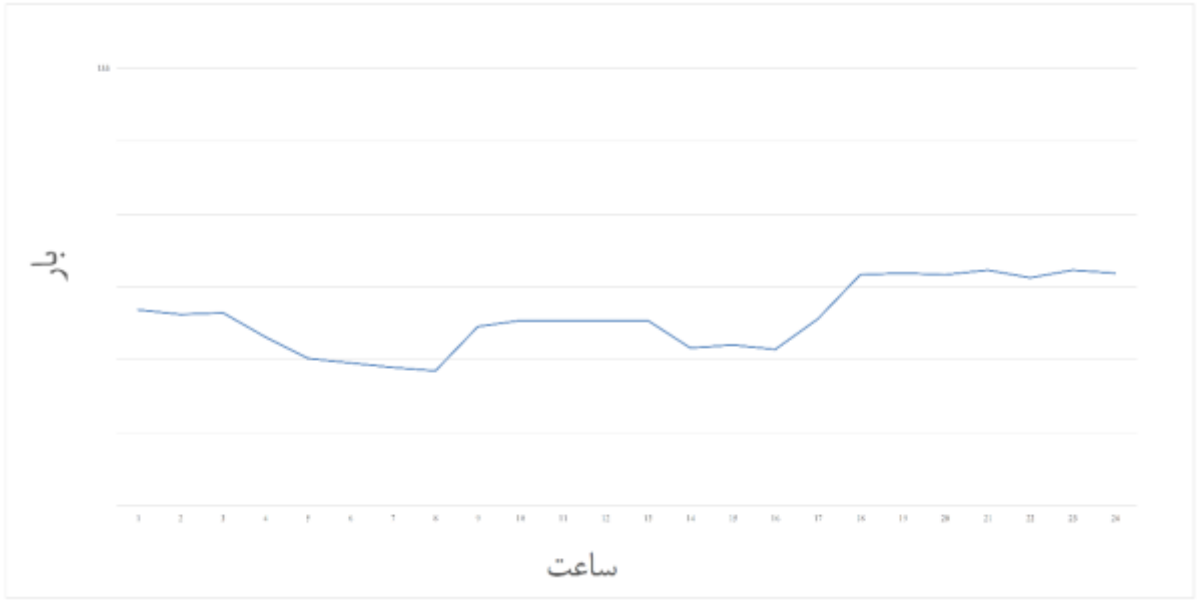
شکل ۲-۵: نمودار مصرف شرکت ۱ در روز یکشنبه، ۱۷ بهمن ۱۴۰۰



شکل ۲-۶: نمودار مصرف شرکت ۲ در روز یکشنبه، ۱۷ بهمن ۱۴۰۰



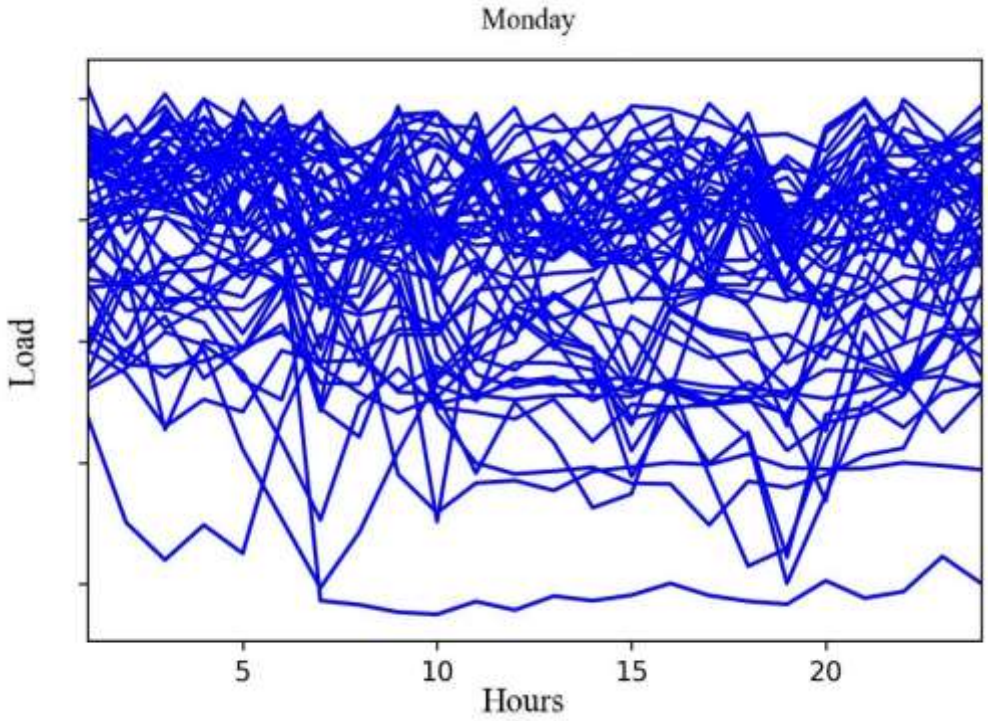
شکل ۲-۷: نمودار مصرف شرکت ۱ در روز یکشنبه، ۹ مرداد ۱۴۰۱



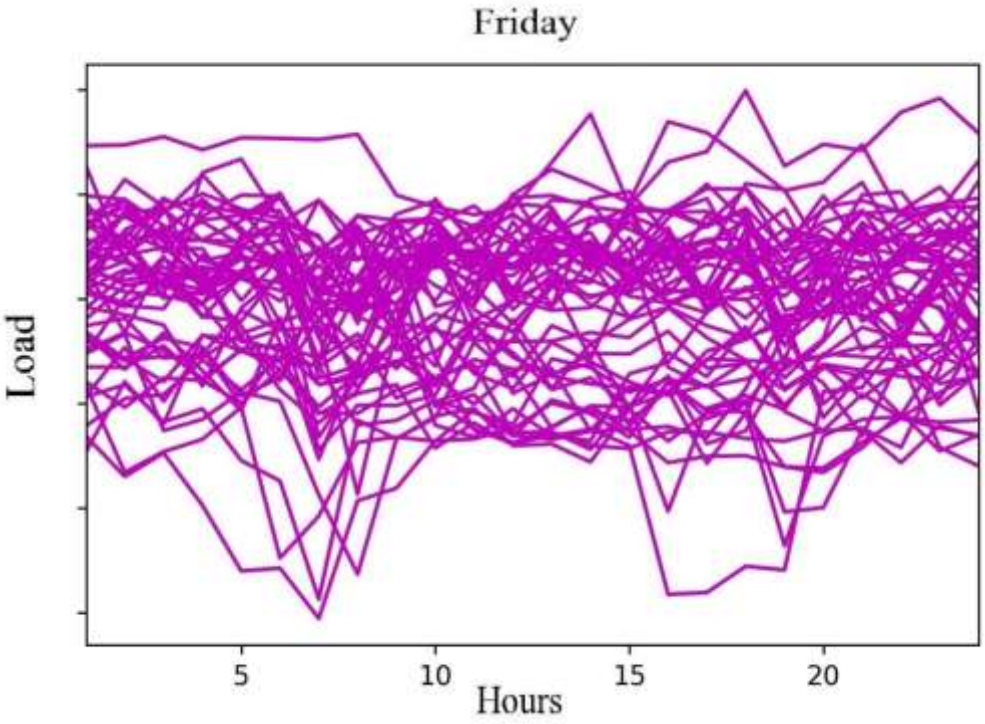
شکل ۲-۸: نمودار مصرف شرکت ۲ در روز یکشنبه، ۹ مرداد ۱۴۰۱

با بررسی نمودارهای (۲-۳) تا (۲-۸)، به نظر می‌رسد در ساعاتی در تابستان برنامه‌ای جهت کاهش بار اعمال شده است که رفتار بار صنایع به طور معناداری تغییر کرده است.

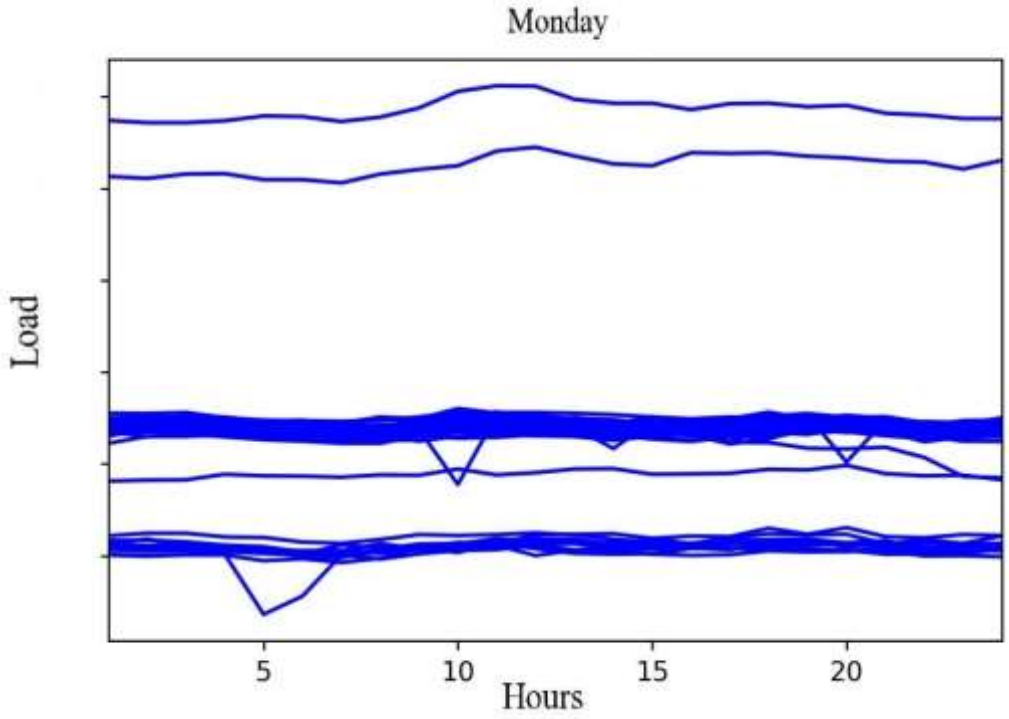
- ✓ رسم نمودار بار مصرفی به تفکیک برای هر روز هفته جهت بررسی شکل مصرف در روزهای مختلف و مقایسه آنها: در این بخش سعی شده تا نمودار بار هر روز خاص در هفته را کنار یکدیگر قرار داد تا بررسی شود آیا می‌توان تحلیل خاصی را روی هر نمودار بدست آورد یا خیر. شکل‌های (۲-۱۵) تا (۲-۱۸) نمودار مربوط به چند روز هفته را برای شرکت‌های ۱ و ۲ نشان می‌دهد.



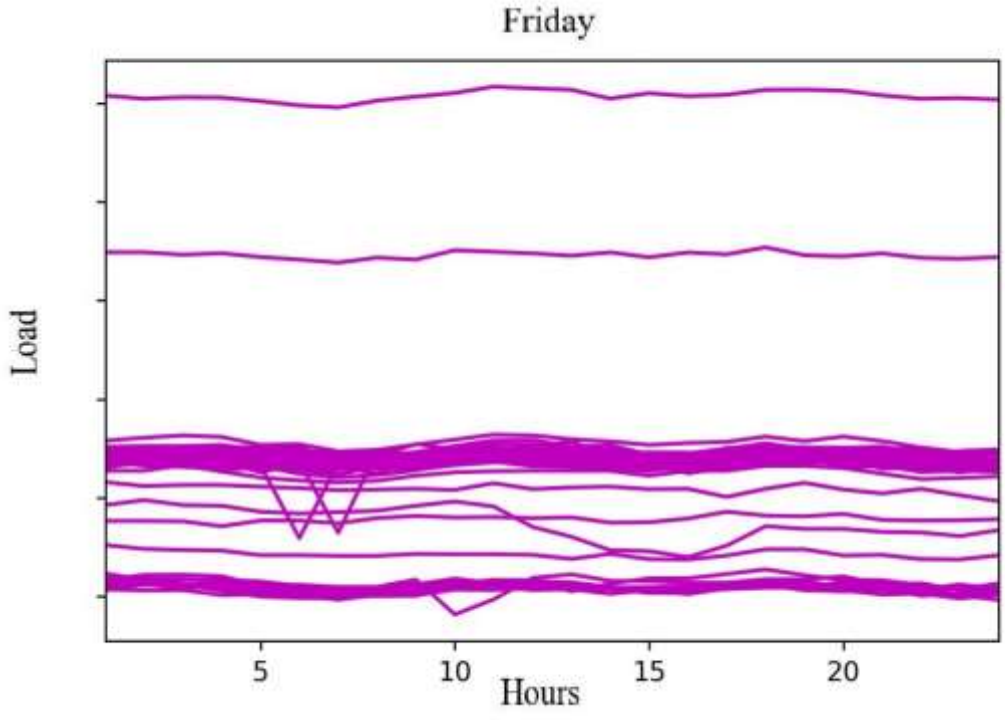
شکل ۲-۱۵: مقدار بار ساعتی روز دوشنبه برای شرکت ۱



شکل ۲-۱۶: مقدار بار ساعتی روز جمعه برای شرکت ۱

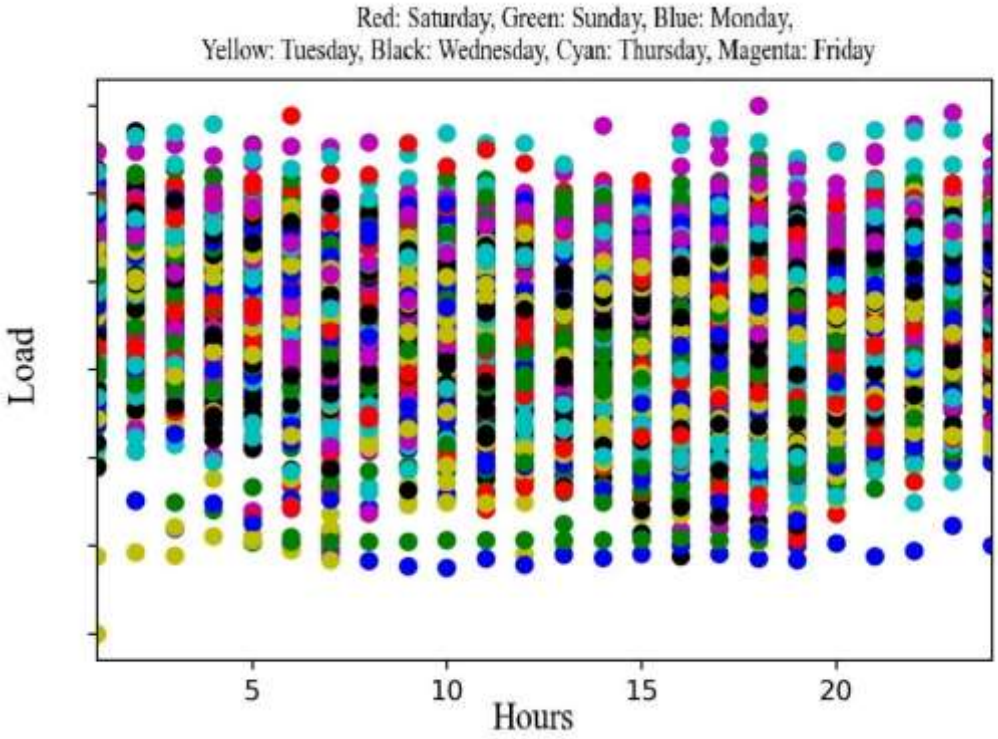


شکل ۲-۱۷: مقدار بار ساعتی روز دوشنبه برای شرکت ۲

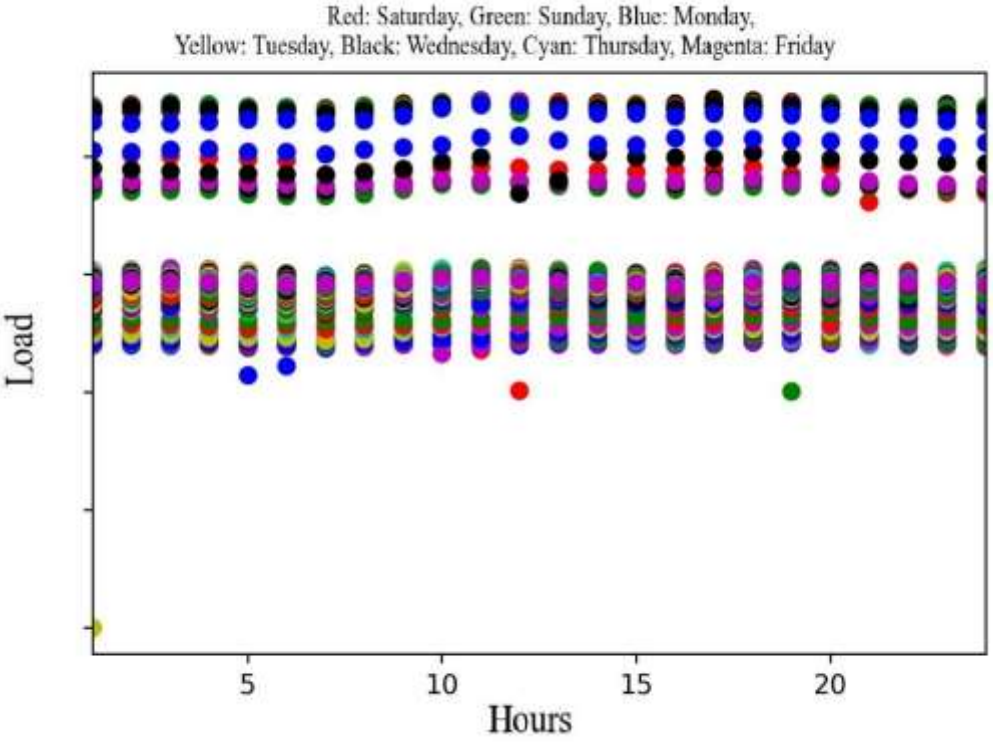


شکل ۲-۱۸: مقدار بار ساعتی روز جمعه برای شرکت ۲

همچنین می‌توان در صورت نیاز نمودار بار روزهای مختلف هفته را نیز روی هم قرار داد و انواع دیگری از نمودارها را رسم کرد (مانند نمودارهای scattered نشان داده شده در شکل‌های (۲-۱۹) و (۲-۲۰) برای شرکت‌های ۱ و ۲).

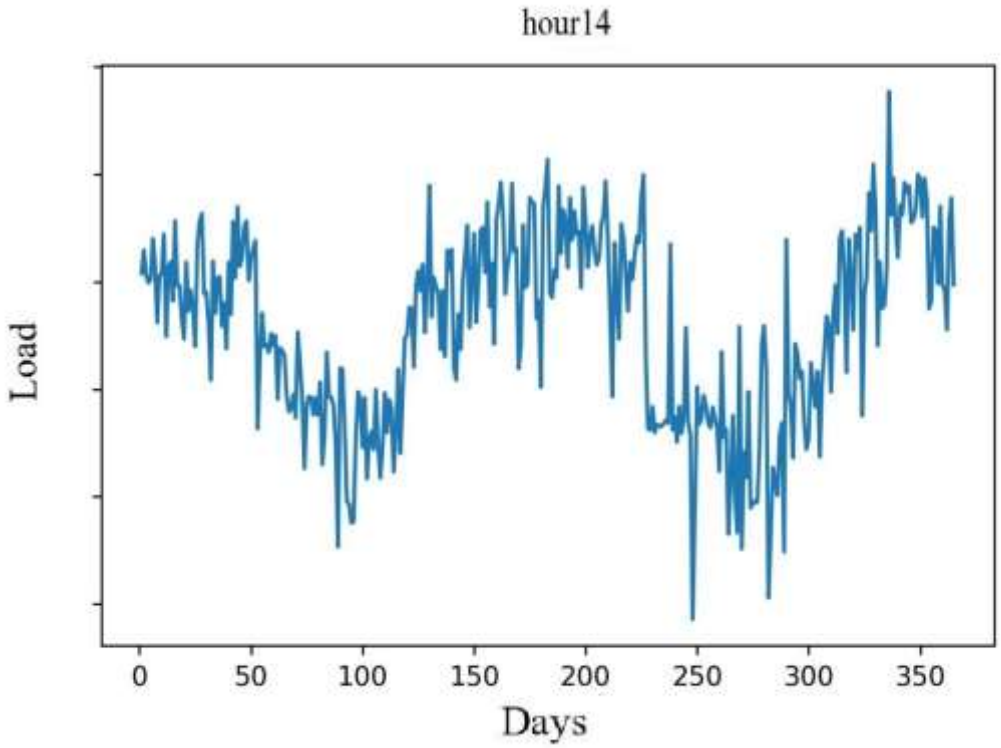


شکل ۲-۱۹: مقدار بار ساعتی روزهای مختلف هفته برای شرکت ۱

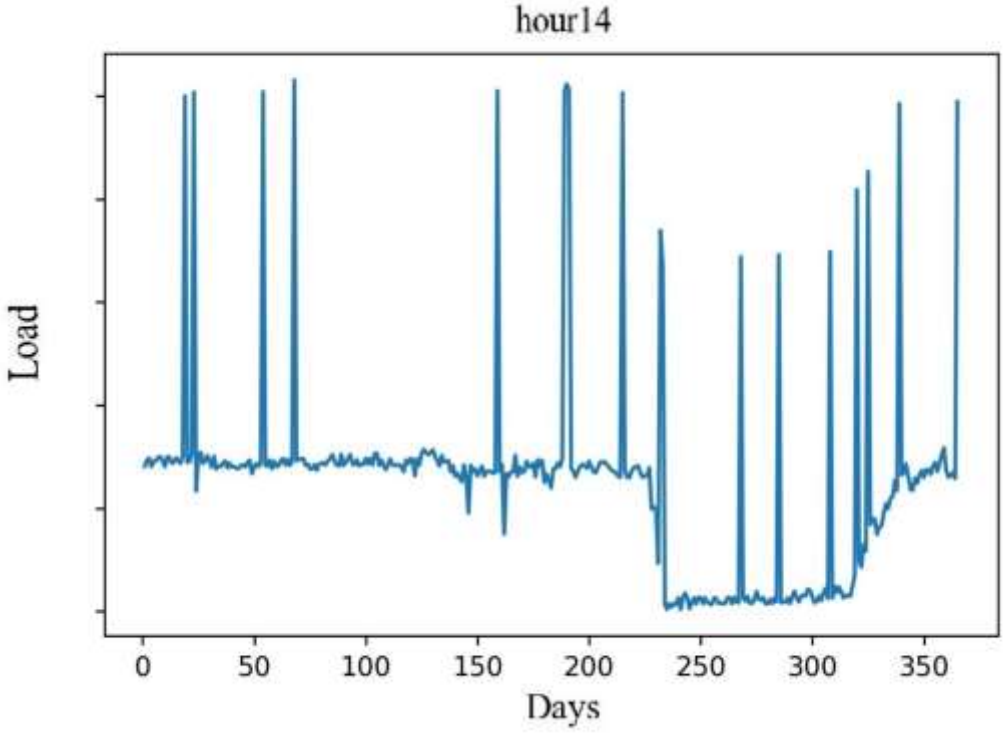


شکل ۲-۲۰: مقدار بار ساعتی روزهای مختلف هفته برای شرکت ۲

- ✓ رسم نمودار بار برای هر ساعت در طول سال جهت بررسی مقدار بار در هر ساعت و مقایسه آن با رفتار بار در ساعت‌های مختلف: در این بخش، جهت بررسی رفتار بار در هر ساعت و مقایسه آن با ساعات دیگر، نمودار بار به ازای هر ساعت به صورت مجزا رسم و بررسی می‌شود (مانند شکل‌های (۲-۱۹) و (۲-۲۰) برای شرکت‌های ۱ و ۲).



شکل ۲-۲۱: مقدار بار ساعت ۱۴ طی سال برای شرکت ۱



شکل ۲-۲۲: مقدار بار ساعت ۱۴ طی سال برای شرکت ۲

- ✓ محاسبه مقدار میانگین بار (میانگین کل بازه تحت بررسی و میانگین به تفکیک هر فصل) و مقایسه آن با مقادیر بیشینه بار در آن بازه، به تفکیک با استفاده از داده‌های سالانه و فصلی و دسته‌بندی شرکت‌ها بر اساس نتایج حاصل[1]: در این بخش، ابتدا مقدار میانگین و بیشینه بار برای دوره سالانه تحت‌بررسی محاسبه شد. سپس، همین محاسبات به تفکیک برای بازه آبان و آذر ۱۴۰۰، زمستان ۱۴۰۰، بهار ۱۴۰۱، تابستان ۱۴۰۱ و مهر ۱۴۰۲ نیز محاسبه شد. مقایسه مقدار میانگین بار در هر بازه با مقدار بیشینه بار آن بازه نشان‌دهنده مقدار فاصله بار مصرفی در هر بازه نسبت به میانگین بار آن شرکت است. در ادامه، از آن‌جا که مدیریت بار در بازه تابستان بر اساس احتیاجات شبکه برق کشور حائز اهمیت بیشتر است، مقایسه‌ای بین میانگین بار در در تابستان ۱۴۰۱ و میانگین بار در کل بازه سالانه مورد بررسی به استثنای تابستان ۱۴۰۱ صورت گرفت. بر این اساس، می‌توان یک دسته‌بندی روی شرکت‌ها انجام داد و شرکت‌هایی که میانگین بار تابستان آن‌ها بیشتر از میانگین بار در بازه زمانی غیر از تابستان است را در یک دسته قرار داد. به عنوان مثال، هیچ کدام از شرکت‌های ۱ و ۲ که در این گزارش بررسی می‌شوند در این دسته قرار نگرفته‌اند که این بدین معنا است که میانگین بار این شرکت‌ها در تابستان کمتر از میانگین بار آنها در سایر فصول است و از این دید چالشی ندارند. البته باید توجه داشت که بر اساس تحلیل‌های پیشین به نظر می‌رسد این شرکت‌ها در برنامه مدیریت بار در تابستان مشارکت داشته‌اند که این مساله می‌تواند داده‌ها را تحت تاثیر قرار دهد (موجب کاهش بار در بازه تابستان شود). برای بررسی دقیق‌تر نیاز است برآورد مقدار مشارکت در برنامه مدیریت بار در تابستان سال

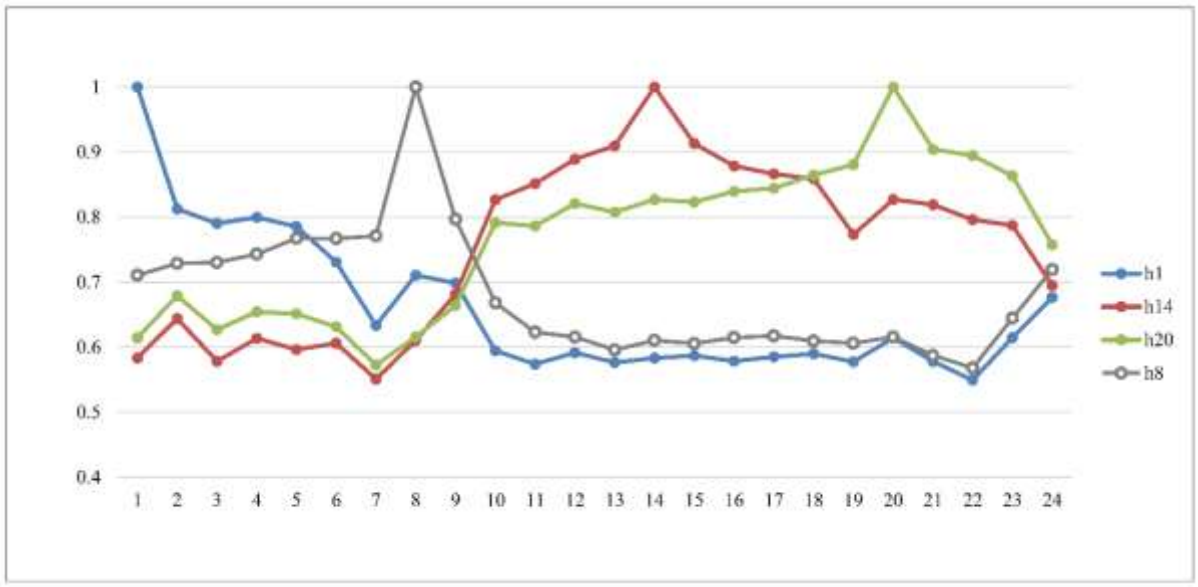
[1] به عنوان مثال، شرکت‌هایی که مقدار میانگی روزانه آنها در تابستان بیشتر از میانگین بار روزانه در سایر فصول است در یک دسته قرار گرفته‌اند.

۱۴۰۱ به تفکیک برای شرکت‌ها در دسترس باشد تا بتوان رفتار واقعی و دستکاری نشده الگوی بار هر شرکت را محاسبه کرد که البته در حال حاضر این امکان وجود نداشته است.

- ✓ استفاده از شاخص همبستگی[۱] جهت بررسی همبستگی ساعتی مقدار بار هر شرکت در طول شبانه‌روز (به تفکیک با استفاده از داده‌های سالانه و فصلی) و رسم نمودار همبستگی برای چند ساعت خاص جهت برآورد یک دید کلی از میزان همبستگی ساعتی: در این بخش ابتدا مقدار شاخص همبستگی به صورت ساعتی با استفاده از داده‌های کل سال محاسبه شد. نتیجه حاصل مقدار همبستگی بار در هر ساعت نسبت به ساعت‌های دیگر شبانه‌روز را نشان می‌دهد. در شکل‌های (۲-۹) و (۲-۱۰) نمودار همبستگی برای ساعات ۱، ۸، ۱۴ و ۲۰ نسبت به ساعات دیگر برای شرکت‌های ۱ و ۲ نشان داده شده است.

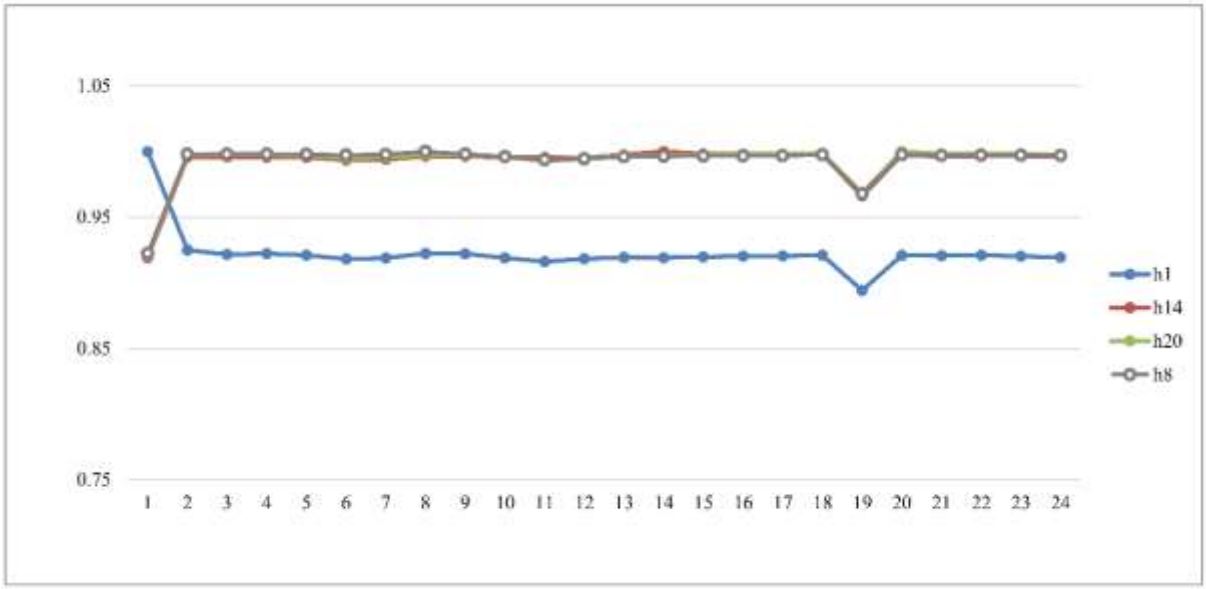
علاوه بر این، مقدار همبستگی ساعتی با استفاده از داده‌های فصلی نیز به صورت مجزا انجام شده است. برای نمونه، شکل‌های (۲-۱۱) و (۲-۱۲) مقدار این شاخص را برای فصل تابستان سال ۱۴۰۱ در شرکت‌های تحت بررسی نشان می‌دهد.

با بررسی نمودارهای فوق می‌توان دریافت که در شرکت ۱ احتمالا وابستگی خاصی بین مصرف در هر ساعت با ساعات دیگر وجود ندارد، در حالی که در بار شرکت ۲ وابستگی بالایی بین تمامی ساعات وجود دارد که نشان‌دهنده مصرف تقریبا یکنواخت این شرکت در طول ساعات مختلف است.

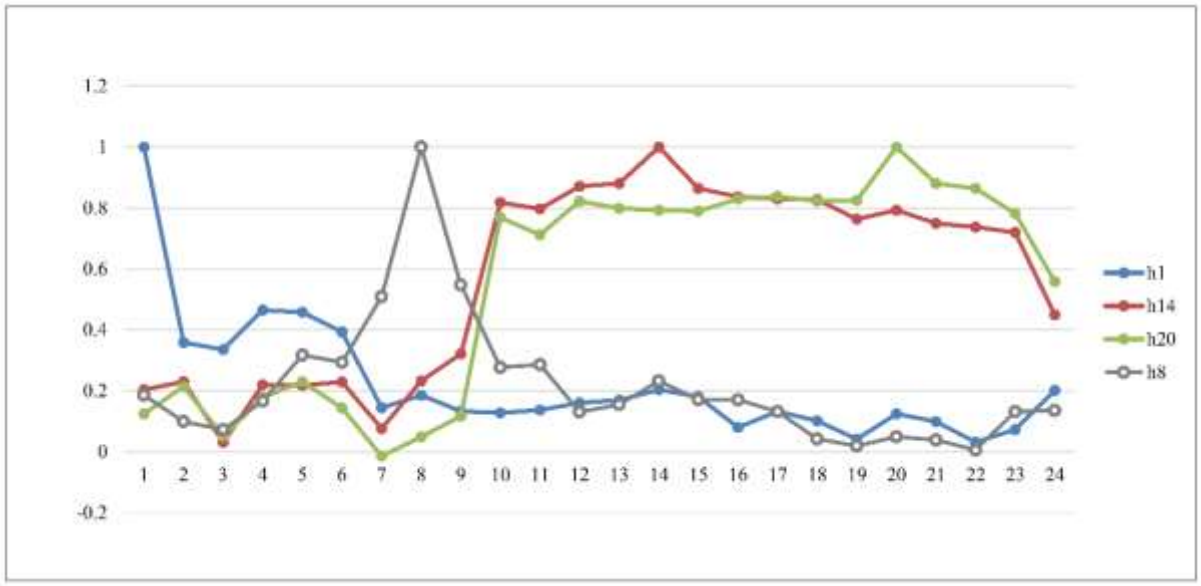


شکل ۲-۹: نمودار شاخص همبستگی ساعتی شرکت ۱ با استفاده از داده‌های کل سال (ساعات ۱، ۸، ۱۴، و ۲۰)

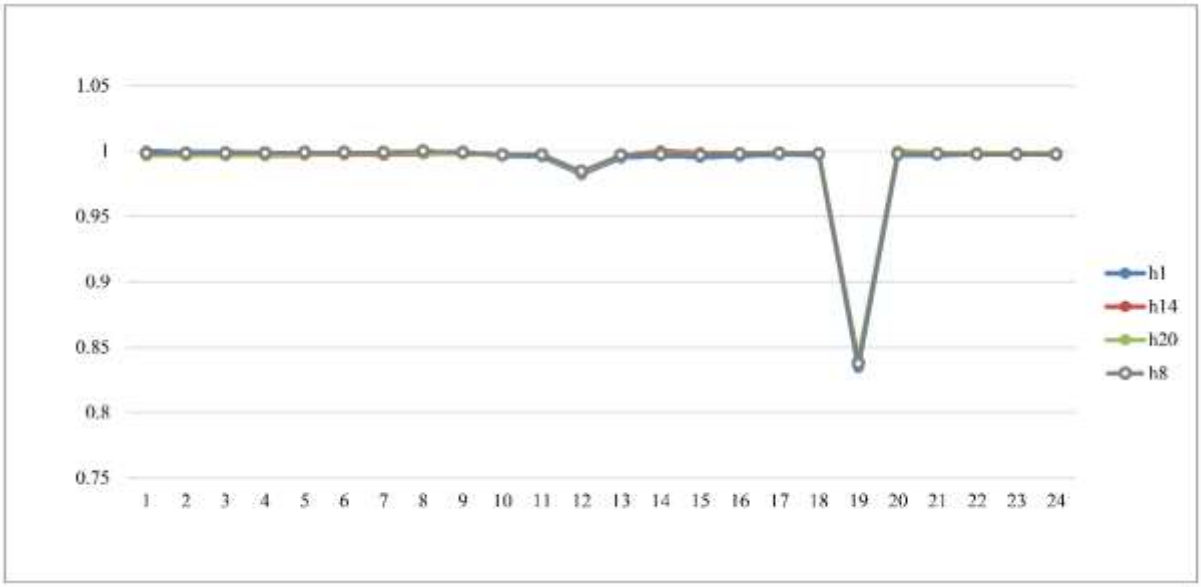
[۱] Correlation



شکل ۲-۱۰: نمودار شاخص همبستگی ساعتی شرکت ۲ با استفاده از داده‌های کل سال (ساعات ۱، ۸، ۱۴، و ۲۰)



شکل ۲-۱۱: نمودار شاخص همبستگی ساعتی شرکت ۱ با استفاده از داده‌های تابستان ۱۴۰۱ (ساعات ۱، ۸، ۱۴، و ۲۰)

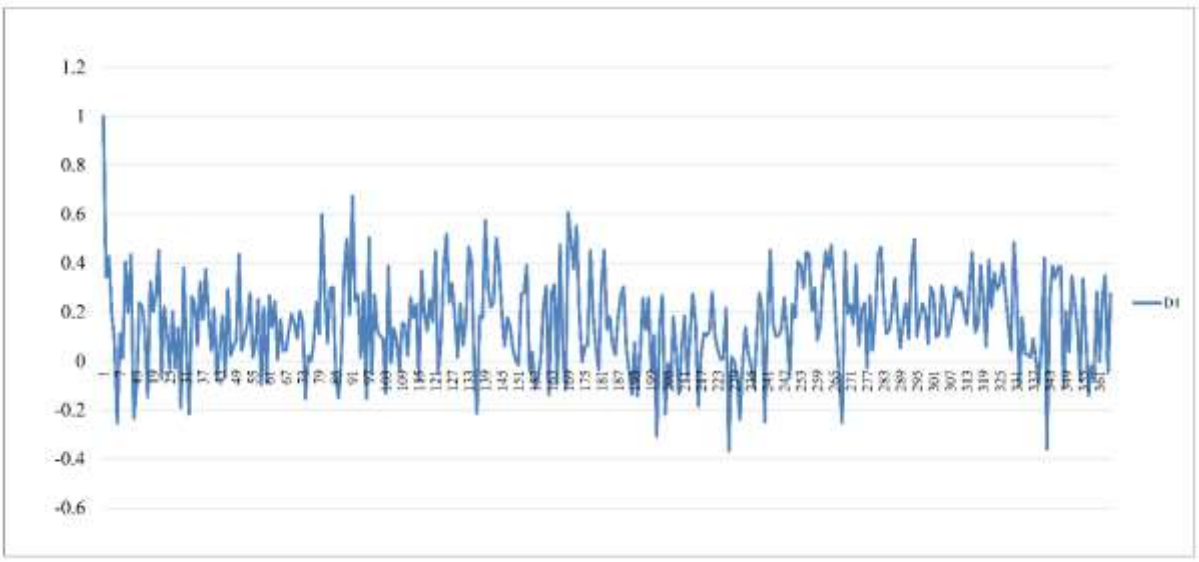


شکل ۲-۱۲: نمودار شاخص همبستگی ساعتی شرکت ۲ با استفاده از داده‌های تابستان ۱۴۰۱ (ساعات ۱، ۸، ۱۴، و ۲۰)

- ✓ استفاده از شاخص همبستگی جهت بررسی همبستگی روزانه مقدار بار هر شرکت با استفاده از داده‌های سالانه و رسم نمودار همبستگی برای یک روز جهت برآورد یک دید کلی از میزان همبستگی روزانه: در این بخش ابتدا مقدار همبستگی بین روزهای مختلف طی یک سال، یعنی برای ۳۶۵ روز، محاسبه شده است. شکل‌های (۲-۱۳) و (۲-۱۴) نمودار همبستگی روزانه را برای شرکت‌های ۱ و ۲ بر اساس داده‌های یک سال نشان می‌دهند.



شکل ۲-۱۳: نمودار شاخص همبستگی روزانه شرکت ۱ با استفاده از داده‌های ۳۶۵ روز سال (برای روز ۱ نسبت به سایر روزها)



شکل ۲-۱۴: نمودار شاخص همبستگی روزانه شرکت ۲ با استفاده از داده‌های ۳۶۵ روز سال (برای روز ۱ نسبت به سایر روزها)

با توجه به نمودارهای فوق هیچ‌گونه همبستگی بین مصرف روزهای مختلف سال نسبت به یکدیگر دیده نمی‌شود.

- ✓ محاسبه مقدار میانگین شاخص همبستگی ساعتی برای هر شرکت به تفکیک بر اساس داده‌های سالانه و فصلی و دسته‌بندی آنها بر اساس نتایج حاصل[1]: در این بخش، ابتدا میانگین مقدار همبستگی ساعتی که با استفاده از داده‌های سالانه و فصلی بدست آمد، محاسبه می‌شوند. با استفاده از این مقدار میانگین همبستگی ساعتی، چند نوع دسته‌بندی قابل انجام است. به عنوان مثال، اگر مقدار میانگین شاخص همبستگی ساعتی شرکتی (با استفاده از داده‌های سالانه) بیشتر از ۰,۹ باشد در دسته اول، بیشتر از ۰,۷ و کمتر یا مساوی ۰,۹ در دسته دوم و کمتر یا مساوی ۰,۷ در دسته سوم قرار می‌گیرند. بر این اساس، شرکت‌هایی که در دسته اول قرار می‌گیرند بیشترین مقدار همبستگی ساعتی را دارند و به همین ترتیب دسته سوم شرکت‌هایی را در بر دارد که کمترین مقدار همبستگی ساعتی را دارند. بر این اساس، مقدار میانگین شاخص همبستگی ساعتی برای شرکت ۱ برابر با ۰,۷۱ و برای شرکت ۲ برابر با ۰,۹۹ محاسبه شد؛ در نتیجه شرکت ۱ در دسته دوم و شرکت ۲ در دسته اول قرار می‌گیرند. همچنین می‌توان این دسته‌بندی را با استفاده از داده‌های ساعتی به تفکیک هر فصل انجام داد. به عنوان نمونه، با استفاده از داده‌های تابستان مقدار میانگین شاخص همبستگی برای شرکت ۱ برابر با ۰,۶۲ و برای شرکت ۲ برابر با ۰,۹۸ محاسبه شد. بر اساس این رویه، شرکت ۱ در دسته سوم و شرکت ۲ همچنان در دسته اول قرار خواهد گرفت.

جهت بررسی بیشتر، جدول (۲-۱) مقادیر همبستگی ساعتی را برای شرکت ۱ نشان می‌دهد.

[1] به عنوان مثال شرکت‌هایی که میانگین روزانه شاخص همبستگی آنها بیشتر از ۰,۹ باشد در یک دسته قرار می‌گیرند.

جدول ۲-۱: همبستگی ساعتی شرکت ۱

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
1	1.00	0.81	0.79	0.80	0.79	0.73	0.63	0.71	0.70	0.59	0.57	0.59	0.58	0.58	0.59	0.58	0.59	0.59	0.58	0.61	0.58	0.55	0.61	0.68
2	0.81	1.00	0.88	0.86	0.85	0.79	0.68	0.73	0.74	0.65	0.63	0.66	0.63	0.64	0.64	0.63	0.65	0.67	0.65	0.68	0.63	0.61	0.67	0.74
3	0.79	0.88	1.00	0.89	0.86	0.81	0.70	0.73	0.73	0.61	0.59	0.60	0.58	0.58	0.60	0.58	0.61	0.62	0.61	0.63	0.57	0.56	0.63	0.71
4	0.80	0.86	0.89	1.00	0.87	0.81	0.72	0.74	0.76	0.63	0.63	0.63	0.61	0.61	0.62	0.62	0.63	0.64	0.65	0.65	0.61	0.58	0.65	0.72
5	0.79	0.85	0.86	0.87	1.00	0.86	0.74	0.77	0.75	0.63	0.61	0.61	0.58	0.60	0.61	0.60	0.63	0.63	0.65	0.65	0.61	0.58	0.63	0.72
6	0.73	0.79	0.81	0.81	0.86	1.00	0.80	0.77	0.77	0.62	0.60	0.60	0.58	0.61	0.61	0.60	0.62	0.63	0.65	0.63	0.60	0.55	0.63	0.71
7	0.63	0.68	0.70	0.72	0.74	0.80	1.00	0.77	0.76	0.60	0.56	0.56	0.54	0.55	0.53	0.56	0.57	0.58	0.62	0.57	0.53	0.51	0.58	0.66
8	0.71	0.73	0.73	0.74	0.77	0.77	0.77	1.00	0.80	0.67	0.62	0.62	0.60	0.61	0.61	0.61	0.62	0.61	0.61	0.62	0.59	0.57	0.64	0.72
9	0.70	0.74	0.73	0.76	0.75	0.77	0.76	0.80	1.00	0.74	0.72	0.71	0.67	0.68	0.67	0.66	0.66	0.65	0.64	0.66	0.63	0.61	0.69	0.74
10	0.59	0.65	0.61	0.63	0.63	0.62	0.60	0.67	0.74	1.00	0.87	0.87	0.83	0.83	0.81	0.80	0.80	0.79	0.73	0.79	0.79	0.76	0.78	0.70
11	0.57	0.63	0.59	0.63	0.61	0.60	0.56	0.62	0.72	0.87	1.00	0.89	0.86	0.85	0.82	0.81	0.80	0.80	0.74	0.79	0.79	0.76	0.75	0.68
12	0.59	0.66	0.60	0.63	0.61	0.60	0.56	0.62	0.71	0.87	0.89	1.00	0.91	0.89	0.86	0.85	0.84	0.85	0.77	0.82	0.82	0.80	0.78	0.69
13	0.58	0.63	0.58	0.61	0.58	0.58	0.54	0.60	0.67	0.83	0.86	0.91	1.00	0.91	0.88	0.85	0.83	0.84	0.77	0.81	0.81	0.79	0.77	0.67
14	0.58	0.64	0.58	0.61	0.60	0.61	0.55	0.61	0.68	0.83	0.85	0.89	0.91	1.00	0.91	0.88	0.87	0.86	0.77	0.83	0.82	0.80	0.79	0.69
15	0.59	0.64	0.60	0.62	0.61	0.61	0.53	0.61	0.67	0.81	0.82	0.86	0.88	0.91	1.00	0.90	0.88	0.86	0.78	0.82	0.83	0.79	0.77	0.68
16	0.58	0.63	0.58	0.62	0.60	0.60	0.56	0.61	0.66	0.80	0.81	0.85	0.85	0.88	0.90	1.00	0.91	0.89	0.82	0.84	0.85	0.81	0.79	0.70
17	0.59	0.65	0.61	0.63	0.63	0.62	0.57	0.62	0.66	0.80	0.80	0.84	0.83	0.87	0.88	0.91	1.00	0.92	0.83	0.84	0.83	0.81	0.79	0.70
18	0.59	0.67	0.62	0.64	0.63	0.63	0.58	0.61	0.65	0.79	0.80	0.85	0.84	0.86	0.86	0.89	0.92	1.00	0.88	0.86	0.86	0.83	0.80	0.71
19	0.58	0.65	0.61	0.65	0.65	0.65	0.62	0.61	0.64	0.73	0.74	0.77	0.77	0.77	0.78	0.82	0.83	0.88	1.00	0.88	0.85	0.84	0.81	0.73
20	0.61	0.68	0.63	0.65	0.65	0.63	0.57	0.62	0.66	0.79	0.79	0.82	0.81	0.83	0.82	0.84	0.84	0.86	0.88	1.00	0.90	0.89	0.86	0.76
21	0.58	0.63	0.57	0.61	0.61	0.60	0.53	0.59	0.63	0.79	0.79	0.82	0.81	0.82	0.83	0.85	0.83	0.86	0.85	0.90	1.00	0.90	0.86	0.73
22	0.55	0.61	0.56	0.58	0.58	0.55	0.51	0.57	0.61	0.76	0.76	0.80	0.79	0.80	0.79	0.81	0.81	0.83	0.84	0.89	0.90	1.00	0.88	0.75
23	0.61	0.67	0.63	0.65	0.63	0.63	0.58	0.64	0.69	0.78	0.75	0.78	0.77	0.79	0.77	0.79	0.79	0.80	0.81	0.86	0.86	0.88	1.00	0.83
24	0.68	0.74	0.71	0.72	0.72	0.71	0.66	0.72	0.74	0.70	0.68	0.69	0.67	0.69	0.68	0.70	0.70	0.71	0.73	0.76	0.73	0.75	0.83	1.00

بر اساس داده‌های جدول (۲-۱) نتایج زیر قابل مشاهده است:

- ✓ تقارن[۱]: ماتریس همبستگی در امتداد قطر اصلی متقارن است؛ زیرا همبستگی بین ساعت X و ساعت Y با همبستگی بین ساعت Y و ساعت X یکسان است.
- ✓ همبستگی قطری[۲]: مقادیر در امتداد قطر اصلی (به عنوان مثال، ۱-۱، ۲-۲، ۳-۳، و غیره) همه برابر با ۱ هستند. این نشان می‌دهد همانطور که انتظار میرفت همبستگی یک ساعت با خودش کامل است (ضریب همبستگی ۱).
- ✓ همبستگی‌های ساعات متوالی: با دور شدن از قطر اصلی، مقادیر همبستگی کاهش می‌یابد. در واقع، ساعات دورتر از هم همبستگی ضعیف‌تری دارند. به عنوان مثال، ساعت ۱ با ساعت ۱۲ همبستگی کمتری نسبت به همبستگی آن با ساعت ۲ دارد. دقت شود ممکن در برخی موارد چنین چیزی دیده نشود؛ ولی در حالت کلی می‌توان این شکل تغییر را در داده‌ها مشاهده کرد. در واقع، بالاترین مقادیر همبستگی (نزدیک به ۱) در نزدیکی قطر اصلی[۳] (از بالا سمت چپ به سمت راست پایین) یافت می‌شود (برای مثال ساعت ۱ با ساعت ۲ همبستگی بالایی دارد و ساعت ۲۳ با ساعت ۲۴ همبستگی بالایی دارد). درکل، همبستگی بین ساعات متوالی (مانند ۱-۲، ۲-۳، ۳-۴، و غیره) به طور کلی بالا است، که نشان‌دهنده یک رابطه مثبت قوی بین این ساعت ها است.
- ✓ ساعات اوج مصرف: برخی ساعات ممکن است با چندین ساعت دیگر همبستگی بیشتری نشان دهند. اینها را می‌توان به عنوان ساعات اوج مصرف یا دوره‌های بحرانی در نظر گرفت که در آن فعالیت‌ها یا رویدادها به طور مداوم در روزهای مختلف مرتبط هستند.

جدول (۲-۲) مقدار میانگین ضرایب همبستگی به تفکیک هر ساعت برای شرکت ۱ را نشان می‌دهد. بر اساس جدول (۲-۲) ملاحظه می‌شود که بیشترین مقادیر میانگین همبستگی از حدود ساعت ۱۰ به بعد وجود دارد که ممکن است نشان‌دهنده ساعات اوج بار باشند.

جدول ۲-۲: میانگین ضرایب همبستگی به تفکیک هر ساعت برای شرکت ۱

۱	0.64	۷	0.62	۱۳	0.73	۱۹	0.73
۲	0.70	۸	0.66	۱۴	0.74	۲۰	0.75
۳	0.67	۹	0.70	۱۵	0.74	۲۱	0.73
۴	0.69	۱۰	0.73	۱۶	0.74	۲۲	0.71
۵	0.68	۱۱	0.72	۱۷	0.74	۲۳	0.73
۶	0.67	۱۲	0.74	۱۸	0.75	۲۴	0.71

- ✓ همبستگی کم: برخی از ساعات ممکن است همبستگی نسبتاً کمی با ساعات دیگر داشته باشند. این می‌تواند نشان‌دهنده دوره‌های فعالیت کم یا ثبات کمتر الگوها در آن ساعات باشد. به عنوان مثال، در جدول (۲-۲) میانگین همبستگی در ساعاتی از قبیل ساعت ۱ بامداد بسیار کم است.

---

[۱] Symmetry

[۲] Diagonal Correlation

[۳] main diagonal

علاوه بر این موضوع، برخی از همبستگی‌ها که مقادیر خیلی کوچکی دارند نشان می‌دهند که رابطه خطی کمی بین برخی ساعات خاص وجود ندارد. این می‌تواند به این معنی باشد که این ساعات از نظر تأثیر آنها بر صنعت نسبتاً مستقل از یکدیگر هستند.

جهت بررسی رفتار شرکت دوم، در جدول (۲-۳) مقادیر همبستگی ساعتی برای این شرکت نشان داده شده است.

جدول ۲-۳: همبستگی ساعتی شرکت ۲

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
1	1.00	0.92	0.92	0.92	0.92	0.92	0.92	0.92	0.92	0.92	0.92	0.92	0.92	0.92	0.92	0.92	0.92	0.92	0.89	0.92	0.92	0.92	0.92	0.92
2	0.92	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.99	0.99	1.00	1.00	1.00	1.00	1.00	1.00	0.97	1.00	1.00	1.00	1.00	1.00
3	0.92	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.99	0.99	1.00	1.00	1.00	1.00	1.00	1.00	0.97	1.00	1.00	1.00	1.00	1.00
4	0.92	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.99	0.99	1.00	1.00	1.00	1.00	1.00	1.00	0.97	1.00	1.00	1.00	1.00	1.00
5	0.92	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.99	0.99	1.00	1.00	1.00	1.00	1.00	1.00	0.97	1.00	1.00	1.00	1.00	1.00
6	0.92	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.97	0.99	0.99	0.99	0.99	0.99
7	0.92	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	1.00	0.97	1.00	1.00	1.00	1.00	1.00
8	0.92	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.99	0.99	1.00	1.00	1.00	1.00	1.00	1.00	0.97	1.00	1.00	1.00	1.00	1.00
9	0.92	1.00	1.00	1.00	1.00	0.99	1.00	1.00	1.00	1.00	0.99	0.99	1.00	1.00	1.00	1.00	1.00	1.00	0.97	1.00	1.00	1.00	1.00	1.00
10	0.92	1.00	1.00	1.00	1.00	0.99	0.99	1.00	1.00	1.00	0.99	0.99	0.99	1.00	1.00	1.00	1.00	1.00	0.96	1.00	1.00	1.00	1.00	1.00
11	0.92	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	1.00	0.99	0.99	1.00	1.00	0.99	1.00	0.99	0.96	0.99	0.99	0.99	0.99	0.99
12	0.92	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	1.00	1.00	0.99	0.99	0.99	0.99	0.99	0.96	0.99	0.99	0.99	0.99	0.99
13	0.92	1.00	1.00	1.00	1.00	0.99	0.99	1.00	1.00	0.99	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.97	1.00	1.00	1.00	1.00	1.00
14	0.92	1.00	1.00	1.00	1.00	0.99	0.99	1.00	1.00	1.00	1.00	0.99	1.00	1.00	1.00	1.00	1.00	1.00	0.97	1.00	1.00	1.00	1.00	1.00
15	0.92	1.00	1.00	1.00	1.00	0.99	0.99	1.00	1.00	1.00	1.00	0.99	1.00	1.00	1.00	1.00	1.00	1.00	0.97	1.00	1.00	1.00	1.00	1.00
16	0.92	1.00	1.00	1.00	1.00	0.99	0.99	1.00	1.00	1.00	0.99	0.99	1.00	1.00	1.00	1.00	1.00	1.00	0.97	1.00	1.00	1.00	1.00	1.00
17	0.92	1.00	1.00	1.00	1.00	0.99	0.99	1.00	1.00	1.00	1.00	0.99	1.00	1.00	1.00	1.00	1.00	1.00	0.97	1.00	1.00	1.00	1.00	1.00
18	0.92	1.00	1.00	1.00	1.00	0.99	1.00	1.00	1.00	1.00	0.99	0.99	1.00	1.00	1.00	1.00	1.00	1.00	0.97	1.00	1.00	1.00	1.00	1.00
19	0.89	0.97	0.97	0.97	0.97	0.97	0.97	0.97	0.97	0.96	0.96	0.96	0.97	0.97	0.97	0.97	0.97	0.97	1.00	0.97	0.97	0.97	0.97	0.97
20	0.92	1.00	1.00	1.00	1.00	0.99	1.00	1.00	1.00	1.00	0.99	0.99	1.00	1.00	1.00	1.00	1.00	1.00	0.97	1.00	1.00	1.00	1.00	1.00
21	0.92	1.00	1.00	1.00	1.00	0.99	1.00	1.00	1.00	1.00	0.99	0.99	1.00	1.00	1.00	1.00	1.00	1.00	0.97	1.00	1.00	1.00	1.00	1.00
22	0.92	1.00	1.00	1.00	1.00	0.99	1.00	1.00	1.00	1.00	0.99	0.99	1.00	1.00	1.00	1.00	1.00	1.00	0.97	1.00	1.00	1.00	1.00	1.00
23	0.92	1.00	1.00	1.00	1.00	0.99	1.00	1.00	1.00	1.00	0.99	0.99	1.00	1.00	1.00	1.00	1.00	1.00	0.97	1.00	1.00	1.00	1.00	1.00
24	0.92	1.00	1.00	1.00	1.00	0.99	1.00	1.00	1.00	1.00	0.99	0.99	1.00	1.00	1.00	1.00	1.00	1.00	0.97	1.00	1.00	1.00	1.00	1.00

بر اساس جدول (۲-۳)، نتایج زیر قابل حصول است:

- ✓ تقارن: مشابه شرکت ۱، در این حالت نیز ماتریس همبستگی در امتداد قطر اصلی متقارن است.
- ✓ همبستگی قطری: مشابه شرکت ۱، در این حالت نیز مقادیر قطری (همبستگی هر ساعت با خودش) همگی ۱ هستند، زیرا هر ساعت کاملاً با خودش همبستگی دارد.
- ✓ همبستگی‌های ساعات متوالی: مشابه شرکت ۱، در این حالت نیز غالبا همبستگی بین ساعات نزدیک به هم بیشتر است. البته برای شرکت دوم، همبستگی بین ساعات نزدیک به هم (مثلاً ۱ و ۲، ۲ و ۳ و غیره) بسیار زیاد است (معمولاً بالای ۰,۹). این نشان‌دهنده یک رابطه مثبت قوی بین ساعات متوالی است. با افزایش زمان تاخیر (به عنوان مثال، ۱ و ۱۰، ۲ و ۱۲، و غیره)، مقادیر همبستگی شروع به کاهش می‌کنند؛ اما همچنان نسبتاً بالا باقی می‌مانند. دقت شود ممکن در برخی حالات این اتفاق نیافتد و با افزایش زمان تاخیر کاهش همبستگی دیده نشود، ولی به طور کلی می‌توان در غالب نمونه‌ها این مساله را مشاهده کرد.
- ✓ همبستگی‌های تاخیری[۱]: همبستگی بین ساعات غیر متوالی (به عنوان مثال، ۱ و ۳، ۲ و ۴، و غیره) نیز بسیار بالا است، که نشان می‌دهد برخی تداوم یا شباهت در عملکرد صنعت بین این ساعت‌ها وجود دارد.
- ✓ همبستگی مثبت قوی[۲]: مقادیر همبستگی بین ساعات روز به طور کلی بسیار بالا است. این نشان‌دهنده یک همبستگی مثبت قوی است؛ به این معنی که با افزایش مقدار بار در یک ساعت، مقدار بار ساعت دیگر نیز افزایش می‌یابد. این نشان می‌دهد که برخی الگوها یا روندهای اساسی وجود دارد که در طول روز ثابت است.
- ✓ فصلی بودن بالقوه[۳]: بسته به زمینه و صنعت موردنظر، این مقادیر همبستگی بالا می‌تواند نشان‌دهنده الگوهای فصلی یا رویدادهای تکراری باشد که بر عملکرد صنعت در ساعات خاصی از روز تأثیر می‌گذارد.
- ✓ پتانسیل برای تجزیه‌وتحلیل سری‌های زمانی: همبستگی بالا بین ساعت های متوالی (به عنوان مثال، ساعت ۱ با ساعت ۲، ساعت ۲ با ساعت ۳، و غیره) نشان می‌دهد که روش‌های تجزیه‌وتحلیل سری‌های زمانی می‌تواند برای مدل‌سازی و پیش‌بینی رفتار این صنعت مفید باشد.

در حالت کلی باید موارد زیر را در نظر داشت:

- ✓ روندهای بالقوه[۴]: بسته به صنعت و ماهیت داده‌ها، ممکن است روندها یا الگوهای بالقوه‌ای وجود داشته باشد که ارزش بررسی بیشتر را دارند. برای مثال، افزایش یا کاهش مداوم مقادیر همبستگی در طول روز ممکن است روندهای زمانی خاصی را نشان دهد.

---

[۱] Lagged Correlations

[۲] Strong Positive Correlation

[۳] Potential Seasonality

[۴] Potential Trends

- ✓ تغییرات فصلی: اگر داده‌ها چندین روز یا فصول را در برگیرند، ممکن است بتوان تغییرات فصلی در همبستگی‌ها را شناسایی کرد؛ که نشان می‌دهد چگونه ساعات خاصی در زمان‌های خاصی از سال به هم مرتبط هستند.
- ✓ استفاده از دانش موجود: به طور کلی، این مقادیر همبستگی را می‌توان برای به دست آوردن بینشی در مورد الگوهای روزانه صنعت، شناسایی ساعات اوج مصرف و به طور بالقوه برنامه‌ریزی عملیاتی یا تصمیم‌گیری بر اساس همبستگی‌های مشاهده شده استفاده کرد. با این حال، برای تجزیه‌وتحلیل بیشتر داده‌ها، در نظر گرفتن دانش آن حوزه و انجام آزمون‌های آماری اضافی برای نتیجه‌گیری معنادار و تصمیم‌گیری آگاهانه بر اساس نتایج داده‌کاوی ضروری است.

  توجه به این نکته مهم است که همبستگی‌های ارائه شده بر اساس داده‌های تاریخی است و برای هر تحلیل یا نتیجه‌گیری معنی‌دار باید عوامل دیگری مانند نوع صنعت، الگوهای فصلی و تأثیرات خارجی بالقوه را در نظر گرفت. توجه شود بدون داشتن اطلاعات پیشین در مورد یک صنعت و داده‌های خاص آن، نتیجه‌گیری قطعی بسیار چالش‌برانگیز است. در واقع، جهت تجزیه‌وتحلیل بیشتر، داشتن دانش کافی و خاص مربوطه برای تفسیر کامل همبستگی‌ها و به دست آوردن بینش عملی ضروری است.

- ✓ محاسبه مقدار میانگین شاخص همبستگی روزانه برای هر شرکت و به تفکیک بر اساس داده‌های سالانه و فصلی و دسته‌بندی آنها بر اساس نتایج حاصل: در این بخش، میانگین مقدار شاخص همبستگی روزانه (بین روزهای مختلف) بر اساس داده‌های سال و همچنین داده‌های هر فصل انجام می‌شود. در این مورد، در تمامی حالات مقدار میانگین شاخص همبستگی برای تمامی شرکت‌ها کمتر از ۰,۷ بوده است که نشان‌دهنده همبستگی پایین مصرف در تمامی موارد است.
- ✓ محاسبه شاخص همبستگی فصلی (جهت برآورد مقدار همبستگی مصرف بین فصول مختلف): در این بخش مقدار همبستگی بین داده‌های بار هر فصل محاسبه شده است. این مقادیر مشخص می‌کنند که الگوی بار هر شرکت در هر فصل به چه میزان با داده‌های بار مصرفی سایر فصول همبستگی دارند. مقادیر این شاخص برای تمامی شرکت‌ها کوچک است و بنابراین همبستگی خاصی بین فصول مختلف دیده نمی‌شود.
- ✓ محاسبه شاخص‌های count، mean، std، ۲۵٪، ۵۰٪، ۷۵٪ و حداقل[۱] (Min) و حداکثر[۲] (Max) با استفاده از داده‌های سالانه و فصلی و محاسبه نسبت انحراف‌معیار بر میانگین با استفاده از داده‌های سالانه و فصلی: در این بخش، مقادیر این شاخص‌ها نیز به تفکیک بر اساس داده‌های ساعتی مربوط به ۳۶۵ روز سال و همچنین به تفکیک برای فصول مختلف محاسبه شدند. در ادمه هر کدام از این شاخص‌ها معرفی می‌شوند.
  - Count تعداد داده‌های موجود در هر بررسی را نشان می‌دهد که به عنوان مثال برای بازه سالانه، ۳۶۵ عدد به ازای هر ساعت وجود دارد.
  - Mean مقدار میانگین بار در هر ساعت را نشان می‌دهد.
  - std مقدار انحراف از معیار یا انحراف استاندارد را برای هر ساعت نشان می‌دهد. انحراف استاندارد در واقع تغییرپذیری یا گسترش مقادیر بار را در اطراف میانگین اندازه‌گیری می‌کند. مقادیر انحراف استاندارد

[۱] Minimum

[۲] Maximum

بالاتر نشان‌دهنده تنوع یا پراکندگی بیشتر بار است. به عنوان مثال، در شرکت اول، ساعت ۶ دارای انحراف استاندارد نسبتاً بالای ۱۶۶,۸۱ است که نشان‌دهنده تغییرات قابل‌توجهی در بار در طول آن ساعت در روزهای مختلف است. حال آنکه برای برای شرکت دوم، مقادیر انحراف استاندارد از حدود ۱۸,۴۸ (کمترین) تا ۱۹,۹۴ (بالاترین) متغیر است.

- Min و Max: مقادیر حداقل و حداکثر بار نشان‌دهنده کمترین و بیشترین مقدار مشاهده شده بار در هر ساعت در طول سال است. به عنوان مثال، بررسی حداقل و حداکثر بار در ساعات مختلف در شرکت اول نشان‌دهنده دامنه وسیعی از نوسانات بار است.
- چارک‌ها[۱] (۲۵٪ (Q1)، ۵۰٪ (Q2)، ۷۵٪ (Q3)): چارک‌ها اطلاعاتی در مورد توزیع مقادیر بار ارائه می‌دهند. صدک ۵۰٪ (میانه)[۲] مقدار باری است که در آن نیمی از نقاط داده در زیر و نیمی در بالای آن قرار دارند. صدک‌های ۲۵ و ۷۵ مقادیری را نشان می‌دهند که به ترتیب ۲۵ و ۷۵ درصد داده‌ها در زیر قرار می‌گیرند. این چارک‌ها به درک تمایل و گسترش مرکزی توزیع بار کمک می‌کنند. در واقع، تجزیه‌وتحلیل چارک به شناسایی ویژگی‌های توزیع بار برای هر ساعت کمک می‌کند. اگر تفاوت بین Q1 و Q3 زیاد باشد، نشان می‌دهد که توزیع بار برای آن ساعت نسبتاً گسترده است (با تغییرات قابل‌توجهی بین دوره‌های بار کم و زیاد). برعکس، اگر تفاوت بین Q1 و Q3 کم باشد، توزیع بار در اطراف میانه متمرکزتر است. در واقع، اگر صدک‌های ۲۵ و ۷۵ نزدیک باشند، نشان‌دهنده بار ثابت و پایدارتر در طول روز است. برعکس، اگر تفاوت معناداری بین Q1 و Q3 وجود داشته باشد، ممکن است الگوهای بار نوسانی را نشان دهد. اگر فاصله بین Q1 و Q3 زیاد باشد، نشان می‌دهد که داده‌های بار توزیع گسترده‌ای دارند و ممکن است نقاط پرت بیشتری داشته باشند. هنگامی که تفاوت بین Q1 و Q2 (صدک ۲۵ تا ۵۰) نسبتاً کم است، نشان می‌دهد که نیمه پایینی داده‌های بار نسبتاً فشرده است. به طور مشابه، زمانی که تفاوت بین Q2 و Q3 (صدک ۵۰ تا ۷۵) نسبتاً کم باشد، نشان می‌دهد که نیمه بالایی داده‌های بار نیز نسبتاً فشرده است. مقایسه میانه‌ها در ساعات مختلف نیز بینش‌هایی را در مورد سطوح بار در آن ساعات ارائه می‌دهد. به عنوان مثال، اگر میانه ساعت ۸ کمتر از میانه ساعت ۱۸ باشد، می‌تواند نشان‌دهنده این موضوع باشد که بار در ساعات بعد از ظهر نسبت به ساعات اولیه صبح بیشتر است. به عنوان مثال، در شرکت دوم، مقادیر Q3 برای هر ساعت از حدود ۱۴۸,۳۰ تا ۱۴۹,۹۲ متغیر است. این نشان می‌دهد که در ۷۵ درصد از ساعات سال، بار کمتر از این مقادیر است و تنها در ۲۵ درصد ساعات بالا، بار بیشتر از این مقادیر است.

به طور کلی، تجزیه‌وتحلیل این شاخص‌ها، الگوهایی را در بار صنعت در طول سال نشان می‌دهد.

همچنین، در شرکت دوم به طور کلی نتایج زیر قابل حصول است:

- ❖ همانطور که توسط مقادیر انحراف استاندارد قابل مشاهده است، مقادیر بار برای هر ساعت از روز تغییرات جزئی را در طول سال نشان می‌دهد.
- ❖ مقادیر بار از یک الگوی روزانه پیروی می‌کنند.
- ❖ به نظر می‌رسد مقادیر بار در یک محدوده مشخص و معقول هستند.

---

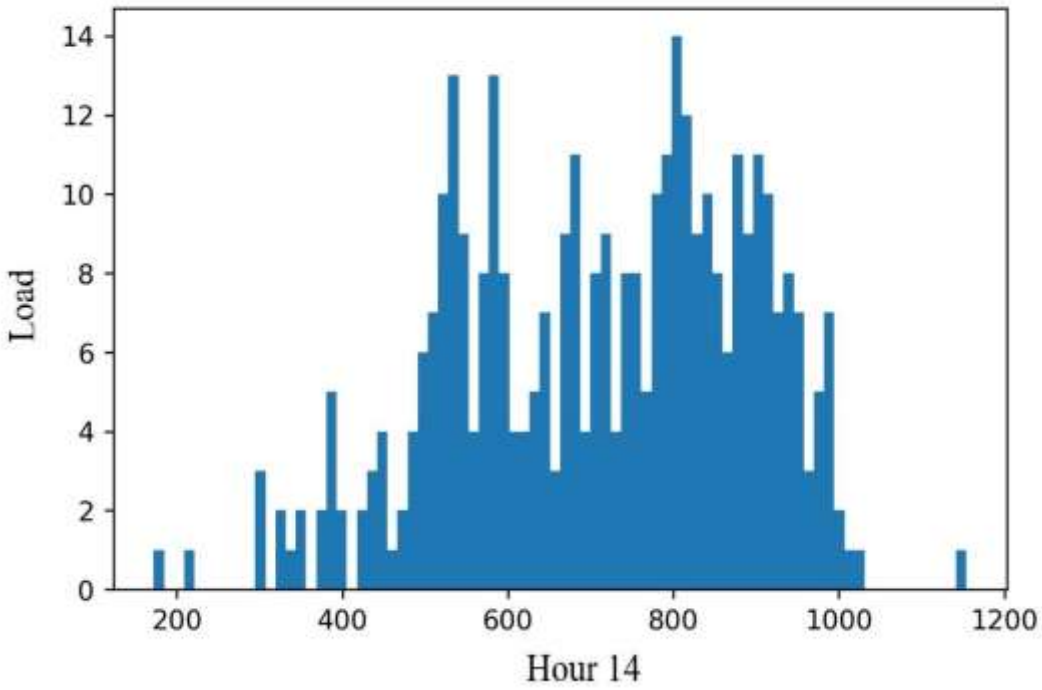
[۱] Quartiles

[۲] 50th percentile (median)

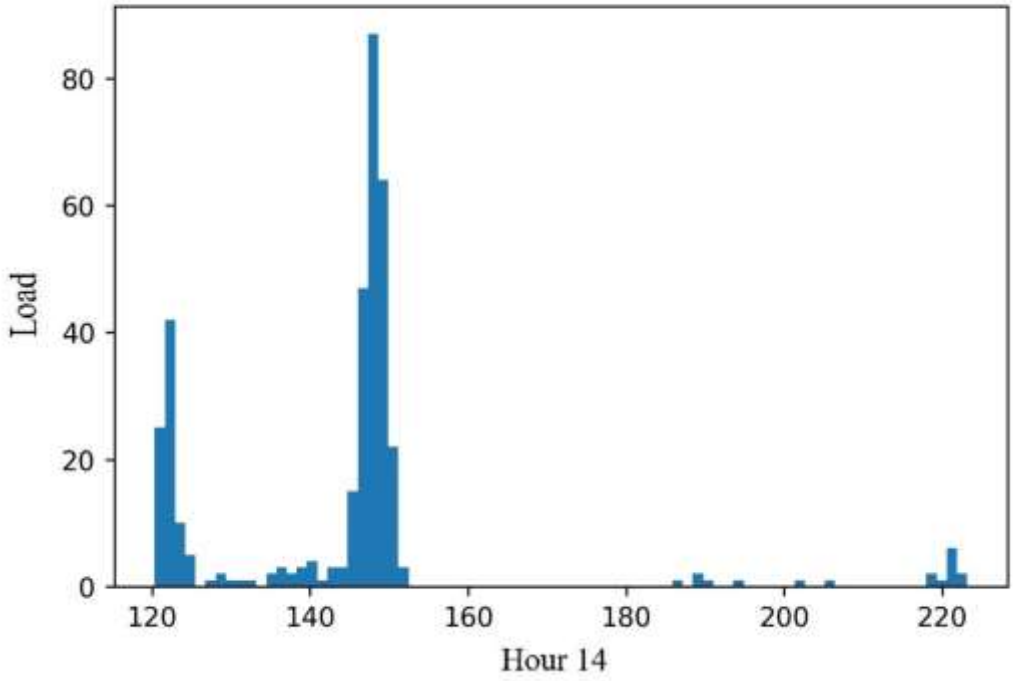
این اطلاعات می‌تواند برای پیش‌بینی بار، برنامه‌ریزی ظرفیت و درک دوره‌های اوج تقاضا در صنعت مفید باشد. هر کدام از این شاخص‌ها می‌توانند در تحلیل‌های مختلف مفید بوده و دید خاصی به کارشناس تحلیل‌گر داده بدهند.

- ✓ محاسبه شاخص ضریب تغییرات (CV)[1]: ضریب تغییرات معیاری برای تغییرپذیری نسبی است و به عنوان نسبت انحراف معیار به میانگین محاسبه می‌شود. به کمک این شاخص می‌توان تغییرپذیری بار را در ساعات مختلف نسبت به سطوح بار متوسط آنها مقایسه کرد. CV بالاتر نشان‌دهنده تنوع نسبتاً بالاتر در مقایسه با میانگین است؛ در حالی که CV پایین‌تر نشان‌دهنده الگوهای بار ثابت‌تر است. بنابراین، در این بخش مقدار نسبت انحراف معیار به میانگین نیز برای هر مورد محاسبه شد. با توجه به مطالب بیان‌شده، هر چقدر مقدار این نسبت کوچک‌تر باشد داده‌ها پراکندگی کمتری داشته و احتمالا نمودار الگوی بار یکنواخت‌تر خواهد بود. دقت شود می‌توان مقدار این شاخص را به ازای هر ساعت محاسبه کرد و در واقع مقادیر CV ساعتی حاصل شود، ولی در اینجا برای اینکه بتوان نوعی دسته‌بندی برای صنایع ارائه داد، نسبت انحراف استاندارد به میانگین برای کل ساعات محاسبه شد تا برای هر صنعت یک عدد حاصل شود و قابل مقایسه با صنایع دیگر باشد. مقدار این شاخص نیز به تفکیک با استفاده از داده‌های کل سال و نیز فصول مختلف محاسبه شد. مقدار این شاخص با استفاده از داده‌های کل سال، برای شرکت ۱ برابر با ۰,۲۳ و برای شرکت ۲ برابر با ۰,۱۳ محاسبه شده است. همانطور که انتظار می‌رفت مقدار این شاخص برای شرکت ۲ که بار یکنواخت‌تری دارد عدد کوچک‌تری است. می‌توان بر اساس این شاخص نیز نوعی دسته‌بندی انجام داد. به عنوان مثال، دسته اول را می‌توان به شرکت‌هایی اختصاص داد که مقدار نسبت انحراف از معیار به میانگین آنها از ۰,۱۵ کمتر است. در این صورت، بر اساس این دسته‌بندی شرکت ۲ در دسته اول و شرکت ۱ در دسته دوم قرار می‌گیرد.
- ✓ رسم نمودار توزیع بار برای هر شرکت جهت بررسی مقدار تکرار سطوح مختلف بار در الگوی مصرف هر شرکت: در این بخش می‌توان نمودار هیستوگرام هر شرکت را رسم نمود تا دیدی کلی نسبت به تکرار سطوح بار هر شرکت بدست آورد. در شکل‌های (۲-۲۳) و (۲-۲۴) نمودار هیستوگرام بار شرکت‌های ۱ و ۲ در ساعت ۱۴ نشان داده شده است.

[1] Coefficient of Variation (CV)

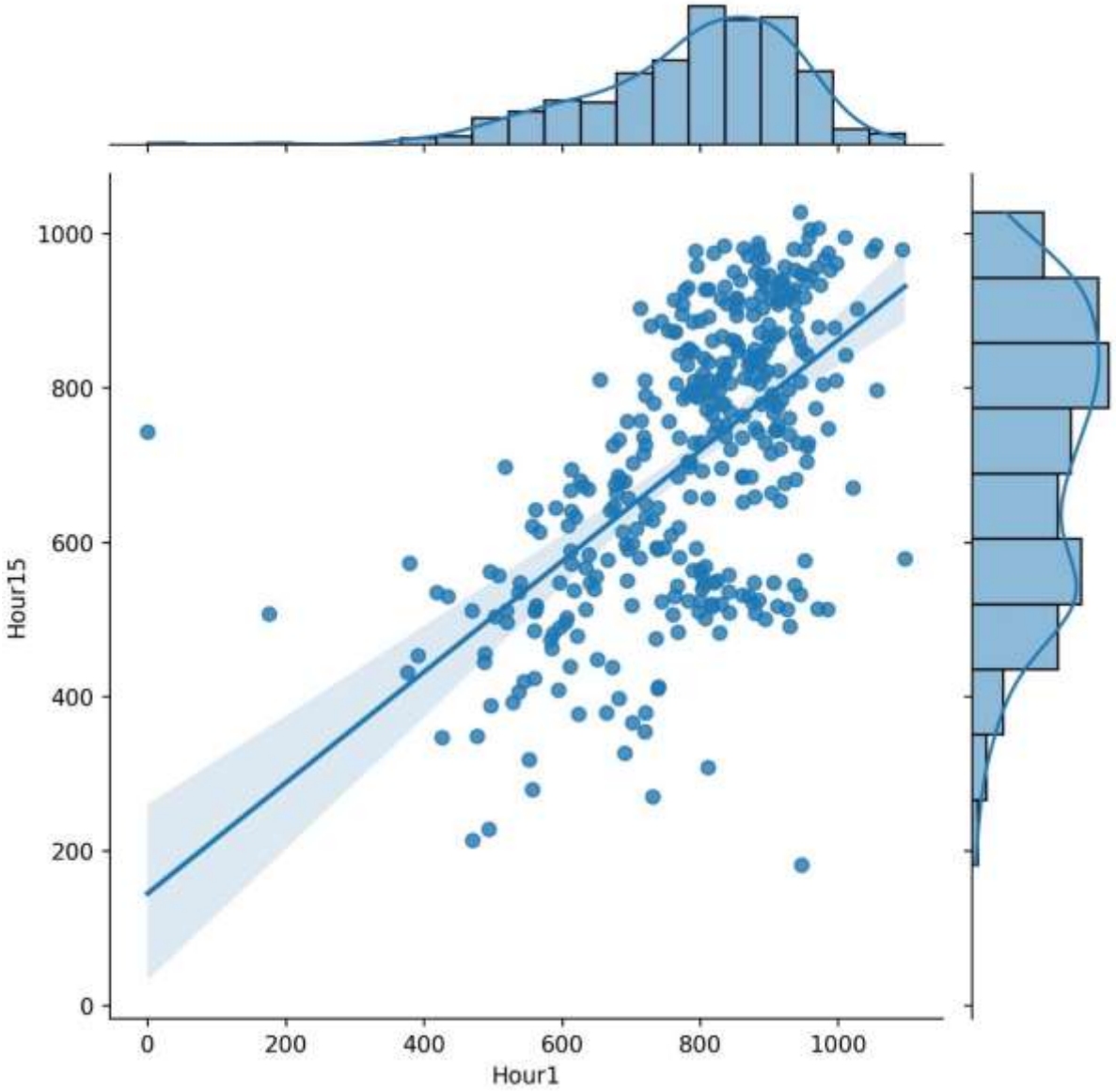


شکل ۲-۲۳: نمودار هیستوگرام ساعت ۱۴ شرکت ۱



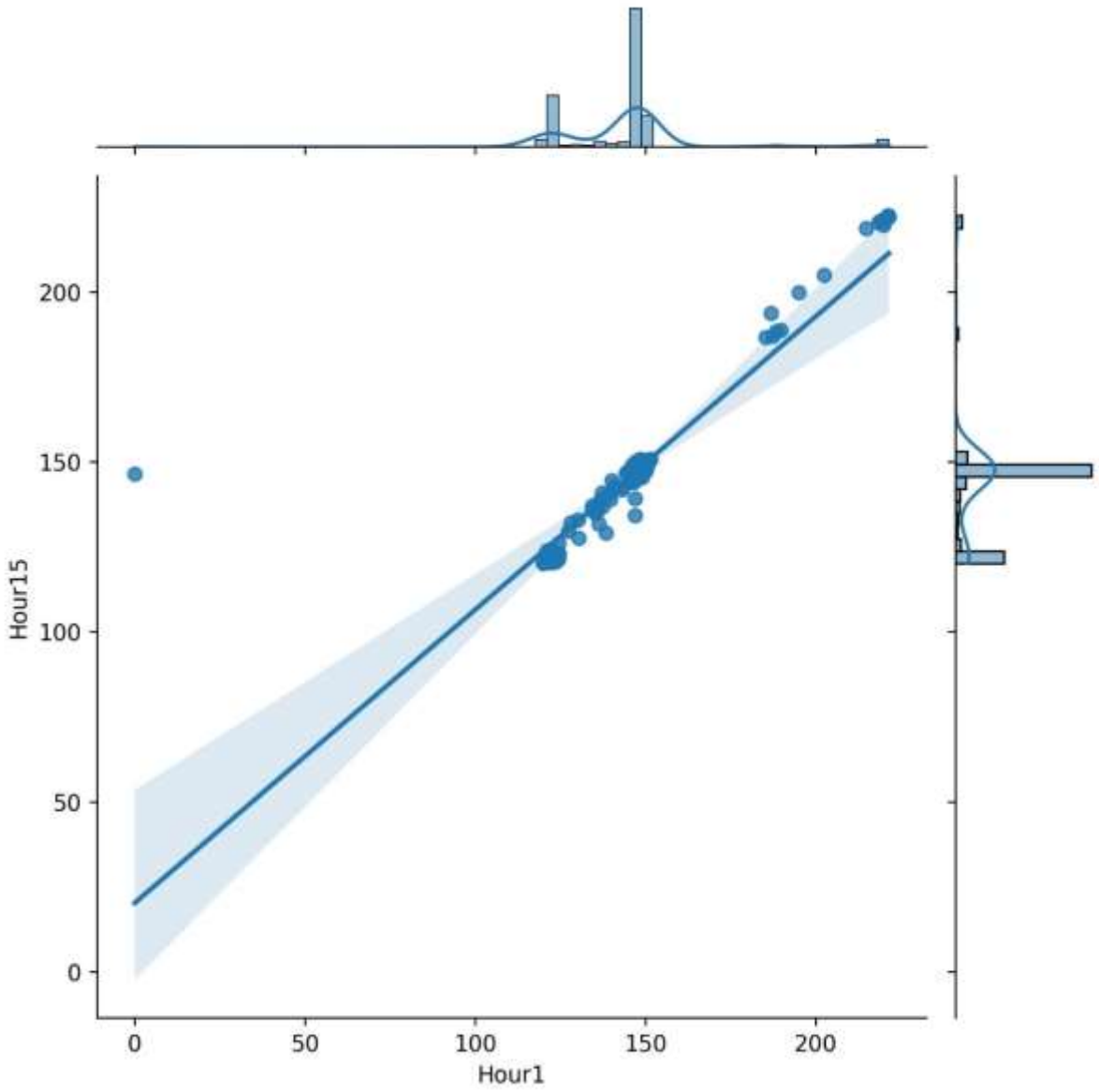
شکل ۲-۲۴: نمودار هیستوگرام ساعت ۱۴ شرکت ۲

- ✓ محاسبه رگرسیون مقدار بار در ساعات مختلف بر اساس داده‌های سالانه و رسم نمودار رگرسیون مربوطه[1]: محاسبه و رسم رگرسیون می‌تواند به خوبی میزان پراکندگی یا یکنواختی داده‌ها را نشان دهد. در شکل‌های (۲-۲۵) و (۲-۲۶) مقدار رگرسیون بار ساعت ۱ نسبت به بار ساعت ۱۵ برای شرکت‌های ۱ و ۲ نشان داده شده است. همانطور که از شکل‌های زیر پیداست، داده‌های بار در شکل مربوط به شرکت ۱ نسبتا پراکنده هستند؛ در حالی که در شکل مربوط به شرکت ۲ داده‌های بار نزدیک به خط رگرسیون بوده و یکنواختی مشخصی را نشان می‌دهند.



شکل ۲-۲۵: نمودار رگرسیون بار ساعت ۱ نسبت به ساعت ۱۵ برای شرکت ۱

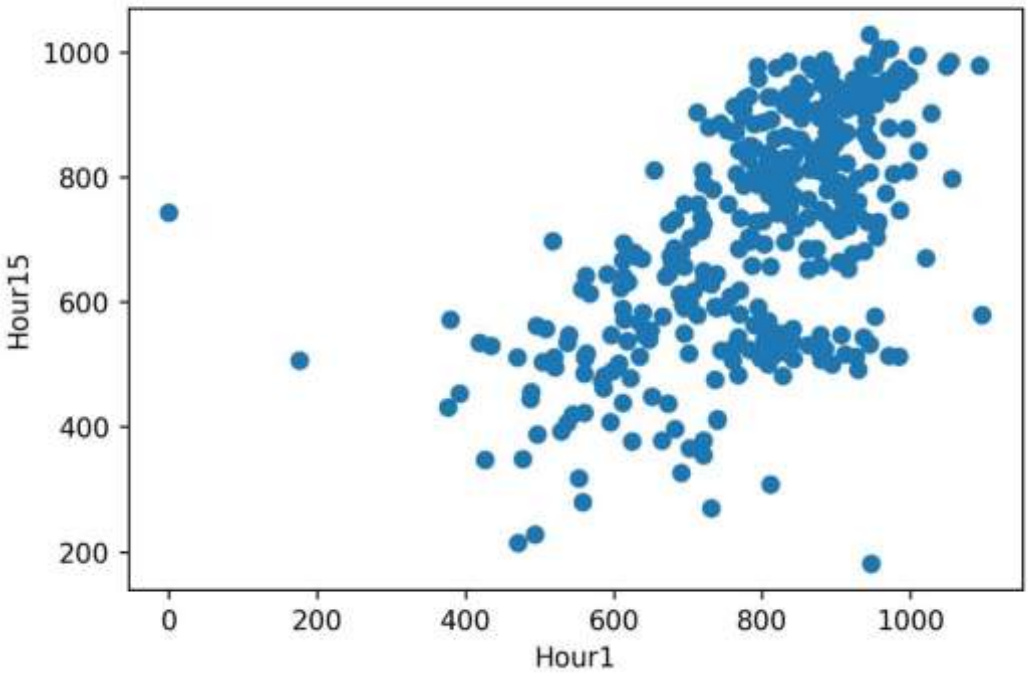
[1] جهت برآورد مقدار وابستگی بار در ساعات مختلف نسبت به یکدیگر



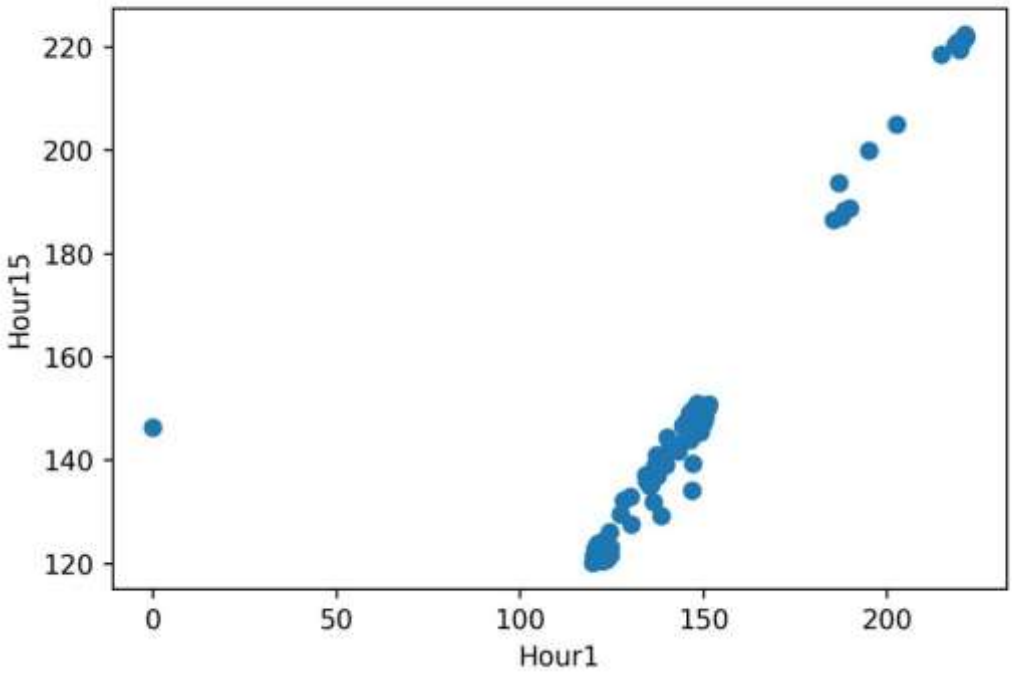
شکل ۲-۲۶: نمودار رگرسیون بار ساعت ۱ نسبت به ساعت ۱۵ برای شرکت ۲

- ✓ رسم مقادیر بار یک ساعت نسبت به مقادیر بار در ساعات دیگر[۱]: در این بخش سعی بر آن است که مقادیر بار در هر ساعت نسبت به مقادیر بار در ساعات دیگر رسم شود که این نمودارها نیز می‌تواند به خوبی میزان پراکندگی یا یکنواختی را نشان دهد (شکل‌های (۲-۲۷) و (۲-۲۸) به ترتیب مربوط به شرکت‌های ۱ و ۲).

[۱] جهت برآورد مقدار وابستگی بار در ساعات مختلف نسبت به یکدیگر

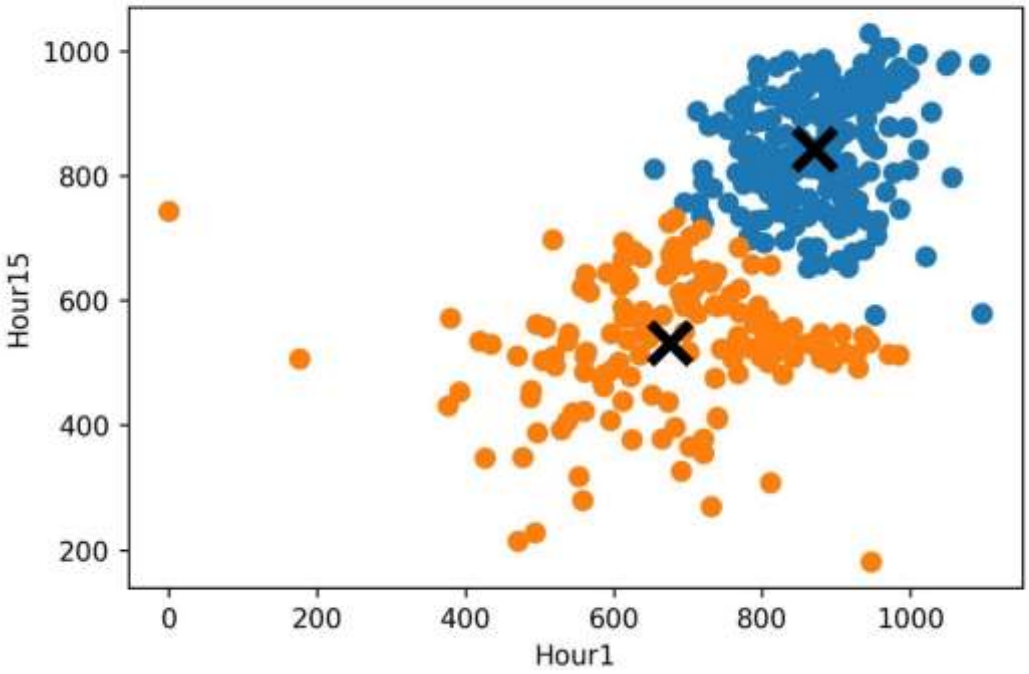


شکل ۲-۲۷: نمودار مقادیر بار ساعت ۱ نسبت به مقادیر بار ساعت ۱۵ برای شرکت ۱

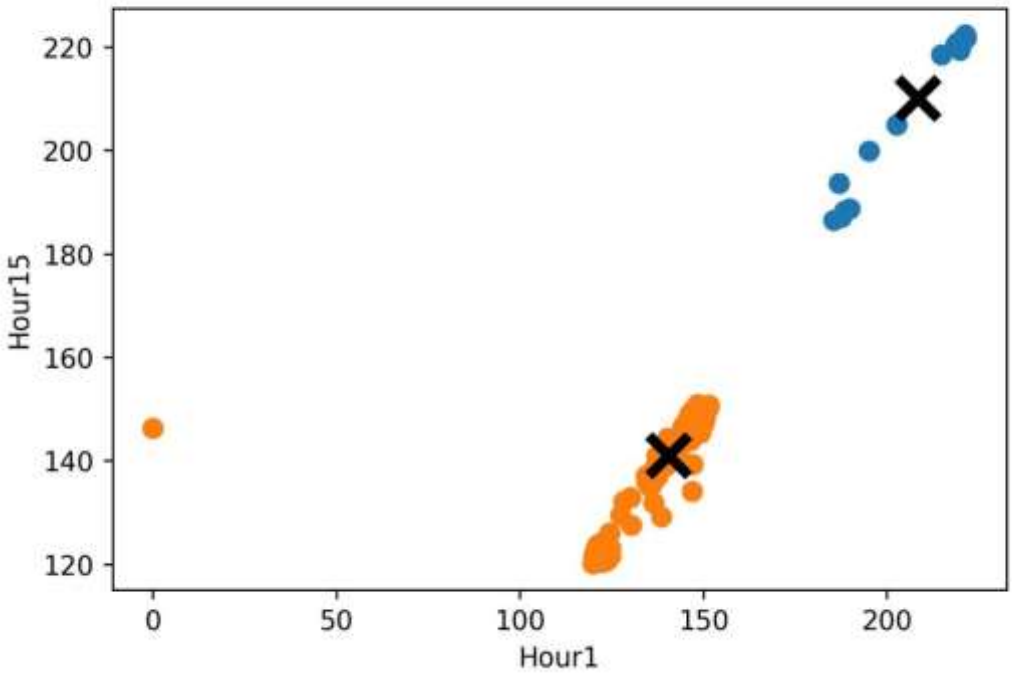


شکل ۲-۲۸: نمودار مقادیر بار ساعت ۱ نسبت به مقادیر بار ساعت ۱۵ برای شرکت ۲

- ✓ خوشه‌بندی مقادیر بار در هر ساعت و رسم نمودارهای مربوطه جهت برآورد دید کلی از سطوح تمرکز بار: در این بخش سعی بر آن است تا مقادیر بار در هر ساعت به چند دسته مشخص تقسیم‌بندی شوند که در این راستا از روش‌های مرسوم خوشه‌بندی استفاده شده است. در شکل‌های (۲-۲۹) و (۲-۳۰) دو نمونه خوشه‌بندی بار در شرکت‌های ۱ و ۲ بر اساس مقادیر بار ساعت ۱ و ۱۵ نشان داده شده است.



شکل ۲-۲۹: خوشه‌بندی بار برای شرکت ۱ بر اساس مقادیر بار در ساعات ۱ و ۱۵



شکل ۲-۳۰: خوشه‌بندی بار برای شرکت ۲ بر اساس مقادیر بار در ساعات ۱ و ۱۵

برخی دیگر از تحلیل‌هایی که می‌تواند در این حوزه صورت پذیرد عبارت هستند از:

چولگی[۱]: چولگی معیاری برای عدم تقارن توزیع مقادیر بار برای هر ساعت است. چولگی مثبت نشان‌دهنده دم سمت راست بلندتر (یعنی تعداد نقاط بار که مقادیر آن‌ها زیاد است بیشتر هستند) است، در حالی که چولگی منفی نشان‌دهنده دم سمت چپ بلندتر (یعنی تعداد نقاط بار که مقادیر آن‌ها کم است بیشتر هستند) است. مقدار چولگی صفر نشان‌دهنده توزیع کاملا متقارن است.

کشیدگی (یا کورتوز[۲]): کورتوز درجه اوج یا مسطح بودن[۳] توزیع مقادیر بار را اندازه‌گیری می‌کند. کشیدگی مثبت نشان‌دهنده توزیع اوج‌تر است، در حالی که کشیدگی منفی نشان‌دهنده توزیع صاف‌تر در مقایسه با توزیع نرمال است.

تجزیه فصلی[۴]: تجزیه فصلی را می‌توان برای استخراج الگوی فصلی و روند داده‌های بار استفاده کرد. این تجزیه‌وتحلیل می‌تواند بینش‌های ارزشمندی در مورد الگوهای بار صنعت ارائه دهد و به شناسایی هر گونه روند فصلی تکرار شونده کمک کند.

نقاط پرت: می‌توان هر یک از نقاط پرت در داده‌های بار در هر ساعت را شناسایی و تجزیه‌وتحلیل کرد. مقادیر پرت مقادیری هستند که به طور قابل‌توجهی از بقیه داده‌ها انحراف دارند و می‌توانند نشانگر رویدادهای بار غیرعادی یا خرابی سیستم باشند.

---

[۱] Skewness

[۲] Kurtosis

[۳] degree of peakedness or flatness

[۴] Seasonal Decomposition

ضریب تغییرات چارکی (CQV[1]): ضریب تغییرات چارکی یکی دیگر از معیارهای پراکندگی نسبی است و محدوده بین چارکی (IQR)[2] را به عنوان درصدی از میانه نشان می‌دهد. CQV را می‌توان با تقسیم محدوده بین چارکی (Q3 - Q1) بر میانه (Q2) و سپس ضرب در ۱۰۰ برای بیان آن به عنوان درصد محاسبه کرد.
میانگین انحراف مطلق (MAD)[3]: MAD میانگین اختلاف مطلق بین هر نقطه داده و میانگین مجموعه داده را اندازه‌گیری می‌کند. MAD برای هر ساعت را می‌توان با جمع کردن اختلاف مطلق بین هر مقدار بار و میانگین و سپس تقسیم بر تعداد نقاط داده محاسبه کرد.
در ادامه و در بخش بعدی، برخی دسته‌بندی‌های ممکن بر اساس تحلیل‌های صورت پذیرفته ارائه خواهد شد.

۳-۲ دسته‌بندی صنایع

در بخش پیشین سعی شد تا با استفاده از ابزار و تحلیل‌های گوناگون بتوان الگوی بار در ساعات، روزها، هفته‌ها یا فصول مختلف برای هر شرکت را تحلیل و بررسی نمود تا بتوان بر این اساس راهکار مناسب جهت سیاست‌گذاری بهینه با هدف اصلاح الگوی مصرف را اتخاذ نمود. همچنین سعی شد با استفاده از این ابزارها و تحلیل‌ها بتوان تمییزی بین شکل بار و الگوی مصرف صنایع مختلف قائل شد تا بتوان دسته‌بندی مناسبی را متناسب با وضعیت هر شرکت ارائه داد. چنین دسته‌بندی و تصمیماتی می‌تواند در سیاست‌گذاری بهینه جهت مدیریت انرژی صنایع راهگشا باشد. لازم است تاکید شود بررسی‌های انجام شده در این پروژه صرفا یک تحلیل اولیه در این مسیر است و جهت اتخاذ تصمیم نهایی لازم است با شناخت هر چه بیشتر مساله پیش‌رو و ویژگی‌های صنایع مختلف، تحلیل‌های جامع‌تر انجام شده و قیود بسیار متنوعی شامل قیود فنی و غیرفنی در نظر گرفته شود. لازم به ذکر است با توجه به ابعاد مساله و دامنه اثرگذاری آن لازم است قیود سایر حوزه‌ها (مانند قیود موجود از سمت وزارت صمت جهت پیشگیری از آثار جانبی از قبیل کاهش تولید یا تورم) با دقت لحاظ شوند. به هر حال در این بخش سعی شده تحلیلی کلی ارائه شود و در واقع این مطالب آغازی بر مسیر پیش‌رو در مطالعات آینده است.
دسته‌بندی صنایع برای طراحی تعرفه برق مستلزم شناسایی معیارها یا ویژگی‌هایی است که آنها را از نظر الگوهای مصرف برق متمایز می‌کند. در ادامه برخی معیارها جهت تعریف دسته‌های مختلف به منظور طراحی تعرفه برق معرفی می‌شود:
۱. الگوهای مصرف برق:
- صنایع با پیک تقاضای برق بالا ممکن است نیاز به ساختارهای تعرفه‌ای متفاوتی برای تشویق جهت مشارکت در مدیریت بار در ساعات اوج مصرف داشته باشند. همچنین، صنایع با تقاضای برق ثابت ممکن است نیازهای تعرفه‌ای متفاوتی نسبت به صنایعی با الگوهای تقاضای متغیر داشته باشند.
۲. ساعت کاری:

[1] Coefficient of Quartile Variation (CQV)
[2] Interquartile range (IQR)
[3] Mean Absolute Deviation (MAD)

- صنایعی که به صورت شبانه‌روزی کار می‌کنند (عملیات ۷/۲۴[۱]) ممکن است در مقایسه با صنایعی که ساعات کاری مشخصی دارند، الزامات تعرفه‌ای متفاوتی داشته باشند. در نظر داشتن آیا صنعت ۷/۲۴ کار می‌کند یا ساعات کاری مشخصی دارد برای طراحی تعرفه بسیار مهم است.

۳. شدت انرژی:

- شدت انرژی بالا[۲]: صنایعی که برای فرآیندهای خود به انرژی قابل‌توجهی نیاز دارند ممکن است برای مدیریت موثر هزینه‌ها به تعرفه‌های متناسب خود نیاز داشته باشند.

صنعت فولاد می‌تواند به عنوان نمونه‌ای از صنایع با بارهای متوسط بالا در نظر گرفته شود که مصرف انرژی آن به طور مداوم در طول روز بالا است. در این موارد طراحی تعرفه‌هایی منظم با هدف ثبات و مدیریت بار می‌تواند مفید باشد.

توجه: بررسی مقادیر همبستگی بین فرآیندهای صنعت و مقدار مصرف برق در بازه‌های زمانی خاص با مقدار بار شبکه در این بازه‌های زمانی می‌تواند به فرآیند دسته‌بندی کمک کند که ممکن است برای مدیریت موثر بار در بازه‌های اوج بار مفید باشد. همچنین، اطلاعات مربوط به شدت بار صنعت در ساعات خاص شبکه (ساعات اوج بار در شبکه) نیز می‌توان دید مناسبی در راستای دسته‌بندی صنایع فراهم کند.

- صنایع با میانگین بارهای متوسط: در این حالت نیز می‌توان تعرفه‌هایی را با هدف تشویق جهت تسطیح بار طراحی کرد.

در واقع، صنایع با ظرفیت نصب شده متفاوت ممکن است به ساختارهای تعرفه‌ای متفاوتی برای تطبیق با هزینه‌های مربوط به ظرفیت خود نیاز داشته باشند

۴. تغییرپذیری فصلی:

- تقاضای فصلی: برای صنایعی که مقدار مصرف آنها بر اساس فصول مختلف متفاوت است، ممکن است بهتر باشد از تعدیل تعرفه‌های فصلی استفاده شود. اگر داده‌های بار صنعت تغییرات قابل‌توجهی را در طول فصول نشان دهد، این نشان‌دهنده تغییرات فصلی در تقاضای برق است. صنایع با تغییرات فصلی مشخص را می‌توان بر این اساس طبقه‌بندی کرد و تعدیل تعرفه‌های فصلی را در این‌گونه موارد در نظر گرفت.

۵. بهره‌وری انرژی[۳]:

- صنایعی که در فناوری‌های کارآمد انرژی سرمایه‌گذاری کرده‌اند ممکن است به تعرفه‌هایی نیاز داشته باشند که تلاش‌های آنها را پاداش دهد. در صورتی که این اقدامات در دسته‌بندی صنایع و تنظیم تعرفه‌ها دیده شود، می‌تواند به ساختارهای تعرفه‌ای منجر شود که به چنین اقداماتی پاداش می‌دهد و بهبود کارایی را تشویق می‌کند.

۶. انعطاف‌پذیری و قابلیت پاسخگویی بار:

- صنایعی که می‌توانند فعالیت‌های خود را به ساعات غیر اوج مصرف تغییر دهند، ممکن است از تعرفه‌های مبتنی بر زمان بهره‌مند شوند. تجزیه‌وتحلیل داده‌های بار و تغییرات آن در ساعات مختلف می‌تواند انعطاف‌پذیری بار صنعت را تعیین کند. صنایعی که توانایی تغییر مصرف برق خود را به ساعات غیر اوج مصرف نشان می‌دهند،

---

[۱] 24/7 Operations

[۲] Energy Intensity

[۳] Energy Efficiency

می‌توانند به عنوان بار انعطاف‌پذیر طبقه‌بندی شوند. به عنوان مثال، برای صنایعی که قادر به تنظیم و کاهش مصرف انرژی خود در ساعات اوج مصرف هستند، می‌توان در این دوره‌ها از تعرفه‌هایی با قیمت بالاتر استفاده کرد.

- برای صنایعی که قادر به تنظیم سریع تقاضای برق خود در پاسخ به سیگنال های شبکه هستند ممکن است بهتر باشد از ساختارهای تعرفه پویا[۱] استفاده شود. همچنین، مقادیر همبستگی در ساعات معین می‌تواند نشان‌دهنده این باشد که صنعت چه میزان دارای پتانسیل پاسخگویی بار ساعتی هست. این مقادیر می‌تواند نشان دهد که آیا صنعت برای ساختارهای تعرفه‌ای پویا مناسب است و می‌تواند مصرف برق خود را در پاسخ به سیگنال‌های شبکه تنظیم کند یا خیر.

۷. ضریب بار[۲]:

- ضریب بار: صنایع با ضرایب بار مختلف (نسبت بار متوسط به بار اوج) ممکن است به تعرفه‌هایی همسو با ضریب بار خود نیاز داشته باشند. این مشخصات می‌تواند بر طراحی تعرفه تأثیر بگذارد، زیرا صنایع با ضرایب بار متغیر ممکن است به تنظیمات تعرفه خاصی نیاز داشته باشند تا بار خود را به طور مؤثر مدیریت کنند.
- صنایع با بارهای پراکنده، می‌توانند تعرفه‌های تنظیم شده در ساعات مختلف را تجربه کنند تا از مصرف انرژی بیشتر در زمان‌های اوج مصرف جلوگیری کنند.

۸. تعاملات با شبکه[۳]:

- صنایعی که دارای منابع تولیدپراکنده یا تجهیزات ذخیره انرژی هستند ممکن است ملاحظات تعرفه‌ای خاصی داشته باشند. نظام دسته‌بندی می‌تواند به نوعی تعرفه‌گذاری منجر شود که با در نظر داشتن امکان تعامل صنعت با شبکه یا ذخیره انرژی در محل، ثبات شبکه را ارتقا دهد یا مشوق‌هایی برای استفاده از ذخیره‌سازی انرژی ارائه دهد.

۹. نوع صنعت[۴]:

- انواع مختلف صنایع دارای پروفایل‌های بار منحصر به فرد هستند. در نظر گرفتن نوع خاص این صنعت، اعم از تولیدی، تجاری، بهداشتی و غیره، می‌تواند به تحلیل[۵] داده‌های بار کمک کند. صنایع از انواع مختلف ممکن است الگوهای بار منحصر به فردی را نشان دهند که منجر به طبقه‌بندی بر اساس نوع صنعت آنها می‌شود.

۱۰. موقعیت جغرافیایی:

- صنایع در مناطق مختلف ممکن است به دلیل پویایی عرضه و تقاضای انرژی منطقه‌ای، ساختارهای تعرفه‌ای متفاوتی داشته باشند. در واقع، صنایعی که در مناطق مختلف فعالیت می‌کنند ممکن است بر اساس موقعیت جغرافیایی خود به ساختارهای تعرفه‌ای متناسب نیاز داشته باشند.

۱۱. تأثیر زیست‌محیطی:

---

[۱] dynamic tariff structures

[۲] Load Factor

[۳] Grid Interactions

[۴] Industry Type

[۵] contextualize

- صنایعی که بر پایایی[1] تمرکز دارند ممکن است به تعرفه‌هایی نیاز داشته باشند که با اهداف زیست‌محیطی آنها هماهنگ باشد.

۱۲. الزامات قانونی[2]:

- صنایع مشمول دستورات نظارتی[3] برای بهره‌وری انرژی یا مدیریت بار ممکن است برای تسهیل انطباق با مقررات موجود به تعرفه‌های خاصی نیاز داشته باشند.

۱۳. عوامل اقتصادی:

- صنایع با حاشیه سود متفاوت ممکن است به ساختارهای تعرفه‌ای نیاز داشته باشند که صرفه‌جویی در هزینه و نیازهای عملیاتی را متعادل کند.

با ادغام این معیارهای کلیدی با مشخصات صنعت و بینش‌های به دست آمده از داده‌ها و تجزیه‌وتحلیل‌ها، می‌توان طبقه‌بندی‌های معناداری برای طراحی تعرفه برق ایجاد کرد که با الگوهای مصرف خاص و ویژگی‌های عملیاتی صنعت همسو باشد. بنابراین، ترکیب و تجزیه‌وتحلیل این معیارها می‌تواند منجر به ایجاد دسته‌های متمایز با ساختارهای تعرفه‌ای متناسب با نیازهای منحصر به فرد هر بخش از صنعت شود. هدف، طراحی تعرفه‌هایی با هدف حصول بهره‌وری انرژی، مدیریت بار و الگوی مصرف انرژی پایا است. این تعاریف و مثال‌ها نشان می‌دهند که بر اساس الگوهای مصرف انرژی و نیازهای متنوع صنایع، می‌توان تعرفه‌های مناسبی را برای هر دسته طراحی کرد.

بنابراین، بر اساس داده‌های هر شرکت و تحلیل‌های صورت‌پذیرفته روی آنها می‌توان ملاک‌هایی را در نظر داشت و بر اساس آن سعی بر دسته‌بندی صنایع شود. با توجه به داده‌های مصرف بار در طول یک بازه زمانی مشخص و تحلیل انجام شده، به همراه دانش کلی در مورد طراحی تعرفه برق و مدیریت بار، می‌توان چندین معیار یا ویژگی خاص را شناسایی کرد که برای طبقه‌بندی صنایع به منظور طراحی تعرفه برق استفاده کرد.

در ادامه چندین معیار خاص به همراه تحلیل بیشتری از آنها آورده شده است.

- تقاضای بار بیشینه
  - ✓ معیار: بیشینه مقدار بار مشاهده شده در طول دوره پرباری.
  - ✓ تحلیل: صنایع با تقاضای بار بیشتر در طول دوره پرباری ممکن است نیازمند ساختار تعرفه‌ای به منظور کاهش مقدار بار در ساعات اوج بار شبکه باشد.
- تغییرپذیری بار
  - ✓ معیار: انحراف معیار بارها که نشان‌دهنده تغییرپذیری مقدار بار نسبت به میانگین است.
  - ✓ تحلیل: صنایع با تغییرپذیری بالاتر می‌توانند از گزینه‌های تعرفه‌ای منعطف برای مدیریت الگوهای متغیر مصرف انرژی خود بهره‌برداری کنند. این صنایع می‌توانند از تعرفه‌های منعطف و مبتنی بر زمان استفاده کنند تا از مزایای مصرف بهینه انرژی در بازه‌های مختلف بهره‌مند شوند.
- توزیع مصرف انرژی

[1] sustainability

[2] Regulatory Requirements

[3] regulatory mandates

- ✓ معیار: تفاوت بین چارک اول و چارک سوم نمایانگر پراکندگی توزیع بار است.
- ✓ تحلیل: توزیع گسترده‌تر ممکن است نشان‌دهنده صنایعی با الگوهای بار متنوع باشد که به طرح‌های تعرفه‌ای متناسب با این تنوع نیاز دارند. برای صنایع با پراکندگی بار بالا، می‌توان تعرفه‌های مختلفی بر اساس بازه‌های زمانی مختلف ایجاد کرد تا به مدیریت پراکندگی بار کمک کرد.

- افزایش ناگهانی بار در بازه مشخصی از روز (مثلا صبح، ظهر یا شب)
  - ✓ معیار: درصد افزایش مقدار بار از حداقل بار آن بازه.
  - ✓ تحلیل: صنایع با افزایش ناگهانی بار در یک بازه زمانی ممکن است نیازمند ساختارهای تعرفه‌ای خاصی باشند تا این افزایش را به خوبی مدیریت کنند.
- نسبت اوج نیمه‌روزی
  - ✓ معیار: نسبت مقدار بار میانگین در اوج نیمه‌روز (مثلاً ۱۲ تا ۱۴) به میانگین کل مصرف بار.
  - ✓ تحلیل: نسبت اوج نیمه‌روزی بالاتر ممکن است برای طراحی تعرفه‌هایی که به بازه‌های اوج بار در این ساعات توجه می‌کند، مورد استفاده قرار گیرد.
- میانه مقدار بار
  - ✓ معیار: میانه مقدار بار برای کل روز.
  - ✓ تحلیل: مقادیر میانه می‌توانند بینش‌هایی را در مورد تمایل مصرف بار ارائه دهد و به طراحی تعرفه‌هایی متناسب با نوع الگوی بار صنایع کمک کند.
- میانگین بار
  - ✓ معیار: میانگین کل مصرف بار روزانه.
  - ✓ تحلیل: این معیار می‌تواند به عنوان معیار میانگین مصرف بار برای هر صنعت استفاده شود. صنایع با میانگین بار بالاتر ممکن است نیازمند تعرفه‌هایی باشند که انگیزه لازم جهت کاراتر نمودن فرآیند مصرف انرژی را ایجاد کند.

با در نظر گرفتن چنین معیارهایی می‌توان صنایع را در طبقات مربوطه برای طراحی تعرفه برق دسته‌بندی کرد. هنگامی که صنایع دسته‌بندی می‌شوند، ساختارهای تعرفه را می‌توان متناسب با ویژگی‌های اصلی هر دسته تنظیم کرد تا به اهداف موردنظر از قبیل بهبود بهره‌وری انرژی، مدیریت بهینه بار و پایداری کلی شبکه رسید.

در ادامه، سعی بر آن است تا بار دو شرکت معرفی‌شده در این گزارش با جزئیات بیشتری بررسی و تحلیل شوند.

با توجه به مطالب بیان شده در بخش پیشین، به نظر می‌رسد شرکتی مانند شرکت ۱ که وابستگی خاصی بین ساعات و روزهای مختلف نشان نمی‌دهد را می‌توان در دسته‌ای قرار داد که می‌توان انواع مختلف تعرفه‌های ساعتی، روزانه، ماهانه یا فصلی را روی آن اعمال کرد و نسبت به هر کدام از این تعرفه‌ها عکس‌العمل مشخصی را ارائه خواهد داد که قطعا جهت اطمینان از پاسخگو بودن و شکل پاسخگویی و اصلاح رفتار نیاز است پروسه فعالیت آن صنعت به دقت بررسی شود.

در تحلیل الگوی بار شرکت ۲ مشاهده می‌شود که این شرکت غالبا رفتار مصرفی یکنواخت دارد و احتمال می‌رود سیاست‌هایی جهت تغییر ساعتی میزان بار برای این شرکت چندان مناسب نباشد. بنابراین، در رابطه با شرکت‌هایی که رفتاری مشابه با شرکت ۲ دارند، احتمالا باید بیشتر بر روی تغییر رفتار روزانه، هفتگی، ماهانه یا فصلی تمرکز داشت.

بنابراین، این دسته از شرکت‌ها در دسته دوم قرار می‌گیرند که سیاستی متفاوت با دسته اول را جهت اصلاح الگوی بار می‌طلبند. در این مورد نیز قطعا جهت تصمیم‌گیری نهایی نیاز به بررسی دقیق‌تر پروسه تولید و نیز میزان پاسخگویی و نحوه پاسخگویی شرکت است.

به عنوان مثال، برخی پیشنهادات برای شرکت‌های ۱ و ۲ که در این گزارش بررسی شدند به شرح زیر است:

- ✓ پیشنهادات برای شرکت اول:

بر اساس داده‌های ارائه شده، چند پیشنهاد برای طراحی تعرفه و برنامه‌های مدیریت بار برای این صنعت وجود دارد:

- ❖ تعرفه زمان استفاده[۱]: با توجه به تغییر بار در طول روز، اجرای تعرفه زمان استفاده می‌تواند برای این صنعت مفید باشد. با قیمت‌گذاری TOU، نرخ‌های متفاوتی را می‌توان برای دوره‌های زمانی مختلف، مانند ساعات اوج، کم‌باری، و ساعت‌های میان‌باری اعمال کرد. این مساله صنعت را تشویق خواهد کرد تا برخی از فرآیندهای انرژی‌بر خود را به ساعات غیر اوج که نرخ برق پایین‌تر است تغییر دهد و منجر به صرفه‌جویی در هزینه شود.
- ❖ اقدامات بهره‌وری انرژی: انجام ممیزی انرژی برای شناسایی حوزه‌هایی که می‌توان در آنها بهبود بهره‌وری انرژی داشت و در ادامه اجرای اقدامات بهره‌وری انرژی می‌تواند منجر به صرفه‌جویی طولانی مدت در هزینه‌ها و مزایای زیست‌محیطی شود.
- ❖ توسعه و استفاده از منابع انرژی تجدیدپذیر: سرمایه‌گذاری و استفاده از منابع انرژی تجدیدپذیر مانند منابع خورشیدی یا بادی می‌تواند سبب کاهش هزینه‌های انرژی در بلندمدت شده و مزایای دیگری (از قبیل کاهش تولید کربن و مزایای زیست‌محیطی) را نیز در بر داشته باشد.
- ❖ توسعه و استفاده از منابع ذخیره انرژی الکتریکی: توسعه و استفاده از سیستم‌های ذخیره‌ساز انرژی برای ذخیره انرژی اضافی در ساعات کم‌باری و استفاده از آن در ساعات اوج بار می‌تواند به مدیریت بار کمک کند.
- ❖ شبکه هوشمند و سیستم‌های مدیریت انرژی: با پیاده‌سازی فناوری‌های شبکه هوشمند و سیستم‌های مدیریت انرژی که داده‌های بلادرنگ مصرف انرژی را ارائه داده و امکان مدیریت خودکار بار را فراهم می‌کند، می‌توان مصرف انرژی را بهینه کرد.

به طور کلی، ترکیبی از طراحی تعرفه، استراتژی‌های مدیریت بار و اقدامات بهره‌وری انرژی می‌تواند به صنعت کمک کند تا مصرف انرژی خود را بهینه کرده و هزینه‌ها را کاهش دهد. با این حال، انجام یک تجزیه‌وتحلیل دقیق از نیازها، اهداف و محدودیت‌های خاص صنعت برای ایجاد بهترین راه‌حل‌ها ضروری است.

- ✓ پیشنهادات برای شرکت دوم:

بر اساس داده‌های بار ارائه شده، در ادامه چند پیشنهاد برای طراحی تعرفه و برنامه‌های مدیریت بار برای این شرکت آورده شده است.

[۱] Time of Use (TOU)

- ❖ برنامه‌های پاسخگویی بار: با استفاده از انواع برنامه‌های پاسخگویی بار می‌توان مشوق‌های مالی برای صنعت به منظور کاهش مصرف برق در زمان اوج بار را ارائه داد. در این راستا می‌توان از سیستم‌های خودکار برای اطلاع‌رسانی به صنعت در دوره‌های اوج بار و مشوق‌های مربوطه برای کاهش بار استفاده کرد. بنابراین، با تشویق به جابجایی بار از طریق بهینه‌سازی زمان‌بندی فرآیندهای پرمصرف برق، این صنعت می‌تواند در زمان‌هایی که تقاضای برق کمتر است، عملیات پرانرژی را فعال کند.
- ❖ بهره‌وری انرژی: انجام ممیزی انرژی و اجرای سیاست‌های افزایش بهره‌وری می‌تواند سبب کاهش کل مصرف انرژی الکتریکی شود.
- ❖ توسعه و استفاده از منابع انرژی تجدیدپذیر: مشابه آنچه پیشتر بیان شد، سرمایه‌گذاری و استفاده از منابع انرژی تجدیدپذیر می‌تواند سبب کاهش هزینه‌های انرژی شده و مزایای دیگری (از قبیل مزایای زیست‌محیطی) را نیز به همراه داشته باشد.
- ❖ استفاده از ذخیره‌ساز انرژی: سرمایه‌گذاری در توسعه و استفاده از تجهیزات ذخیره‌سازی انرژی مانند باتری‌ها، به منظور ذخیره انرژی الکتریکی در ساعات کم‌باری و استفاده از آن در ساعات اوج مصرف می‌تواند به کاهش پیک بار و مدیریت مصرف کمک کند.
- ❖ شبکه هوشمند و سیستم‌های مدیریت انرژی: همانطور که بیان شد، استفاده از ظرفیت و امکانات شبکه هوشمند می‌تواند منجر به مدیریت بهینه بار شرکت شود.

در کنار موارد بیان شده در فوق، تلاش‌های جمعی مانند همکاری صنایع یا تاسیسات منطقه برای به اشتراک گذاشتن اطلاعات و استراتژی های مدیریت بار می‌تواند به شیوه‌های مدیریت بار مؤثرتر و حفظ انرژی کلی منجر شود.

هدف این پیشنهادها بهینه‌سازی مصرف انرژی در صنعت، کاهش پیک بار و ارتقای بهره‌وری انرژی است. با اجرای این طراحی تعرفه و برنامه‌های مدیریت بار، صنعت می‌تواند به یک سیستم انرژی پایدارتر و انعطاف‌پذیرتر کمک کند و در عین حال به طور بالقوه هزینه‌های عملیاتی را نیز کاهش دهد.

۴-۲ جمع‌بندی و نتیجه‌گیری

در این فصل سعی شد تا بتوان تحلیلی اولیه را نسبت به مقادیر بار شرکت‌های مختلف ارائه داد. در این راستا چندین شاخص و روش مختلف ارائه شد تا بر اساس آن بتوان مصارف صنایع مختلف را مقایسه و دسته‌بندی کرد. همچنین، هدف دیگر از انجام چنین تحلیل‌هایی بررسی شکل مصرف و احتمال تغییر رفتار مصرف شرکت‌های مختلف در بازه‌های زمانی متنوع (ساعتی، روزانه، هفتگی، ماهیانه یا فصلی) بوده است. به عنوان مثال، با انجام چنین تحلیل‌هایی این‌گونه نتیجه شد که شکل مصرف شرکت ۱ نشان از مقداری پراکندگی و عدم وابستگی بین بار در ساعات یا روزهای مختلف است؛ حال آنکه به نظر می‌رسد شرکت ۲ الگوی مصرفی نسبتا یکنواخت دارد و احتمالا باید سیاست اصلاح الگوی مصرف آن به گونه‌ای تنظیم شود که در افق‌هایی بلندتر از افق ساعتی باشد (مثلا برنامه‌هایی با پتانسیل اثرگذاری هفتگی یا ماهیانه).

با توجه به مطالب بیان شده در این فصل نتیجه می‌شود که جهت اصلاح تعرفه هر شرکت یا صنعت نیاز است با استفاده از تحلیل‌هایی از قبیل موارد بیان شده در این فصل، دید مناسبی از شکل مصرف و بار هر شرکت داشت و سپس بایستی صنایع مختلف در دسته‌های متناسبی تقسیم‌بندی شوند. سپس، سعی شود الزامات کلی هر دسته بررسی و به دقت مشخص شود تا بتوان بر اساس آن سیاست‌گذاری کلی در رابطه با نحوه برخورد با هر دسته را تنظیم نمود. مسلما

جهت اتخاذ تصمیم نهایی بایستی ابعاد مختلف فنی و غیرفنی (از قبیل ابعاد اجتماعی، اقتصادی، قانونی و قیود حاکم بر مساله از سمت سایر نهادها از قبیل قیود اعمالی از سمت وزارت صمت جهت پیشگیری از کاهش محصول در بازار و ایجاد تورم) را به دقت بررسی و مدلسازی نمود. در نهایت یک مساله بهینه‌سازی حاصل خواهد شد که تابع هدف آن با توجه به اهداف مدنظر از اصلاح الگوی مصرف و تنظیم تعرفه‌ها تعیین می‌شوند و قیود گسترده‌ای را در بر خواهد داشت که پوشش‌دهنده ابعاد وسیع فنی و غیرفنی خواهد بود. صرفا با انجام چنین تحلیل جامعی می‌توان نسبت به اصلاح تعرفه و تنظیم الگوی مصرف مشترکین گام برداشت و ریسک‌های ناشی از ابعاد گوناگون فنی و غیرفنی را کاهش داد.

جمع‌بندی و نتیجه‌گیری

در سالیان اخیر به دلیل رشد زیاد مصرف انرژی الکتریکی و نبود توسعه کافی در سمت تولید، ناترازی بین میزان تولید و مصرف رخ داده است. یکی از اقدامات مهم که می‌تواند در افق زمانی کمتر (نسبت به توسعه زیرساخت تولید، انتقال و توزیع انرژی الکتریکی) جهت کنترل مصرف و خصوصا عبور از پیک مصرف مفید باشد، مدیریت انرژی سمت مشترکین است. در این پروژه، مدیریت انرژی مشترکین صنعتی مورد نظر بوده است. در این راستا، هدف بر آن بوده است تا با تحلیل برخی داده‌های در دسترس بتوان دیدی کلی از شکل مصرف داشت تا بتوان برنامه‌ای متناسب و کارا را جهت اصلاح الگوی مصرف این دسته از مشترکین پیشنهاد داد. لازم به ذکر است بررسی دقیق و اتخاذ تصمیمات راهبردی نیازمند مدل‌سازی دقیق مساله، تعریف صحیح تمامی توابع هدف و به خصوص در نظر داشتن تمامی قیود فنی و غیر فنی در شبکه برق کشور است. در واقع، علاوه بر بعد فنی کار، بایستی تمامی جوانب دیگر از قبیل اجتماعی، قانونی و اقتصادی را نیز در نظر داشت و آثار جانبی تصمیمات ممکن را نیز بررسی کرد. به عنوان نمونه، در صورت کاهش تولید یک صنعت در بازه‌ای از سال در نتیجه تغییر تعرفه برق آن صنعت، آیا ممکن است کمبود آن کالا در بازار رخ دهد و در نتیجه تورم ایجاد شود؟ آیا نارضایتی اجتماعی ممکن است در نتیجه یک تصمیم کلان و آثار جانبی آن ایجاد شود؟ یا اینکه ممکن است کالایی حیاتی بوده و کاهش آن سلامتی افراد جامعه را تهدید کند. بنابراین، بایستی مساله موردنظر و آثار جانبی آن به طور کامل شناخته شده و به عنوان قیود مساله به خوبی مدل‌سازی شوند تا اصلاح الگوی مصرف برق یک مشترک منجر به بروز خسارت و آسیب در حوزه‌های دیگر نشود. در این راستا نیاز است از دانش و تجربه تمامی متخصصین و صاحب‌نظران در حوزه‌های مختلف بهره جست. مساله دیگر این است که آیا می‌توان تصمیمی یکسان برای تمامی مشترکین صنعتی قائل شد یا بایستی با توجه به متنوع بودن شکل مصرف انواع مختلفی از تصمیمات و تعرفه‌ها را اعمال نمود؟ به نظر می‌رسد ممکن است با توجه به متفاوت بودن فرآیند تولید (و بسیاری از دلایل دیگر که حتی احتمال دارد در گام اول ناشناخته باشند)، بهتر باشد برنامه‌های متنوعی با اهداف اصلاح ساعتی، روزانه، هفتگی، ماهیانه یا فصلی مصرف برای انواع مختلف مشترکین در نظر گرفته شود. در این صورت سوال مشابهی می‌تواند شکل گیرد که آیا تصمیمات اتخاذ شده برای هر صنعت ثابت خواهد بود یا لازم است برای هر شرکت در حوزه یک صنعت یکسان نیز متفاوت باشد؟ آیا اساسا از لحاظ قانونی می‌توان تصمیماتی متنوع به تفکیک هر شرکت یا هر صنعت اتخاذ و اجرا نمود؟ در صورت امکان‌پذیر بودن، بعد اجتماعی آن به چه صورت است و آیا از طرف مشترکین مختلف قابل‌قبول خواهد بود یا ممکن است اعتراضاتی به چنین تصمیماتی وجود داشته باشد.

علاوه بر موارد فوق، در نحوه اجرایی نمودن تصمیمات نیز لازم است درک کاملی از مساله وجود داشته باشد. آیا نتایج حاصل از مطالعات در گام اول برای تمامی مشترکین صنعتی اعمال خواهد شد یا اینکه یک یا چند نمونه از صنایع در مرحله اول به عنوان نمونه مطالعاتی انتخاب خواهند شد؟ همچنین نحوه انتخاب صنایع پایلوت بر چه اساسی خواهد بود؟ آیا صنعتی با شدت بار بالا و در نتیجه اثرگذاری بیشتر در رفتار بار صنایع در سطح ملی انتخاب خواهد شد یا صنعتی کوچک‌تر با شدت انرژی کمتر؟ دقت شود در صورت انتخاب صنایع با شدت بار بالا و تاثیرگذار، بررسی اثرگذاری و مشاهده نتایج حاصل به طور محسوسی قابل رصد خواهد بود ولی از طرف دیگر ممکن است این صنایع حساس و حیاتی باشند و در صورت اجرای نامناسب یا هر گونه خطا یا کاستی در انجام تحلیل‌ها اثرگذاری زیادی در ابعاد مختلف

فنی و غیرفنی در جامعه خواهد داشت. در صورت انتخاب صنایع کوچک‌تر، احتمالا حساسیت کمتری در ابعاد فنی و غیرفنی نسبت به نتایج وجود خواهد داشت و اثرگذاری کمتری در گام‌های ابتدایی اجرایی شدن این برنامه‌ها حاصل خواهد شد، ولی در صورت وجود خطایی در تحلیل‌ها یا سیاست‌ها احتمالا آثار مخرب کمتری رخ خواهد داد.

این پروژه گامی ابتدایی در زمینه مدیریت انرژی صنایع و تنظیم تعرفه و نیز شناخت فضای مساله، چالش‌ها و سوالات پیش‌رو است. در ابتدا لازم است شرکت‌ها و صنایع مختلف به خوبی شناخته شوند. در این راستا، ابتدا راه‌کاری مبتنی بر تحلیل داده پیشنهاد شده است که بر اساس نتایج حاصل از آن بتوان دسته‌بندی‌های مناسبی روی شرکت‌ها و صنایع مختلف صورت داد که این امر به تنهایی نیازمند مطالعاتی وسیع و عمیق است. در ادامه، لازم است پروسه‌ها و سیکل‌های کاری هر شرکت یا صنعت نیز جهت تدقیق دسته‌بندی‌ها به‌خوبی شناخته شده و با نتایج و تحلیل‌های پیشین ادغام شود. همچنین، لازم است میزان کشش قیمتی بار نیز به خوبی مدل‌سازی شود. سپس، لازم است به مدل‌سازی مساله با در نظر گرفتن تمامی قیود فنی و غیرفنی پرداخت تا نتیجه چنین مطالعات و تحلیل‌هایی، تصمیماتی راهبردی را شکل دهند که به چه صورت می‌توان تعرفه شرکت‌های مختلف را اصلاح نمود؛ به طوری که الگوی مصرف در راستای اهداف اساسی کشور اصلاح‌شده و پیامدهای احتمالی در حوزه‌های دیگر نیز کاملا تحت کنترل باشد.

همانطور که اشاره شد، این پروژه شروعی بر مطالعات در این حوزه بوده است. بدین‌منظور، ابتدا در فصل اول مفاهیم داده‌کاوی و ضرورت بحث و جایگاه آن در حوزه دانشی و صنعت بیان شد. سپس روند انجام یک فرآیند داده‌کاوی شرح داده شد و برخی ابزار و روش‌های موجود نیز معرفی شد. در این بخش، برخی چالش‌ها از قبیل حفظ حریم خصوصی و ملاحظات اخلاقی، در کنار برخی راه حل‌های آن، بررسی شدند. هم‌چنین، نگاهی اجمالی نیز بر چالش‌ها و جهت‌گیری‌های آینده در این حوزه ارائه شد. سپس، در فصل دوم، با استفاده از برخی از داده‌های در دسترسی از بار ساعتی صنایع، کاربرد داده‌کاوی در مدیریت انرژی صنایع بررسی شد و برخی اقدامات انجام شده در این پروژه نیز تشریح شد. ملاحظه شد که در نتیجه این اقدامات و تحلیل‌های ابتدایی نیز می‌توان برخی دسته‌بندی‌ها را روی صنایع مختلف اعمال نمود که به چند نمونه نیز اشاره شد. در پایان، نتیجه‌گیری شده که لازم است چنین رویه‌هایی مورد بررسی بیشتر قرار گیرند و تحلیل‌های جامع‌تری نیز انجام شوند تا در نهایت بتوان پس از انجام مدل‌سازی مساله و با در نظر داشتن توابع هدف و قیود موجود (حوزه‌های مختلف فنی و غیرفنی) بتوان تصمیماتی راهبردی و کارا را ارائه داد. در نتیجه چنین مطالعات جامع و عمیق و تحلیل‌های کیفی و کمی گسترده می‌توان تصمیماتی دقیق در راستای اصلاح تعرفه بخش صنایع پیشنهاد داد و به مرحله اجرا درآورد.

مراجع

[1] Han, J., Kamber, M., & Pei, J. (2011). Data Mining: Concepts and Techniques. Morgan Kaufmann.

[2] Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2016). Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann.

[3] Tan, P. N., Steinbach, M., & Kumar, V. (2014). Introduction to Data Mining. Pearson.

[4] Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer.

[5] Aggarwal, C. C. (2015). Data Mining: The Textbook. Springer.

[6] Berry, M. J., & Linoff, G. (2011). Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management. Wiley.

[7] Han, J., Kamber, M., & Pei, J. (2011). Data Mining: Concepts and Techniques. Morgan Kaufmann.

[8] Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2016). Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann.

[9] Tan, P. N., Steinbach, M., & Kumar, V. (2014). Introduction to Data Mining. Pearson.

[10] Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer.

[11] Aggarwal, C. C. (2015). Data Mining: The Textbook. Springer.

[12] Berry, M. J., & Linoff, G. (2011). Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management. Wiley.

[13] Chen, M., Mao, S., & Liu, Y. (2014). Big Data: A Survey. Mobile Networks and Applications, 19(2), 171-209.

[14] Hand, D. J., Mannila, H., & Smyth, P. (2001). Principles of Data Mining. MIT Press.

[15] Mitchell, T. M. (1997). Machine Learning. McGraw-Hill.

[16] Jain, A. K., Murty, M. N., & Flynn, P. J. (1999). Data Clustering: A Review. ACM Computing Surveys, 31(3), 264-323.

[17] Berry, M. J., & Linoff, G. (2004). Data Mining Techniques: For Marketing, Sales, and Customer Support. Wiley.

[18] Fayyad, U. M., Piatetsky-Shapiro, G., & Smyth, P. (1996). From Data Mining to Knowledge Discovery in Databases. AI Magazine, 17(3), 37-54.

[19] Chen, H., Chiang, R. H., & Storey, V. C. (2012). Business Intelligence and Analytics: From Big Data to Big Impact. MIS Quarterly, 36(4), 1165-1188.

[20] Kim, J., & Mueller, C. (2015). Introduction to Data Mining and Its Applications. CRC Press.

[21] Hand, D., Mannila, H., & Smyth, P. (2001). Principles of Data Mining. MIT Press.

[22] Vijay Kotu, "Predictive Analytics and Data Mining", Concepts and Practice with Rapidminer, Chapter 2: Data Mining Process, 2015 , Pages 17-36, DOI: http://dx.doi.org/10.1016/B978-0-12-801460-8.00002-1.

[23] Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). CRISP-DM 1.0: Step-by-step Data Mining guide. SPSS Inc. Retrieved from ftp://ftp. software.

ibm.com/software/analytics/spss/support/Modeler/Documentation/14/UserManual/ CRISP-DM.pdf.
[24] SAS Institute (2013). Getting Started with SAS Enterprise Miner 12.3. (pp. 1–3). Shearer, C. (2000). The CRISP-DM Model: The New Blueprint for Data Mining. Journal of Data Warehousing, 5(4), 13–22.
[25] Kubiak, T., & Benbow, D. W. (2005). The Certified Six Sigma Black Belt Handbook. Milwaukee, WI: ASQ, Quality Press.
[26] Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From Data Mining to Knowledge Discovery in Databases. AI MAGAZINE, 17(3), 37–54.
[27] Piatetsky, G. (2014). KDnuggets 15th Annual Analytics, Data Mining, Data Science Software Poll: RapidMiner Continues To Lead. Retrieved August 01, 2014, from http://www.kdnuggets.com/2014/06/kdnuggets-annual-software-poll-rapidminer-continues-lead.html.
[28] Shearer, C. (2000). The CRISP-DM Model: The New Blueprint for Data Mining. Journal of Data Warehousing, 5(4), 13–22.
[29] Lidwell, W., Holden, K., & Butler, J. (2010). Universal Principles of Design, Revised and Updated: 125 ways to Enhance Usability, Influence Perception, Increase Appeal, Make Better Design Decisions, and Teach Through Design. Beverly, MA: Rockport Publishers.
[30] Tan, P.-N., Steinbach, M., & Kumar, V. (2005). Introduction to Data Mining. Journal of School Psychology, 19, 51–56. http://dx.doi.org/10.1016/0022-4405(81)90007-8.
[31] Weisstein, E. W. (2013). Least Squares Fitting. MathWorld - Wolfram Research, Inc. Retrieved from http://mathworld.wolfram.com/LeastSquaresFitting.html.
[32] Guazzelli, A., Zeller, M., Lin, W., & Williams, G. (2009). PMML: An Open Standard for Sharing Models. The R Journal, 1(1), 60–65.
[33] M.A.P. Chamikara, P. Bertok, D. Liu, S. Camtepe, I. Khalil, "Efficient Privacy Preservation of Big Data for Accurate Data Mining", Information Sciences (2019), doi: https://doi.org/10.1016/j.ins.2019.05.053.
[34] Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2016). Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann. URL: https://books.google.com.au/books?isbn=0128043571.
[35] Chamikara, M. A. P., Bertok, P., Liu, D., Camtepe, S., & Khalil, I. (2018). Efficient data perturbation for privacy preserving and accurate data stream mining. Pervasive and Mobile Computing, 48 , 1–19. doi:https://doi.org/10.1016/j.pmcj.2018.05.003.
[36] Helbing, D., Brockmann, D., Chadefaux, T., Donnay, K., Blanke, U., Woolley-Meza, O., Moussaid, M., Johansson, A., Krause, J., Schutte, S. et al. (2015). Saving human lives: What complexity science and information systems can contribute. Journal of statistical physics, 158 , 735–781. doi:https://doi.org/10.1007/s10955-014-1024-9.
[37] Jalili, M., & Perc, M. (2017). Information cascades in complex networks. Journal of Complex Networks, 5 , 665–693. doi:https://doi.org/10.1093/comnet/cnx019.
[38] Capraro, V., & Perc, M. (2018). Grand challenges in social physics: In pursuit of moral behavior. Frontiers in Physics, 6 , 107. doi:https://doi.org/10.3389/fphy.2018.00107.
[39] Vatsalan, D., Sehili, Z., Christen, P., & Rahm, E. (2017). Privacy-preserving record linkage for big data: Current approaches and research challenges. In Handbook of Big Data Technologies (pp. 851–895). Springer. doi:https://doi.org/10.1007/978-3-319-49340-4_25.
[40] Wen, Y., Liu, J., Dou, W., Xu, X., Cao, B., & Chen, J. (2018). Scheduling workflows with privacy protection constraints for big data applications on cloud. Future Generation Computer Systems, . doi:https://doi.org/10.1016/j.future.2018.03.028.
[41] Chen, K., & Liu, L. (2005). A random rotation perturbation approach to privacy preserving data classification. The Ohio Center of Excellence in Knowledge-Enabled Computing, . URL: https://corescholar.libraries.wright.edu/knoesis/916/.
[42] Chen, K., & Liu, L. (2011). Geometric data perturbation for privacy preserving outsourced data mining. Knowledge and Information Systems, 29 , 657–695. doi:https://doi.org/10.1007/s10115-010-0362-4.

[43] Kerschbaum, F., & H¨arterich, M. (2017). Searchable encryption to reduce encryption degradation in adjustably encrypted databases. In IFIP Annual Conference on Data and Applications Security and Privacy (pp. 325–336). Springer. doi:https://doi.org/10.1007/978-3-319-61176-1_18.
[44] Li, P., Li, J., Huang, Z., Gao, C.-Z., Chen, W.-B., & Chen, K. (2017). Privacy-preserving outsourced classification in cloud computing. Cluster Computing, (pp. 1–10.). doi:https://doi.org/10.1007/s10586-017-0849-9.
[45] Dwork, C., Roth, A. et al. (2014). The algorithmic foundations of differential privacy. Foundations and Trends in Theoretical Computer Science, 9 , 211–407. doi:http://dx.doi.org/10.1561/0400000042.
[46] Tang, J., Korolova, A., Bai, X., Wang, X., & Wang, X. (2017). Privacy loss in apple's implementation of differential privacy on macos 10.12. arXiv preprint arXiv:1709.02753, URL:https://arxiv.org/abs/1709.02753.
[47] Kairouz, P., Oh, S., & Viswanath, P. (2014). Extremal mechanisms for local differential privacy. In Advances in neural information processing systems (pp. 2879–2887). URL: http://papers.nips.cc/paper/5392-extremal-mechanisms-for-local-differential-privacy.
[48] Buccafurri, F., Lax, G., Nicolazzo, S., & Nocera, A. (2016). A threat to friendship privacy in facebook. In International Conference on Availability, Reliability, and Security (pp. 96–105). Springer. doi:https://doi.org/10.1007/978-3-319-45507-5_7.
[49] Wei, Z., Wu, Y., Yang, Y., Yan, Z., Pei, Q., Xie, Y., & Weng, J. (2018). Autoprivacy: automatic privacy protection and tagging suggestion for mobile social photo. Computers & Security, doi:https://doi.org/10.1016/j.cose.2017.12.002.
[50] Cuzzocrea, A. (2015). Privacy-preserving big data management: The case of olap. Big Data: Algorithms, Analytics, and Applications, (pp. 301–326;). URL: https://books.google.com.au/books?isbn=1482240564.
[51] Kieseberg, P., & Weippl, E. (2018). Security challenges in cyber-physical production systems. In International Conference on Software Quality (pp. 3–16). Springer. doi:https://doi.org/10.1007/978-3-319-71440-0_1.
[52] Torra, V. (2017). Data Privacy: Foundations, New Developments and the Big Data Challenge. Springer. doi:https://doi.org/10.1007/978-3-319-57358-8.
[53] Aldeen, Y. A. A. S., Salleh, M., & Razzaque, M. A. (2015). A comprehensive review on privacy preserving data mining. SpringerPlus, 4 , 694. doi:https://doi.org/10.1186/s40064-015-1481-x.
[54] Muralidhar, K., Parsa, R., & Sarathy, R. (1999). A general additive data perturbation method for database security. management science, 45 , 1399–1415. doi:https://doi.org/10.1287/mnsc.45.10.1399.
[55] Okkalioglu, B. D., Okkalioglu, M., Koc, M., & Polat, H. (2015). A survey: deriving private information from perturbed data. Artificial Intelligence Review, 44 , 547–569. doi:https://doi.org/10.1007/s10462-015-9439-5.
[56] Hasan, A., Jiang, Q., Luo, J., Li, C., & Chen, L. (2016). An effective value swapping method for privacy preserving data publishing. Security and Communication Networks, 9 , 3219–3228. doi:https://doi.org/10.1002/sec.1527.
[57] Torra, V. (2017). Fuzzy microaggregation for the transparency principle. Journal of Applied Logic, 23 , 70–80. doi:https://doi.org/10.1016/j.jal.2016.11.007.
[58] Aggarwal, C. C., & Yu, P. S. (2004). A condensation approach to privacy preserving data mining. In EDBT (pp. 183–199). Springer volume 4. doi:https://doi.org/10.1007/978-3-540-24741-8_12.
[59] Chen, K., & Liu, L. (2005). A random rotation perturbation approach to privacy preserving data classification. The Ohio Center of Excellence in Knowledge-Enabled Computing, URL: 675 https://corescholar.libraries.wright.edu/knoesis/916/.
[60] Liu, K., Kargupta, H., & Ryan, J. (2006). Random projection-based multiplicative data perturbation for privacy preserving distributed data mining. IEEE Transactions on knowledge and Data Engineering, 18 , 92–106. doi:https://doi.org/10.1109/TKDE.2006.14.
[61] Agrawal, R., & Srikant, R. (2000). Privacy-preserving data mining. In ACM Sigmod Record (pp.439–450). ACM volume 29. doi:https://doi.org/10.1145/335191.335438.

[62] Dimitrios Papakyriakou, Ioannis S. Barbounakis, “Data Mining Methods: A Review”, International Journal of Computer Applications (0975 – 8887), Volume 183, No. 48, January 2022.
[63] Trupti, A. Kumbhare, and Santosh, V. Chobe, (2014). An Overview of Association Rule Mining Algorithms, International Journal of Computer Science and Information Technologies, Vol.5(1), pp. 927-930.
[64] Sudhir, M. Gorade, Ankit Deo and Pritesh Purohit, (2017). A Study of Some Data Mining Classification Techniques. International Research Journal of Engineering and Technology. Vol. 4, Issue. 4, pp. 31123115.
[65] J. Han, M. Kamber and J. Pei, J (2010). Data Mining Concepts and Techniques (3rd ed.) University ofIllinois.Chapter 8, pp. 99-117.
[66] Huh, Myung-Hoe, & Lee, Yonggoo. (2006). ―LMS and LTS-type Alternatives to Classical Principal Component Analysis‖. Communications for Statistical Applications and Methods, 13 (2), 233–241. https://doi.org/10.5351/CKSS.2006.13.2.233.
[67] R. Agrawal and R. Srikant., (March 1995).―Mining Sequential Patterns‖. In Proc. of the 11th Int'l Conference on Data Engineering, Taipei, Taiwan.
[68] Thabet Slimani, and Amor Lazzez. (2013). ―Sequential Mining: Patterns and Algorithms Analysis‖, International Journal of Computer and Electronics Research, Volume 2, Issue 5, pp 639-647.
[69] Kum, H.-C., Chang, J. H., & Wang, W. (2006). ―Sequential Pattern Mining in MultiDatabases via Multiple Alignment‖. Data Min. Knowl. Discov., 12(23), 151-180.