www.collegeprozheh.ir کالج پروژه

۱-۳- ابعاد تحقیق

۱-۱-۳- بعد پژوهش از نظر نتایج یا پیامد

پژوهش کاربردی تلاشی است برای یافتن پاسخی برای حل یک معضل و مشکل عملی، که در دنیای واقعی وجود دارد. در این گونه پژوهش‌ها از نتایج پژوهش‌های بنیادی به منظور بهبود و به کمال رساندن رفتارها، ابزارها، وسایل، تولیدات، ساختارها و الگوهای مورد استفاده‌ی جوامع انسانی استفاده می‌نماید[۲۱]. از این رو این تحقیق از بعد نتایج یا پیامد، پژوهش کاربردی می‌باشد.

۲-۱-۳- بعد پژوهش از نظر منطق اجرا

طرح‌های پژوهشی از نظر منطق اجرا، در عمل به دو گروه اساسی پژوهش‌های قیاسی[1] و استقرایی[2] تقسیم می‌گردند. از نظر منطقی در پژوهش‌های قیاسی، استدلال از کل به جزء است و در پژوهش‌های استقرایی برعکس، حرکت از جزء به کل است[۲۱]. بنابراین تحقیق حاضر، از بعد منطق اجرا، از نوع پژوهش‌های استقرایی می‌باشد.

۳-۱-۳- بعد پژوهش از نظر زمان اجرا

پژوهش مقطعی برای توصیف ویژگی‌ها، نگرش‌ها، عقاید، اندیشه و رفتار افراد در یک جامعه در مقطع معینی از زمان به کار می‌رود. به علاوه، به منظور گردآوری داده‌ها در باره‌ی یک یا چند صفت در یک مقطع خاص از زمان از طریق نمونه گیری از جامعه انجام می‌شود. این گونه پژوهش به

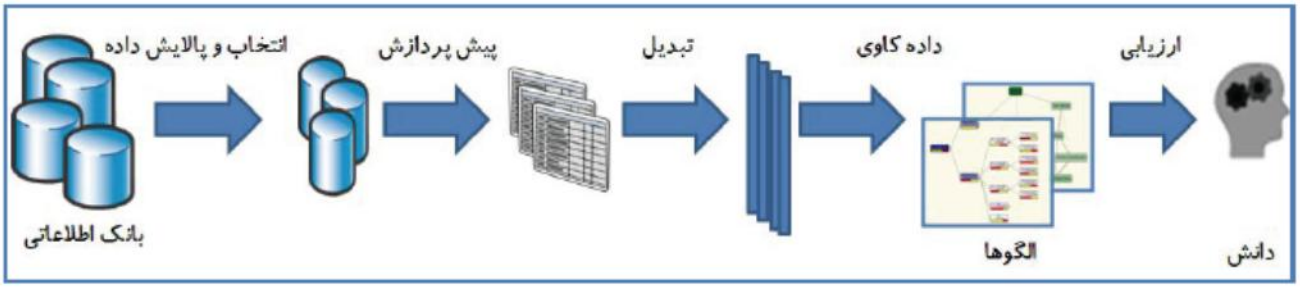
1. Deductive
2. Inductive

www.collegeprozheh.ir کالج پروژه

توصیف جامعه بر اساس یک یا چند متغیر می‌پردازد[۲۱] و این تحقیق از بعد زمان انجام، تحقیق مقطعی به شمار می‌آید.

۲-۳-روش تحقیق

در انجام این تحقیق، جهت خوشه بندی مشتریان شرکت سیمان از فرآیند کشف دانش از پایگاه داده استفاده گردیده است. فرآیند کشف دانش از پایگاه داده، یافتن اطلاعات و الگوهای مفید از داده می‌باشد که باید معتبر، از قبل ناشناخته و بالقوه مفید (برای برنامه های کاربردی معلوم و معین) باشند. الگوهای جالب توجه به کاربر ارائه می‌شوند یا ممکن است به عنوان یک دانش جدید سازمانی ذخیره گردند. این فرآیند یک فرآیند تکرار پذیر و تعاملی می‌باشد و شامل مراحل هفت‌گانه‌ی می‌باشد که در ادامه به تفضیل بحث می‌شوند. فرآیند کشف دانش از پایگاه داده در شکل (۱-۳) آمده است.



شکل ۱-۳: فرآیند کشف دانش از پایگاه داده [۲۲].

۱-۲-۳- پالایش داده (پاک‌سازی داده نادرست)

پاک‌سازی داده، فرآیند تشخیص، اصلاح و حذف خطاهای موجود در داده‌هاست. خطاهای داده شامل داده‌های غلط، ناقص، تکراری، متناقض و یا با ساختار نامناسب هستند. برای بیان این تعریف از عبارات تمیز کردن داده یا پالایش داده هم استفاده می‌شود. ابزارهای پیچیده‌ای با استفاده از

www.collegeprozheh.ir کالج پروژه

الگوریتم‌ها، قوانین و جداول جستجو پاک‌سازی داده را انجام می‌دهند. این ابزارها قابلیت اصلاح خودکار برخی خطاها مانند پیدا کردن و حذف داده ای تکراری را دارند[۵۱].

هدف اصلی پاک‌سازی داده‌ها، از بین بردن ناسازگاری، مقادیر نادرست و سایر کمبود های یکپارچگی داده‌ها از بانک‌های اطلاعاتی موجود است. علاوه بر آن، باعث ایجاد سازگاری بین مجموعه‌های مختلف داده، می‌شود که با یکدیگر ادغام شده‌اند، که این مسئله می‌تواند هدف دیگری برای پاک‌سازی داده باشد. عمل پاک‌سازی، تبدیل داده های موجود در سیستم‌های فعلی را به فرمت و ساختار مورد نیاز برای سیستم‌های جدید تسهیل می‌کند. بدیهی است، افرادی که درگیر پروژه پاک‌سازی داده می‌شوند، لازم است تا با ساختار داده های موجود و همچنین ساختار داده‌ها در سیستم هدف آشنایی کافی داشته باشند. در نتیجه این امکان برای متخصصین سیستم‌های موجود فراهم می‌شود تا خود را با پایگاه داده هدف و ابزار های توسعه و عملکرد این سیستم آشنا سازند. از سوی دیگر توسعه دهندگان سیستم هدف، با ساختار و محتوی سیستم‌های موجود در سازمان آشنا می‌شوند[۵۱].

پاک‌سازی داده، زمانی مورد نیاز است که مشکلات داده ای در سیستم رخ دهد. استراتژی‌های جلوگیری از خطا می‌تواند بسیاری از مشکلات داده ای را کاهش دهد، ولی نمی‌تواند آنان را از بین ببرد. روش‌های مختلفی در فرآیند پاک‌سازی داده مورد استفاده قرار می‌گیرد. در اینجا به مهم‌ترین آنان اشاره می‌شود[۵۱]:

الف) ترکیب: در فرآیند پاک‌سازی داده از ترکیب برای شناسایی خطاهای نحوی استفاده می‌شود. یک ترکیب کننده‌ی گرامرG، برنامه ایست که تشخیص می‌دهد آیا رشته ورودی، متعلق به زبان تعریف شده توسط گرامر هست یا خیر. در زبان‌های برنامه سازی، رشته یک برنامه است؛ و در

www.collegeprozheh.ir کالج پروژه

پاک‌سازی داده، رشته‌ها می‌توانند رکورد و یا مقادیر داده ای باشند. رشته‌هایی که خطای نحوی دارند، باید اصلاح شوند. تعداد خطاهای نحوی موجود در داده بستگی به محدودیت‌های محیطی دارد که داده در آن ذخیره شده است. اگر داده در فایل بدون ساختار ذخیره شده باشد دارای خطاهای نحوی و دامنه‌ای بیشتری خواهد بود[۵۱].

ب) تبدیل داده: تبدیل داده، به عمل نگاشت داده از یک قالب به قالب مورد نظر با استفاده از یک نرم افزار گفته می‌شود. این تبدیل هم الگوی رکورد و هم دامنه مقادیر را تغییر می‌دهد؛ به این ترتیب که ابتدا داده های چند منبع به یک الگوی مشترک که نیازها را به نحو مطلوبی برآورده می‌سازند، تبدیل می‌شوند؛ سپس اصلاح مقادیر در صورتی انجام می‌شود که داده های ورودی با الگوی مشترک مطابقت نداشته باشد و این عدم تطابق باعث شکست تبدیل داده شود. استاندارد سازی و نرمال سازی، تبدیل‌هایی هستند که در سطح نمونه برای از بین بردن ناهنجاری‌های استفاده می‌شوند[۵۱].

ج) اعمال محدودیت‌های جامعیت: این روش، محدودیت‌های جامعیت را پس از انجام تراکنش‌های تغییر داده (شامل حذف، اضافه و بروز رسانی) تأمین و تضمین می‌کند. دو رویکرد مختلف برای این روش، کنترل محدودیت جامعیت و حفظ محدودیت جامعیت است. در روش اول، تراکنش‌هایی که نقض کننده جامعیت هستند، برگشت داده می‌شوند و در روش دوم، تراکنش‌های بروز رسانی و اعمال تغییرات در داده های اصلی شناسایی می‌شوند تا داده‌ها پس از تغییر، هیچ یک از محدودیت‌های جامعیت را نقض نکنند[۵۱].

د) از بین بردن داده های تکراری: روش‌های متفاوتی برای حذف داده های تکراری وجود دارد که در همه این روش‌ها، باید الگوریتمی وجود داشته باشد که تشخیص دهد دو یا چند رکورد،

www.collegeprozheh.ir کالج پروژه

نمایش‌های تکراری از یک موجودیت می‌باشند. برای یک تشخیص کارا، هر رکورد باید با همه رکورد های دیگر مقایسه شود. به عنوان مثال با استفاده از روش همسایگی مرتب شده، می‌توان تعداد مقایسه‌ها را به حداقل رساند. به این ترتیب که رکوردها بر اساس کلیدی مرتب می‌شوند که رکورد های تکراری نزدیک هم قرار گیرند. سپس رکوردهایی که در یک پنجره کوچک شناور قرار دارند، با یکدیگر مقایسه می‌شوند. تشخیص رکورد های تکراری بر اساس قوانین موجود در دانش حوزه مورد مطالعه است[۵۱].

ه) **روش‌های آماری:** روش‌های آماری هم برای بررسی داده و هم برای اصلاح ناهنجاری داده مورد استفاده قرار می‌گیرند. تشخیص و از بین بردن خطاهای پیچیده با استفاده از کنترل و اعمال محدودیت‌های جامعیت امکان پذیر نیست. با تحلیل داده‌ها بر اساس مقادیر میانگین، انحراف معیار و الگوریتم‌های خوشه بندی، اشخاص خبره ممکن است مقادیر داده‌های پیش بینی نشده‌ای را پیدا کنند که نشان دهنده رکورد های نامعتبر است. معمولاً اصلاح این خطاها غیر ممکن است، زیرا صفات جدول دارای مقادیر داده‌ای صحیح هستند. یک راه حل ممکن با استفاده از روش‌های آماری، قرار دادن مقادیر آماری از قبیل میانگین در این صفات ست. ناهنجاری دیگری که به وسیله روش‌های آماری اصلاح می‌شود، مقادیر خالی است که با داده های قابل قبول جایگزین می‌شود. تولید این داده‌ها، نیازمند الگوریتم‌های تولید داده وسیع است[۵۱].

www.collegeprozheh.ir کالج پروژه

۲-۲-۳- یکپارچه سازی و تجمیع داده[1] (ترکیب منابع داده چندگانه)

برای یک فرد متوسط، فناوری اطلاعات یک دنیای اسرارآمیز به حساب می آید که با زبان‌های برنامه نویسی غیرقابل فهم و سخت افزارهای گران قیمت پر شده است. گوش دادن به صحبت های تکنیسین های فناوری اطلاعات مانند این است که تصادفاً به مکالمه ای در یک زبان بیگانه گوش می دهید. اما علیرغم این موانع ظاهراً غیرقابل نفوذ، شناخت دنیای فناوری اطلاعات برای تصمیم گیرندگان در بنگاه های تجاری و سازمان‌های گوناگون از اهمیت حیاتی برخوردار است. یکی از مهم‌ترین تصمیمات فناوری اطلاعات، تجمیع داده‌ها است[۵۲].

در ظاهر، تجمیع داده ها چیزی شبیه به یک ایده ساده به نظر می رسد. از آنجا که بسیاری از سازمان‌ها اطلاعات خود را بر روی بانک‌های اطلاعاتی متعدد نگهداری می‌کنند، به روشی برای بازیابی داده ها از منابع مختلف و هم‌گذاری آنان با یک شیوه یکپارچه نیاز دارند. برای مثال، فرض کنید که یک شرکت الکترونیکی برای ارائه ی یک ابزار موبایل جدید آماده می شود. بخش بازاریابی این شرکت احتمالاً می‌خواهد اطلاعات مشتریان را از یک بانک اطلاعاتی بخش فروش استخراج کرده و آن را با اطلاعات بخش محصول مقایسه نماید تا یک فهرست فروش هدفمند را ایجاد کند. یک سیستم خوب تجمیع داده‌ها به بخش بازاریابی اجازه خواهد داد تا اطلاعات هر دو منبع را با یک شیوه‌ی یکپارچه مشاهده نموده و هر اطلاعاتی که به جستجوی مورد نظر مربوط نمی‌شود را کنار بگذارد [۵۲].

تجمیع داده ها اصولاً بر بانک های اطلاعاتی تمرکز دارد. یک بانک اطلاعاتی، یک مجموعه سازماندهی شده از داده ها است. این مجموعه به یک سیستم فایل شباهت دارد که یک ساختار

1. Data Integration

www.collegeprozheh.ir کالج پروژه

سازماندهی برای فایل ها است تا یافتن، دسترسی و دست کاری آنان آسان باشد. روش های مختلفی برای دسته بندی بانک‌های اطلاعاتی وجود دارد. برای مثال گاهی آنان را بر اساس نوع داده‌هایی که در بانک‌های اطلاعاتی ذخیره می‌شوند، طبقه بندی می‌کنند و یا در روش دیگری به نحوه‌ی سازماندهی داده‌ها در بانک اطلاعاتی توجه می‌شود[۵۲].

۳-۲-۳- انتخاب داده (بازیابی داده های مرتبط با وظایف تحلیل)

برای فرایند داده کاوی داده های مورد نیاز موجود در انبار داده‌ها باید انتخاب شوند. درک این مطلب که برای ارزیابی یک مدل که بعداً برای تست و به‌کارگیری آن مدل بکار می‌رود، موفقیت آمیز باشد، بسیار مهم است در غیر این صورت نتایج درستی حاصل نمی‌گردد[۵۲].

مثلاً انبار داده‌ها شامل انواع مختلف و گوناگونی از داده‌ها است به عنوان مثال در یک پایگاه داده های مربوط به سیستم فروشگاهی، اطلاعاتی در مورد خرید مشتریان، خصوصیات آماری آن‌ها، توزیع کنندگان، مشتریان، حسابداری و ... وجود دارند که همه آنان در داده کاوی مورد نیاز نیستند[۵۲].

٤-۲-۳- تبدیل داده یا مهندسی داده‌ها

زمانی که داده های مورد نیاز از پایگاه داده های موجود در انبار داده‌ها جمع آوری شدند و داده‌های مورد کاوش مشخص گردیدند، لازم است داده‌ها به فرمی که برای تکنیک‌های داده کاوی مناسب باشد، ترجمه گردند و تبدیلات خاصی به روی آنها انجام گردد که شامل حداقل دو مرحله متداول می‌باشد[۵۲]:

www.collegeprozheh.ir کالج پروژه

الف: آشکارسازی (حذف) داده های غیرعادی: داده های غیرعادی یا غیر معمول در حقیقت نتیجه‌ی سنجش خطاها، کد نویسی و ثبت خطاها است. برای این کار یکی از دو راهکار ذیل را باید بکار برد[۵۲]:

- ❖ داده های غیرعادی را تشخیص داد و حذف کرد؛
- ❖ باید روش‌های قوی مدل سازی را به گونه ای توسعه داد که نسبت به این نوع داده‌ها غیر حساس باشند.

ب) ویژگی‌های مقیاس بندی، رمزگذاری و انتخاب: در تبدیل داده‌ها توصیه می‌شود که داده‌ها را جهت تحلیل و بررسی مقیاس بندی و و رمزگذاری کرد؛ مثلاً یک مشخصه با دامنه[۱و۰] و دیگری با دامنه [-100,1000] دارای ارزش مشابهی در تکنیک‌های اعلام شده نیستند که در صورت نادیده گرفتن همین تفاوت در دامنه داده‌ها، روی نتایج نهایی داده کاوی تأثیر خواهند گذاشت[۵۲].

۵-۲-۳- داده کاوی بر پایه‌ی مهندسی الگوریتم و تعیین استراتژی‌های کاوش

تعداد الگوریتم‌های بسیار زیادی که در زیر هر کدام از متدها قرار می‌گیرند، می‌توانند حتی برای افراد حرفه‌ای هم گیج‌کننده باشند. این نکته باید ذکر شود که همه‌ی این الگوریتم‌ها دارای تکنیک‌های پایه‌ای یکسانی هستند. انواع مختلفی از الگوریتم‌های داده‌کاوی وجود دارند. مدل‌های مشهور، از قوانین و درخت‌های تصمیم، رگرسیون و کلاسه‌بندی غیرخطی، متدهای مبتنی بر مثال (شامل متدهای نزدیک‌ترین همسایه و استنتاج بر اساس مورد)، مدل‌های آماری (نظیر شبکه‌های بیز، تابع توزیع احتمال نرمال، تابع توزیع احتمال χ^2 و...) و مدل‌های یادگیری رابطه‌ای (نظیر برنامه‌نویسی منطقی استنتاجی)، استفاده می‌کنند. لازم به ذکر است که هر تکنیک برای انواع مختلفی از مسایل بهتر از سایر

www.collegeprozheh.ir کالج پروژه

تکنیک‌ها عمل می‌کند. برای مثال کلاسه بندی‌های درخت تقسیم برای پیدا کردن ساختار در فضاهای با ابعاد بالا و همچنین در مسایل با داده‌هایی که می‌توانند پیوسته یا طبقه بندی شده باشند، مفید خواهد بود (زیرا متدهای بر پایه‌ی درخت نیاز به فواصل متریک ندارند)؛ در حالی که این درخت‌ها برای مسایلی که نیاز به مرزبندی دقیق بین کلاس‌ها دارند، مفید نمی‌باشد. بنابراین هیچ متد داده کاوی واحدی نمی‌تواند وجود داشته باشد و انتخاب یک الگوریتم مشخص برای کاربردهای ویژه نیاز به مهارت‌های خاصی دارد. در عمل به علت گستردگی، پیچیدگی و حجم بسیار بالای داده‌های موجود برای یک کاربرد خاص و نیاز به بهینه‌سازی آن‌ها، دامنه متدهای مفید محدود شده است. این مرحله یعنی اینکه مفیدتر است که برای یک منظور خاص و بر روی یک سری داده‌های ویژه، چه الگوریتمی اعمال شود که بهترین کارایی را برای ما داشته باشد[٤٥].

٦-٢-٣-اجرای الگوریتم کاوش و ارزیابی الگوها

انتخاب و پیاده سازی مناسب داده کاوی وظیفه اصلی این مرحله است. در عمل چندین مدل به طور همزمان پیاده سازی شده و سپس بهترین آنان انتخاب می‌شود. شاید بتوان به طور خلاصه گفت که مأموریت اصلی کاوش داده‌ها به عهده این گام است [٤٥].

بخش‌های مختلف این گام عبارتند از:

- ❖ انتخاب و استفاده از تکنیک مدل سازی مناسب
- ❖ دست کاری و تنظیم مدل و استفاده از الگوریتم‌ها برای دستیابی به نتایج بهینه
- ❖ در صورت نیاز برگشت به گام پیش پردازش

www.collegeprozheh.ir کالج پروژه

۷-۲-۳- ارائه‌ی دانش

بصری سازی دانش به معنی استفاده از بازنمایی‌های بصری[1] برای انتقال دانش بین حداقل دو انسان است. هدف بصری سازی دانش، بهبود انتقال دانش از طریق شیوه‌های نمایش کامپیوتری یا غیر کامپیوتری است. نمونه‌هایی از این بازنمایی‌های بصری، طراحی‌ها، نمودارها، تصاویر، اشیاء، بصری سازی تعاملی، برنامه های بصری سازی اطلاعات و داستان‌ها هستند. بصری سازی دانش، علاوه بر مفاهیم، به دنبال انتقال بینش، تجربه، تمایلات، ارزش‌ها، انتظارات، دیدگاه‌ها، نظرات و پیش بینی‌ها نیز می‌باشد. برای بصری سازی دانش، شیوه های مختلفی از زمینه های متفاوت مانند ارتباطات بصری[2]، علم ارتباطات[3]، ادراک بصری[4] و مدیریت دانش به کار گرفته می‌شود[۳۳].

به کمک تحلیل‌های بصری می‌توان به توانمندی‌های مختلف در عرصه های گوناگون علمی پی برد. مثلا در زمینه‌های علمی همچون ریاضیات، فیزیک، زیست شناسی، اخترشناسی و ... . حین اجرای تحلیل‌ها، برخی از الگوها به سرعت نمایان می‌شوند. به کمک بصری سازی می‌توان مفاهیم زیادی را در قالب یک صفحه نشان داد. استفاده از این مفاهیم در مدیریت دانش کاربرد فراوان داشته و می‌توان به کمک آن‌ها، تصمیمات سریع و بهینه ای را اتخاذ کرد. یکی از نتایج این امر، بازگشت سریع‌تر سرمایه خواهد بود.

۳-۳- متغیرهای کلیدی

متغیرهای کلیدی به کار رفته در این تحقیق به شرح ذیل می‌باشد:

1. Visual Representations
2. Visual Communication
3. Communication Science
4. Visual Perception

www.collegeprozheh.ir کالج پروژه

- **محل تخلیه‌ی سیمان :** که با توجه به داده‌های شرکت سیمان، شامل استان آذربایجان غربی، آذربایجان شرقی، کردستان، تهران و همچنین برخی از کشورهای همسایه می‌باشد.
- **گروه مشتری:** شامل پیمانکاران، عاملین توزیع، انبوه سازان، سازمان‌های دولتی، سیمانبران، جواز تاسیس، تاجر(کشورهای همسایه) و سایر می‌باشد.
- **نوع سیمان:** که سیمان پرتلند تیپ یک(۳۲۵-۱ ، ٤۲٥-۱) و سیمان پوزولانی می‌باشد.
- **بسته بندی :** که به صورت پاکتی و یا فله ای می‌باشد.
- **مقدار تحویلی سیمان :** که مقدار تحویل داده شده‌ی سیمان به تن را در چهار زیر گروه، مشخص می‌کند.

٤-۳- جامعه آماری

جامعه‌ی آماری این تحقیق، کلیه‌ی مشتریان شرکت سیمان ارومیه می‌باشد که اطلاعات آنان از طریق پایگاه داده‌ی قسمت فروش به ثبت رسیده است.

٥-۳- روش نمونه گیری

در این پژوهش اطلاعات همه‌ی مشتریان شرکت سیمان ارومیه در بازه‌ی زمانی سال ۱۳۸۸ و سه ماهه‌ی اول سال ۱۳۹۰ بعنوان نمونه از جامعه‌ی آماری مورد بحث، جهت انجام خوشه بندی در نظر گرفته شده است. دلیل در نظر گرفتن این نمونه، موجود بودن اطلاعات مربوط به مشتریان شرکت سیمان ارومیه در برهه‌ی مذکور و ارائه‌ی آن توسط قسمت فروش شرکت می‌باشد و از روش نمونه گیری خاصی برای انتخاب رهه‌ی مذکور استفاده نشده است.

www.collegeprozheh.ir کالج پروژه

٦-٣- ابزار گردآوری داده‌ها

در این پژوهش برای تدوین کلیات، مبانی نظری و پیشینه‌ی تجربی تحقیق از روش کتابخانه‌ای و استفاده از منابع اطلاعاتی، مانند: اینترنت، مقالات، کتاب‌های مربوط به زمینه‌ی تحقیق استفاده شده است. علاوه بر این برای پاسخ به سوالات تحقیق، گردآوری داده‌ها به صورت اسنادی از پایگاه داده‌ی قسمت فروش شرکت سیمان ارومیه، صورت گرفته است.

٧-٣- روش‌های تجزیه و تحلیل اطلاعات

روش‌های آماری قابل استفاده در این تحقیق آمار توصیفی و تحلیلی است. از آمار توصیفی برای تنظیم جدول فراوانی و رسم نمودارها استفاده می‌شود و برای تجزیه و تحلیل تکنیک خوشه بندی داده کاوی نیز از الگوریتم شبکه‌ی عصبی کوهونن استفاده شده است.