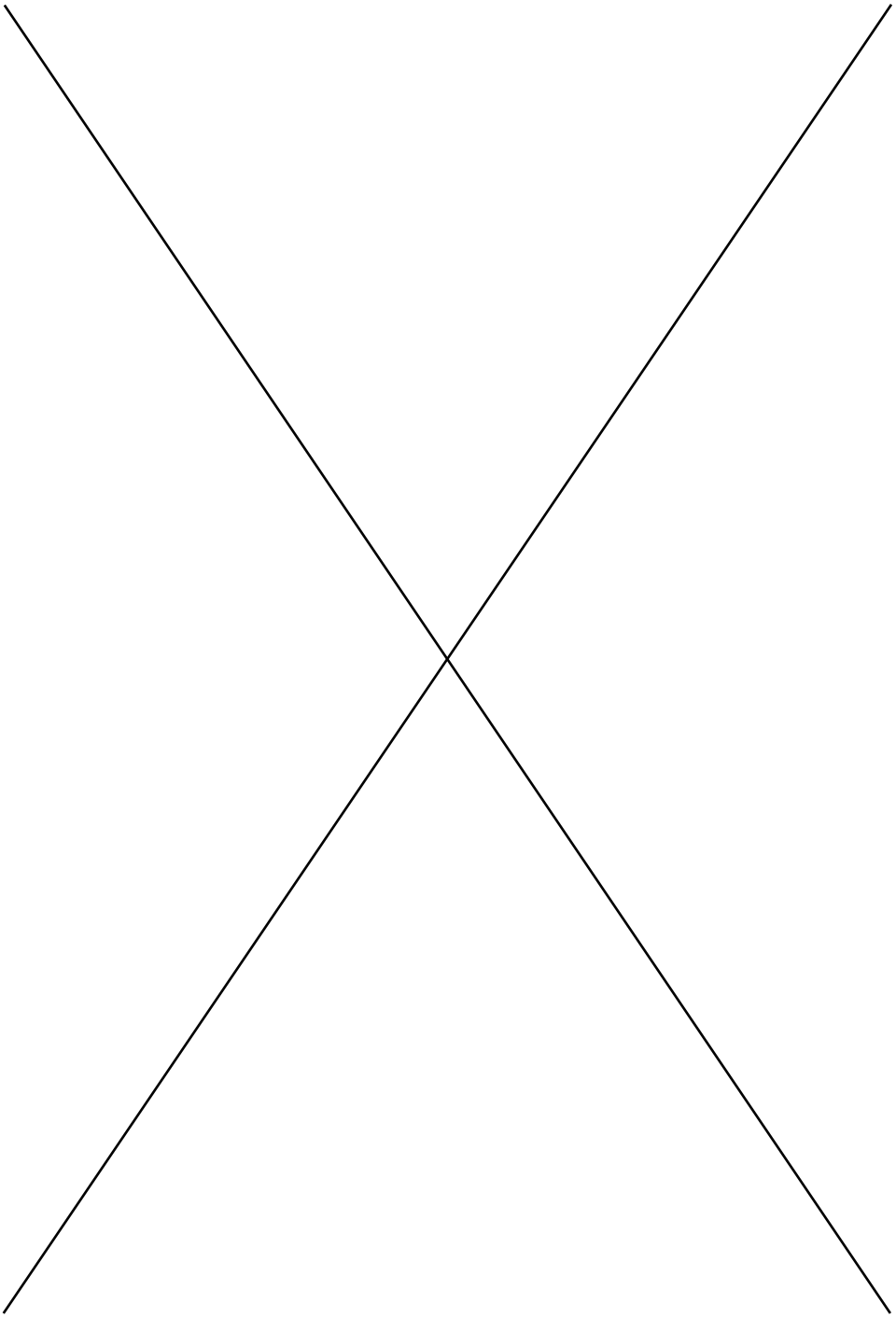
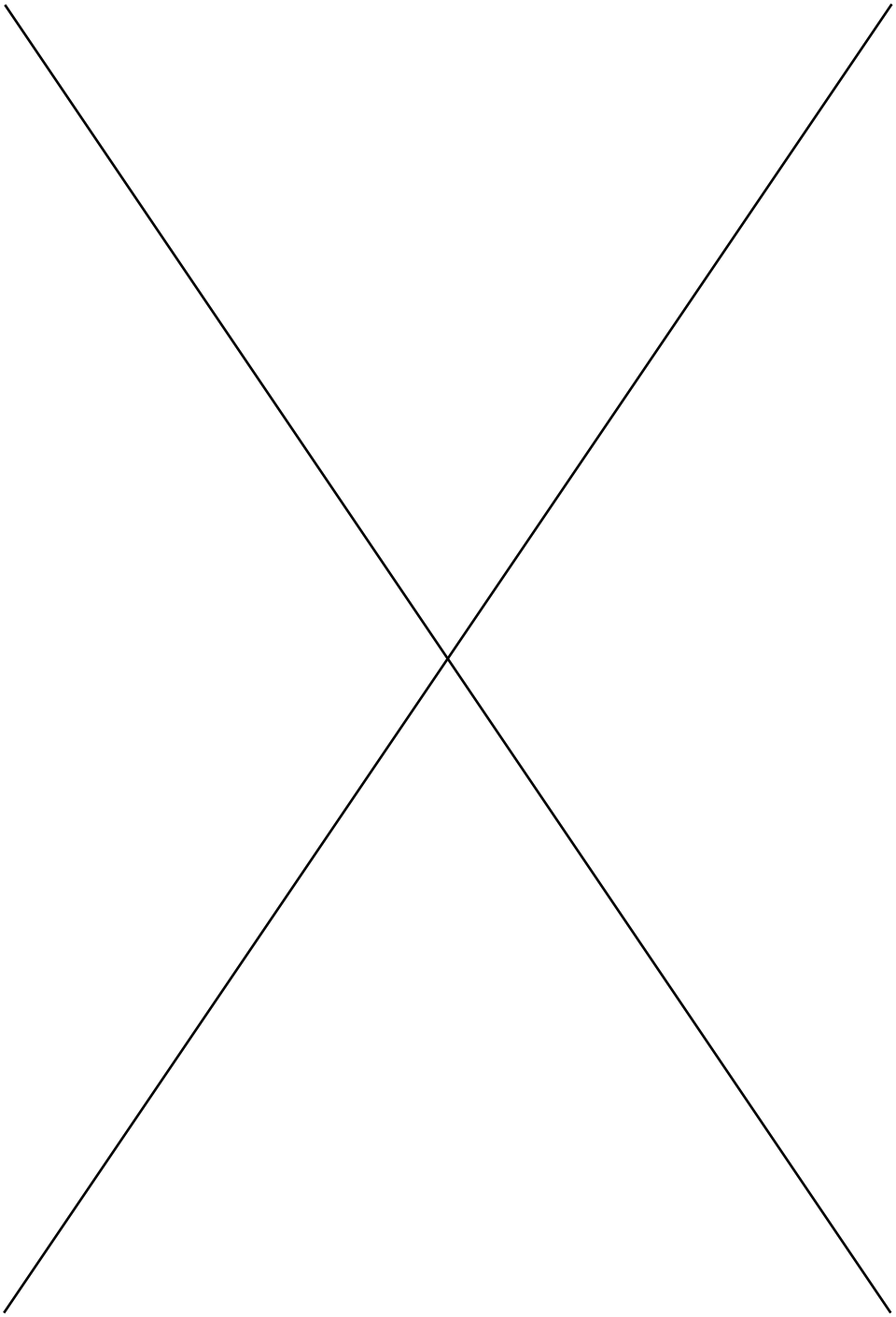
فرآیند استاندارد فرآصنعتی داده‌کاوی

راهنمای گام به گام داده‌کاوی

مترجم: محمد روزبه

دای‌کهه
DAYCHE





در مـدت فعالیـت من ، بعنوان پژوهش‌گر ارشـد و رئیس دپارتمان داده‌کاوی در شـرکت بنز آلمان، این دپارتمان در چند پروژه داده‌کاوی نیز شرکت داشت که توسط بازار مشترک اروپا پشتیبانی مالی می‌شد.

CRISP-DM بنظر من مهم‌ترین آن‌ها بود. ایده این پروژه که در انجام آن علاوه بر دپارتمان داده‌کاوی شـرکت بنز سه شـرکت دیگر از جمله شـرکت SPSS نیز سهیم بودند در سـال ۱۹۹۶ میلادی شکل گرفت و انجام آن سـه سـال بطول انجامید. این شـرکت که چند سال پیش توسـط IBM خریداری شد هماهنگی پروژه را نیز بعهده داشت.

امروزه که بیش از چهارده سال از پایان این پروژه می‌گذرد، بدون اغراق می‌توان گفت که CRISP-DM مهم‌ترین استاندارد موجود جهت انجام پروژه‌های کاربردی داده‌کاوی در جهان می‌باشد.

بسـیاری از پروژه های کاربردی داده‌کاوی که در مدت حضور من در دپارتمان داده‌کاوی شرکت بنز انجام گرفت، با اسـتفاده از این اسـتاندارد بود. افزون بر این، تاجایی‌که من اطلاع دارم، در ایران نیز تاکنون این اسـتاندارد در بسـیاری از رساله های کارشناسی ارشـد که در زمینه داده‌کاوی نوشته شده اند، مورد اسـتفاده قرار گرفته اسـت. همچنین مدت زیادی است که این استاندارد در ابزار داده‌کاوی SPSS Modeler نیـز تلفیـق گردیـده و در دسـترس کاربـران ایـن ابـزار قـرار دارد.

برای من بعنوان شخصی که نقش کوچکی در بوجود آمدن این استاندارد داشته، بسیار مسرت بخش اسـت که اکنون ترجمه توصیف این اسـتاندارد به زبان فارسـی، به همت گروه داده‌کاوی دایکه و بویژه با کوشـش مسـتمر آقای محمد روزبه، در اختیار پژوهش‌گران و کاربران فارسـی زبان داده‌کاوی قرار می‌گیرد. ضمن ابراز قدردانی از زحمات ایشان برای تمام خوانندگان و بویژه برای دانشجویان عزیز آرزوی بهره‌گیری فراوان از این اثر را دارم.

غلامرضا نخعی زاده
استاد دانشگاه کارلسروهه آلمان
مدیر اسبق دپارتمان داده‌کاوی شرکت بنز

مقدمه مترجم

تصمیم به ترجمه و چاپ این استاندارد تقریباً همزمان با شروع فعالیت‌های ما در قالب دپارتمان داده‌کاوی در زمستان ۱۳۸۶ بوده است. کمتر از ۵ ماه بعد نسخه اولیه ترجمه با همکاری اعضای اولیه دپارتمان آماده شد. ولی نوپا بودن این دانش در سطح کشور و نداشتن تجربه اجرایی مناسب اعضای تیم در آن مقطع، باعث شد نسخه اولیه صرفاً ترجمه ای از متن اصلی بدون درک درستی از مفاهیم بکار رفته در آن باشد. بنابراین تصمیم به چاپ آن معلق شد تا در وقتی مناسب با کیفیتی شایسته محتوای پر اهمیت آن ارائه گردد.

طی سال های بعد متن اولیه بارها دستخوش تغییر شد و با توجه به تجربیات روزافزون که در قالب آموزش، مشاوره و اجرای پروژه های متعدد بدست آمده بود مورد بازنگری قرار می گرفت. ولی گذر زمان و مشغله های کاری باعث کمرنگ شدن تصمیم اولیه برای چاپ شده بود؛ تا اینکه برگزاری مستمر دوره های آموزشی «پیاده سازی فرآیند داده‌کاوی در ابزارهای تخصصی داده‌کاوی» از جمله (Clementine SPSS Modeler) در دو سال اخیر و مشاوره های متعددی که برای پیاده سازی و اجرای پروژه های داده‌کاوی به مدیران و کارشناسان شرکت ها، سازمان ها و همچنین دانشجویان مقاطع تکمیلی ارائه می‌شد، ضرروت چاپ این استاندارد را بیش از پیش نمایان ساخت.

با توجه به ماهیت استاندارد و همچنین طیف مخاطبان آن، سعی شده تا در کنار پایبندی به ساختار اصلی متن و همچنین نوع نگارش آن، متن ترجمه شده بصورت روان و شناور بوده تا کمترین میزان خستگی را در مطالعه آن ایجاد نماید.

برای رسیدن به این هدف، اساتید محترم و دوستان گرانقدری کمک نموده اند که صمیمانه از همه آن‌ها تشکر می کنم. جناب پروفسور نخعی زاده که راهنمایی ها و همراهی های ایشان مسلماً برای من فراموش نشدنیست. همکاران خوبم در هسته اولیه دپارتمان داده‌کاوی، خانم زهرا زجاجی و آقای عباسعلی ناطقی که در تهیه نسخه اولیه این ترجمه از اعضای اصلی بوده اند. همچنین دوستان با ارزش و عزیزم خانم زهرا ذوالقدر وآقایان احسان اژدری، خشایار مقدم، صادق طولابی فر، پویان رمضانی و سایر دوستانی که در بازبینی متن، ویرایش و همچنین حمایت‌هایشان برای تسریع در اتمام این کار نقش بسزایی داشتند.

با تمام این صحبت ها معتقدم متن حاضر خالی از اشکال نیست و بهبود کیفیت آن مستلزم ارائه پیشنهادات و انتقادات خوانندگان عزیز می باشد.

محمد روزبه

مرداد ۱۳۹۲

فهرست مطـالب

پیشگفتار

CRISP-DM محصول تلاش‌های سه شرکت پر سابقه بازار نوپا و ناپخته داده کاوی، در اواخر سال ۱۹۹۶ است. دایملر کرایسلر (که بعداً دایملر بنز شد) پیش از دیگران در فرآیند های صنعتی و امور تجاری‌اش از داده کاوی استفاده کرده بود. SPSS (یا ISL) از سال ۱۹۹۰ خدماتی بر اساس داده کاوی ارائه داده بود و در سال ۱۹۹۴، نخستین نرم افزار تجاری داده کاوی (کلمنتاین) را معرفی کرد. NCR نیز در راستای ارائه ارزش افزوده بیشتر برای مشتریان انبار داده ترادیتای خود، تیمی را متشکل از مشاوران داده کاوی و متخصصان صنعت، برای رفع نیازمندی های آنان، به کار گرفته بود.

گرایش های بازار آن زمان حاکی از گسترش روزافزون داده کاوی بود؛ گسترشی که هیجان برانگیز بود و البته نگران کننده. همه ما با همراهی هم دستاوردهای داده کاوی را توسعه داده بودیم اما آیا ما درست عمل کرده بودیم؟ آیا همه کسانی که برای اولین بار از داده کاوی استفاده می‌کردند می‌بایستی از طریق سعی و خطا (مانند ما!) آن را بیاموزند؟ و ما که ارائه دهنده آن بودیم چگونه باید مشتریان خود را به این باور می رساندیم که داده کاوی می تواند راهنمای امور تجاری آن‌ها باشد؟

سپس به این نتیجه رسیدیم یک مدل فرآیند استاندارد که به جایی اختصاص نداشته باشد و به صورت مجانی در دسترس باشد می تواند این موضوع را به ما و همه کاربران نشان دهد.

یک سال بعد کنسرسیومی تجاری تشکیل دادیم (با عنوان فرآیند استاندارد فرا صنعتی داده کاوی[۱]) که با تأمین بودجه خود از جانب کمیسیون اروپا همان اهداف اولیه ما را دنبال می‌کرد. از آنجا که نبایست CRISP-DM از نظر صنعت و ابزار و کاربرد متعلق به گروه خاصی می بود، دریافتیم که برای ادامه کار می‌بایست از گروهی به اندازه کاربران و سایر سرویس دهندگان (مانند ارائه دهندگان انبار داده‌ها و مدیران مشاوری که علاقمند به داده کاوی بودند) انرژی بگیریم. این کار را با ایجاد گروه علاقمندان ویژه CRISP-DM (SIG)[۲] انجام دادیم. به این صورت که از طریق تبلیغات رادیویی از علاقمندان دعوت کردیم در کارگاه یک روزه ای که در آمستردام برگزار شد شرکت کنند. ما می‌خواستیم ایده های خود را با آن‌ها در میان بگذاریم و از آن‌ها بخواهیم ضمن معرفی خود به بحث و تبادل نظر در مورد پیشبرد CRISP-DM بپردازند.

---

۱ Cross Industrial Process for Data Mining (CRISP-DM)

۲ Special Interest Group

در روز برگزاری کارگاه اعضای گروه نگران بودند نکند کسی در کارگاه شرکت نکند؟ و اگر کسی شرکت کرد برای این آمده باشد که بگوید نیازی ضروری به یک فرآیند استاندارد احساس نمی‌کند. یا این که تلاش‌های ما آنقدر پراکنده بوده که هرگونه ایده استاندارد کردن مشتی تخیلات غیر عملی باشد. نتیجه کارگاه فراتر از انتظار ما بود. آنچه بیش از همه جلب نظر می‌کرد سه مورد زیر بود:

- جمعیت حاضرین دو برابر انتظار اولیه ما بود.
- همگی متفقالقول بر این باور بودند که صنعت امروز نیاز مبرم به فرآیند استاندارد دارد که بایستی هم-اکنون انجام پذیرد.
- هر بار که یکی از حضار دیدگاه خود را در مورد داده کاوی با توجه به تجربیات پروژه اش ارائه می‌کرد، علی رغم تفاوت های ظاهری به ویژه در تأیین حد و مرز فازها و مجموعه لغات مورد استفاده، وجوه قابل توجه مشترک در دیدگاه های آن‌ها وضوح بیشتری پیدا می‌کرد.

با خاتمه کارگاه اطمینان کامل داشتیم که با تکیه بر امیدبخشی و نقدهای SIG خواهیم توانست یک مدل فرآیند استاندارد برای جامعه داده کاوی ارائه دهیم.

در طی دو سال و نیم پس از آن تلاش ما معطوف به توسعه و رفع اشکالات CRISP-DM بود. تلاش‌هایمان را روی پروژه‌های واقعی و عظیمی مانند داده کاوی برای مرسدس بنز و شریک بخش بیمه مان OHRA ادامه داده و طی آن روی یکپارچه سازی CRISP-DM با ابزارهای تجاری کار کردیم. SIG بسیار ارزشمند از آب درآمده بود و اعضای آن به بیش از ۲۰۰ نفر رسیده بود و کارگاه هایی در لندن، نیویورک و بروکسل برگزار کرده بود.

پس از پایان یافتن بخش تخصیص اعتبار پروژه در اواسط سال ۱۹۹۹ یک پیش نویس با کیفیت از مدل فرآیند تهیه کردیم و برای بهبود هر چه بیشتر آن تلاش کردیم که نتیجه آن تلاش‌ها CRISP-DM 1.0 بود. در حین انجام کار می‌دانستیم که فرآیند مدل CRISP-DM هنوز در حال طی کردن مراحل ساخت امی‌باشد و امتحان خود را تنها روی پروژه‌های انگشت شمار پس داده است. در سال گذشته دایملر کرایسلر فرصت پیدا کرد تا CRISP-DM را به طور گسترده به کار گیرد. همچنین گروه های خدماتی حرفه ای مانند SPSS و CRISP-DM، NCR را پذیرفته اند، و از آن برای حل مسائل تجاری و صنعتی مشتریانشان استفاده می کنند. در این مدت سرویس‌های خدماتی زیادی که CRISP-DM را پذیرفته اند با تحلیلگران مختلف مکرراً به آن به عنوان یک استاندارد غیر رسمی صنعت مراجعه کرده اند. و در کنار همه اینها افزایش آگاهی مشتریان از اهمیت CRISP-DM نکته ایست که به آسانی نمی‌توان از کنار آن گذشت (CRISP-DM هم اکنون مرتباً به مناقصه گذاشته می‌شود).

بر این باوریم که ابتکار عمل ما کاملاً به اثبات رسیده است و اگرچه گسترش ها و بهسازی هایی که بعدها صورت می‌گیرد مطلوب و البته اجتناب ناپذیر است اما نسخه CRISP-DM-1.0 آنقدر ارزشمند

هست که نشر و گسترش داده شود.

CRISP-DM نـه از طریـق نظریه پردازی و کار آکادمیک بر روی اصول فنی شـکل گرفته اسـت و نه به وسـیله گروه نخبه ای که پشـت درهای بسـته در مورد آن تحقیق کرده باشند. این دو مسیر به توسعه روش‌هایی منجر شـده اسـت که پیش از این امتحان شـده اند اما به ندرت به یک اسـتاندارد مورد اقبال عموم که عملی و موفق باشـد منجر شـده اند. دلیل موفقیت CRISP-DM اسـتوار بودن بر مبنای تجربه های عملی و واقعی مرسومی است که پروژه‌های داده کاوی با استفاده از آن‌ها انجام شده اند. از این رو ما بی‌نهایت مرهون کاربرانی هستیم که در طول پروژه ایده ها و تلاش‌های خود را در اختیار ما قرار داده اند.

کنسرسیوم CRISP-DM
آگوست ۲۰۰۰

فصل اول
مقدمه

روش شناسی CRISP-DM

جداسازی سلسله مراتبی

روش داده کاوی CRISP-DM در شمای کلی با مدلی که شامل چهار سطح از وظایف است نشان داده می‌شود. این چهار سطح عبارتند از: فاز[۳]، وظیفه عمومی[۴]، وظیفه تخصصی[۵] و نمونه فرآیند[۶].

در واقع این مدل نشـان دهنده فرآیندی اسـت که داده کاوی از طریق آن صورت می گیرد. در سـطح نخسـت، فرآیند شـامل چند فاز است؛ و هر فاز دربرگیرنده چندین وظیفه عمومی (سطح دوم). سطح دوم به این دلیل عمومی نامیده می‌شود که باید شامل همه حالات ممکن داده کاوی باشد. دو ویژگی مهم هر سطح عمومی «کامل بودن» و «پایداری» است. «کامل بودن» موجب می‌شود این وظیفه دربردارنده همه قسمت های فرآیند داده کاوی و همه کاربردهای ممکن آن باشد؛ و «پایداری» قابلیت توسعه یافتن آن را در مواجهه با موارد پیش بینی نشده (مانند تکنیک های نوین مدلسازی) تضمین می‌کند.

سطح سوم شامل وظایف تخصصی ست. این وظایف نحوه انجام وظایف عمومی را در شرایط مختلف شرح می دهند. به عنوان مثال اگر وظیفه عمومی «پاکسازی» باشد، وظیفه تخصصی (که توضیح دهنده نحـوه اجرای وظیفه عمومی در موقعیت های مختلف اسـت) اجرای پاکسـازی در مـورد مقادیر عددی یا مقادیر گسسته را توضیح خواهد داد.

اجرا شـدن وظایف و فازها به صورت مراحل مجزا و با ترتیب مشـخص حالتی ایده آل اسـت. در عمل ترتیب اجرای بسـیاری از وظایف با حالت ایده‌آل تفاوت دارد و در موارد زیادی لازم اسـت مجدداً به مراحل پیشین برگشته و عملیات خاصی را تکرار کنیم. با این توضیحات باید این نکته را هم خاطرنشان کنیم که مدل حاضر درصدد پوشش دادن همه مسیرهای موجود در فرآیند داده کاوی نیست، چون این امر مستلزم توصیف مدلی پیچیده خواهد بود.

سـطح چهارم نمونه فرآیند نام دارد و شامل ثبت دقیق عملیات، تصمیم ها و نتایج یک مسئله واقعی داده کاوی اسـت. نمونـه فرآیندها بر اسـاس وظایفی که در سـطوح بالاتر تعریف شـده اند سـازمان دهی می‌شوند و شامل جزئیات عملیات یک مسئله واقعی هستند.

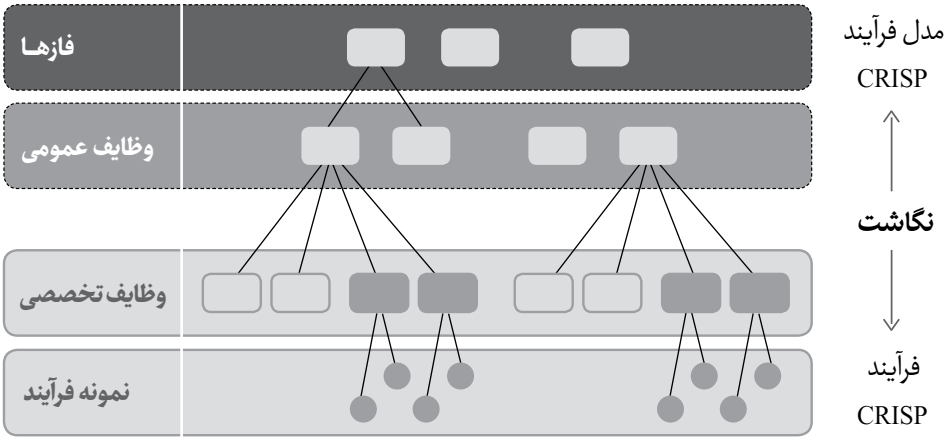
---

۳ Phase

۴ Generic Task

۵ Specialized Task

۶ Process Instance



شکل ۱: تقسیم بندی چهارسطحی روش شناسی CRISP-DM

مدل مرجع[۷] و راهنمای کاربر[۸]

در روش شناسی[۹] CRISP-DM مدل مرجع و راهنمای کاربر با هم تمایز دارند. مدل مرجع شمایی کلی از فازها، وظایف وخروجی آن‌هاست و عملیاتی را که در یک پروژه داده کاوی صورت می گیرد توضیح می‌دهد. اما راهنمای کاربر نکات و راهنمایی های جزئی تری در مورد هر فاز و وظایف مربوط به آن در اختیار می گذارد و روش انجام پروژه داده کاوی را شرح می‌دهد. این کتاب مدل مرجع و راهنمای کاربر را به صورتی کلی پوشش می‌دهد.

نگاشت مدل های عمومی به مدل های اختصاصی

چهارچوب داده کاوی[۱۰]

چهارچوب داده کاوی، نگاشت بین سطوح عمومی و اختصاصی را به عهده دارد. چهار بعد چهارچوب داده کاوی عبارتند از:

- **دامنه کاربردی**، که حوزه خاصی است که داده کاوی در آن رخ می‌دهد.
- **نوع مسئله داده کاوی**، که بیان کننده طبقه بندی خاصی از نوع موضوعات پروژه داده کاوی است.

---

۷ Reference Model
۸ User Guide
۹ Methodology
۱۰ Data Mining Context

- **جنبه های تکنیکی**، که در بر گیرنده نکات خاصی است که در یک پروژه داده کاوی صورت می‌گیرند.
- **ابزار و تکنیک ها**، که مشخص کننده ابزارها و تکنیک های مورد استفاده در پروژه داده کاوی است.

در جدول ۱ چهار بعد داده کاوی همراه با مثال شرح داده شده اند.

هـر چهارچوب خاص با نسـبت دادن مقادیر واقعی به یک یا چند بعد از ابعـاد مذکور به وجود می‌آید. برای مثال یک پروژه داده کاوی که مسئله آن از نوع پیش بینی است، یک چهارچوب داده کاوی خاص اسـت. هر چه مقادیر مشـخص شده بیشتری به ابعاد داده کاوی نسـبت داده شود چهارچوب داده کاوی روشن تر و محکم تر خواهد بود.

	چهارچوب داده‌کاوی			
ابعـــاد	دامنه کاربرد	نوع مسئله داده‌کـاوی	جنبه‌های تکنیکی	ابزارهـــا و تکنیک‌ها
مثال‌ها	مدل پاسخ	توصیف و خلاصه سازی	مقادیر گمشده	کلمنتاین
	پیش‌بینی رویگردانی مشتریان	خوشه‌بندی	مقادیر پرت	MineSet
	...	توصیف مفاهیم		درخت‌های تصمیم
		طبقه‌بندی		
		پیش‌بینی		
		تحلیل وابستگی‌ها		

جدول ۱: ابعاد چهارچوب داده کاوی همراه با مثال

نگاشت[۱۱] با چهارچوب ها

در CRISP-DM نگاشت های عمومی و تخصصی از هم متمایزند:

نگاشت برای وضعیت کنونی:

اگر مدل کلی فرآیند فقط برای یک پروژه داده کاوی خاص در نظر گرفته شده باشد و هدف ما تطبیق وظایف عمومی و توصیفاتشان با خواسته ها و نیازهای آن پروژه خاص باشد، آنگاه پای یک انطباق یگانه در میان است که (به احتمال زیاد) فقط یک بار مورد استفاده قرار می گیرد.

نگاشت برای آینده:

هر گاه یک مدل را به صورت سیستماتیک و بر اساس یک چهارچوب از پیش تعریف شده تخصصی کنیم (به عبارت دیگر تجربیات یک پروژه را در راستای اختصاصی سازی مدل، تحلیل و جمع بندی کنیم) تا بعدها از آن در چهارچوب های دیگر استفاده کنیم، یک مدل ویژه CRISP-DM ایجاد کرده ایم.

این که کدام انطباق مفید تر خواهد بود، به چهارچوب های خاص داده کاوی و نیازهای شرکت بستگی دارد.

چگونگی نگاشت

استراتژی پایه برای نگاشت مدل کلی به تخصصی برای هر دو نگاشت یکسان است:

- چهارچوب خاص خود را تحلیل کنید.
- جزئیات غیر قابل اعمال در چهارچوب مورد نظر را حذف کنید.
- جزئیات مورد نظر خود را اضافه کنید.
- بخش های کلی را بر اساس چهارچوب مورد نظر تنظیم کنید.
- در صورت امکان نام قسمت های کلی را با نام های تازه ای که القا کننده معانی گویاتری در چهارچوب‌اند، عوض کنید.

توضیح فصل ها

مطالب

مدل فرآیند CRISP-DM در پنج فصل مختلف تنظیم شده است.

---

۱۱ Mapping

- فصـل اول مقدمه ای بر روش CRISP-DM اسـت که راهنمایی های کلی برای نگاشـت مدل عمومی فرآیند به مدل تخصصی فرآیند را فراهم آورده است.
- فصل دوم مدل مرجع CRISP-DM و اجزای آن از قبیل فازها، وظایف عمومی و خروجی ها را شرح می‌دهد.
- فصل سوم راهنمای کاربر CRISP-DM را توضیح می‌دهد و شامل توصیف کامل مباحث فازها، وظایف عمومی و خروجی هاست. در این فصل همچنین انجام یک پروژه داده کاوی به صورت دقیق شرح داده شده است.
- فصل چهارم درباره گزارش هایی ست که در حین انجام پروژه و پس از آن بایستی نوشته شوند و طرح کلی این گزارش ها را ارائه می‌دهد. همچنین جدول متقاطع خروجی ها و وظایف نمایش داده شده است.
- فصل پنجم پیوست است که شامل توصیف اصطلاحات و واژگان است و خصوصیات انواع مسائل داده کاوی را شرح می‌دهد.

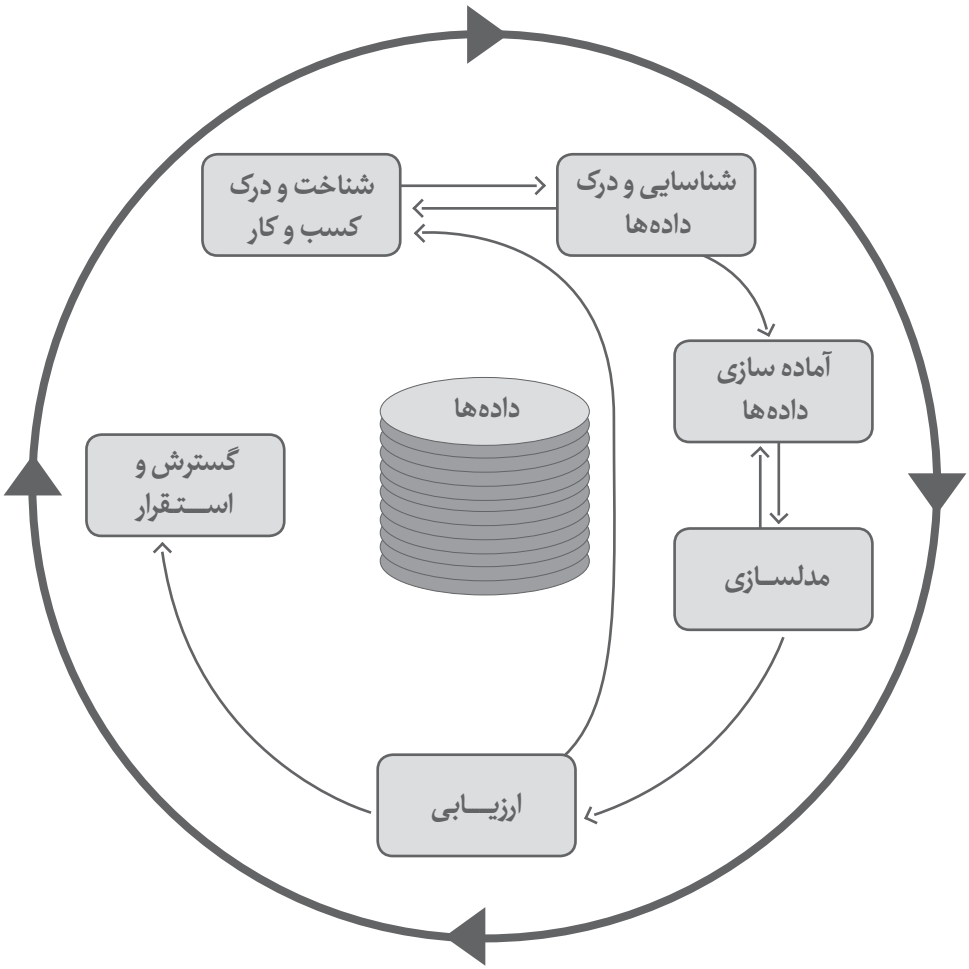
اهداف

نکات زیر امکان استفاده هر چه بهتر از این کتاب را فراهم می‌کند:

- اگـر نخسـتین باری اسـت کـه مـدل CRISP-DM را مطالعه می کنیـد، مطالعه کتـاب را از فصل اول (مقدمه) شروع کنید تا درک مناسب تری از مفاهیم این روش و ارتباط بین این مفاهیم به دست آورید. در این صورت در پروژه‌های بعد از مطالعه مجدد آن معاف خواهید بود!
- اگر دسـتیابی هر چه سـریع‌تر به یک شـمای کلی از مدل CRISP-DM مد نظر شماست، به فصل دوم (مـدل مرجـع CRISP-DM) مراجعـه کنید. همچنین اگر قصد دارید به سـرعت یک پـروژه داده کاوی را شروع کنید برای آشنایی مقدماتی با راهنمای کاربر مطالعه این فصل را از دست ندهید.
- اگـر برای انجام پـروژه داده کاوی به راهنمایی های دقیق و کامل نیازمندید به فصل سـوم (راهنمای کاربـر CRISP-DM) مراجعـه کنیـد. این فصـل مهم‌ترین و با ارزشـترین فصل این کتاب اسـت (پیش از مطالعه این فصل بایستی دو فصل نخست را مطالعه کرده باشید).
- اگر در حال نوشـتن گزارش های داده کاوی هسـتید به فصل چهارم مراجعه کنید. در صورت تمایل به تهیه گزارش ها در حین انجام پروژه، از مطالب دو فصل ۳ و ۴ به صورت همزمان استفاده کنید.
- پیوست پایان کتاب می تواند برای به دست آوردن اطلاعاتی اولیه در مورد داده کاوی مفید واقع شود. اگر هنوز تخصص چندانی در زمینه داده کاوی ندارید برای پیدا کردن معانی و مفاهیم واژه های تخصصی از این فصل استفاده کنید.

فصل دوم
مدل مرجع
CRISP-DM

مدل مرجع CRISP-DM (شکل ۲) شمایی اجمالی از مراحل مختلف پروژه‌های داده کاوی است. این مدل شامل فازهای یک پروژه، وظایف مربوط به آن‌ها و روابط بین این وظایف است. بنابراین تعیین دقیق همه روابط بین وظایف غیر ممکن می‌باشد. اساساً وجود ارتباط بین هر کدام از وظایف داده کاوی، به هدف، زمینه، علایق کاربر و اهمیت داده‌ها بستگی دارد.



شکل ۲: فازهای مدل مرجع CRISP-DM

مراحل مختلف یک پروژه داده کاوی شـامل ۶ فاز اسـت. این فازها در شـکل ۲ نمایش داده شده اند. البته توالی اجرای فازها از پیش تعیین شده نیست و معمولاً لازم است به برخی فازها مجدداً مراجعه شود. لـزوم انجـام یا عدم انجام این عمل بسـتگی به نتیجه هر فاز دارد و اینکه کدام فـاز یا کدام وظیفه خاص از یک فاز، تعیین کننده گام بعدی خواهد بود. پیکان های نمایش داده شـده در شـکل ۲ مهم‌ترین و پر کاربرد ترین ارتباطات بین فازها را نشان می‌دهد.

دایره بیرونی در شـکل ۲ نشـان دهنده طبیعت چرخشـی داده کاوی است. در واقع داده کاوی چیزی بیش از یک راهکار تعمیم یافته برای حل مسئله نیست. مطالبی که در حین فرآیند داده کاوی و از طریق راهکارهای تعمیم یافته آموخته می‌شـوند بیشـتر تمرکز خود را به سـؤالات کسـب و کار معطوف می کنند. بر همین اسـاس است که بیشـترین بهره برداری در فرآیندهای داده کاوی از تجربیات پیشین صورت می گیرد.

در بخش های زیر هر کدام از فازها را به طور خلاصه مرور می کنیم.

شناخت و درک کسب و کار[۱۲]

نخسـتین فاز بر به دسـت آوردن درک مناسـبی از اهداف و ملزومات پروژه از منظر کسـب و کار تمرکز می‌کنـد و از آن برای تعریف یک مسـئله داده کاوی اسـتفاده می‌شـود تا طرحی ابتدایی برای رسـیدن به اهداف مسئله فراهم گردد.

شناسایی و درک داده ها[۱۳]

فاز شناسایی و درک داده‌ها با جمع آوری مقدماتی داده‌ها آغاز شده و با فعالیت هایی به منظور آشنایی بـا داده‌هـا برای تعیین کیفیت آن‌ها ادامه می یابد تا به بینش اولیـه ای از داده‌ها یا فرضیه هایی در مورد زیرمجموعه های مستتر در داده‌ها بینجامد.

آماده سازی داده ها[۱۴]

فاز آماده سـازی داده‌ها شامل همه فعالیت هایی اسـت که مجموعه داده‌های نهایی(داده‌هایی که در اختیار ابزار مدلسـازی قرار می گیرند) را از داده‌های اولیه ایجاد می‌کند. معمولاً فاز آماده سـازی داده‌ها به زمان بیشـتری نسـبت به سـایر فازها نیاز دارد و البته هیچ راهکار از پیش تعیین شده ای برای وظایف ایـن فـاز وجود ندارد. وظایف این فاز عبارتند از: جدول بنـدی، انتخاب و گزینش رکورد و صفت، تبدیل و پاکسازی داده‌ها برای مورد استفاده قرار گرفتن به وسیله ابزارهای مدلسازی.

---

۱۲ Business Understanding

۱۳ Data Understanding

۱۴ Data Preparation

مدلسازی[۱۵]

در ایـن فاز تکنیـک های مختلف داده کاوی انتخاب و اسـتفاده می‌شـوند و پارامترهـای مدل برای رسیدن به مقادیر بهینه تنظیم می‌شوند. به عنوان مثال برای یک مسئله داده کاوی چندین نوع تکنیک وجود دارد. برخی تکنیک ها به اعمال تغییرات خاصی در مورد فرم داده‌ها نیازمندند و این موضوع ایجاب می‌کند مجدداً به فاز آماده سازی داده‌ها رجوع شود.

ارزیابی[۱۶]

در این مرحله مدل سـاخته شـده از نظر تحلیل داده‌ها به کیفیت مطلوبی رسـیده است. قبل از ادامه مسـیر و گسـترش و استقرار نهایی مدل، ارزیابی کاملی از مدل و بررسـی مراحل ساخت آن ضروری است تا در سایه آن، اطمینان از دستیابی مناسب به اهداف فعالیت تجاری حاصل شود. یک نکته کلیدی در این مرحله شناسـایی فعالیت های تجاری مهمی اسـت که به صورت مناسـبی بررسی نشده اند. در پایان این فاز باید شرایط مناسب برای تصمیم گیری در مورد قابل استفاده بودن نتایج داده کاوی فراهم شده باشد.

گسترش و استقرار[۱۷]

سـاختن مدل پایان پروژه نیسـت. حتی اگر کسـب اطلاعات بیشـتر از داده‌ها هدف مدلسـازی باشد، اطلاعات به دست آمده به سازماندهی و ارائه شدن نیازمندند به نحوی که مشتری بتواند از آن‌ها استفاده کند.

شکل ۳ شمایی کلی از فازها با وظایف عمومی (تیره) و خروجی ها را (رنگ خاکستری) نشان می‌دهد. در بخش های بعد، هر یک از وظایف و خروجی ها با جزئیات بیشتری توضیح داده شده اند.

---

۱۵ Modeling

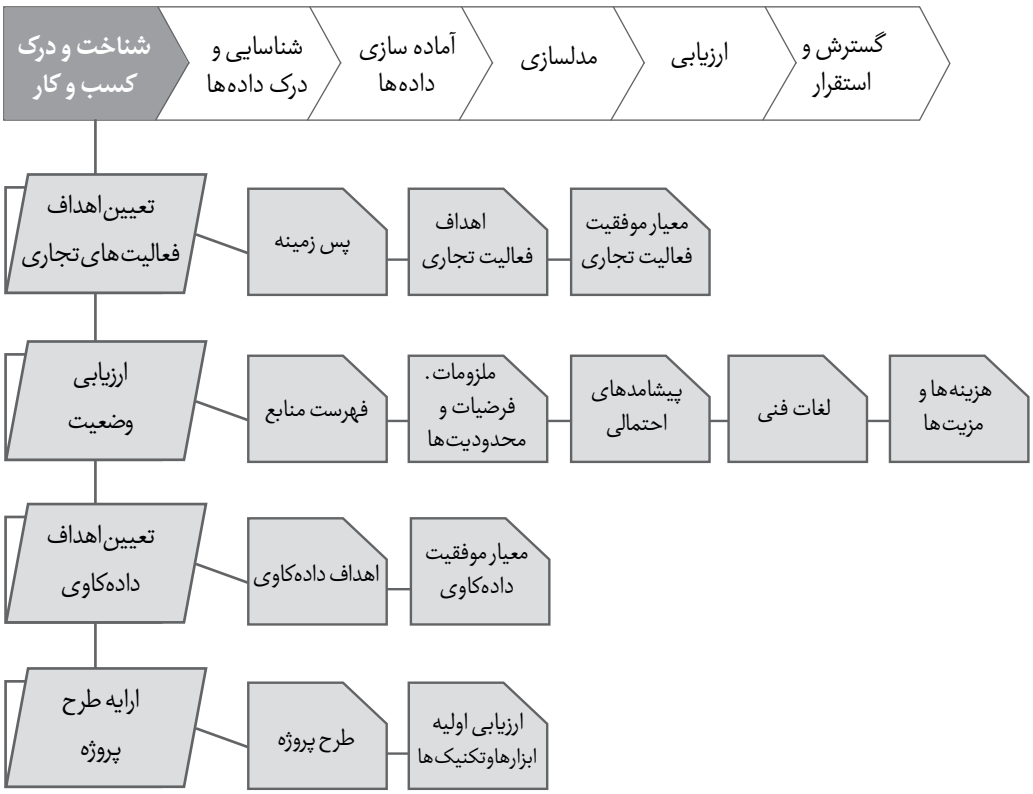
۱۶ Evaluation

۱۷ Deployment

شناخت و درک کسب و کار	**تعیین اهداف فعالیت تجاری** (پس زمینه، اهداف فعالیت تجاری، معیار موفقیت فعالیت تجاری) **ارزیابی وضعیت** (فهرست منابع، ملزومات، فرضیات و محدودیت‌ها، پیشامدهای احتمالی، لغات فنی، هزینه‌ها و مزیت‌ها) **تعیین اهداف داده کاوی** (اهداف داده کاوی، معیار موفقیت داده‌کاوی) **ارائه طرح پروژه**(طرح پروژه، ارزیابی اولیه ابزارها و تکنیک‌ها)
شناسایی و درک داده‌ها	**جمع‌آوری اولیه داده‌ها** (گزارش جمع‌آوری اولیه داده‌ها) **توصیف داده‌ها** (گزارش توصیف داده‌ها) **کاوش داده‌ها** (گزارش کاوش داده‌ها) **بررسی کیفیت داده‌ها** (گزارش بررسی کیفیت داده‌ها)
آماده سازی داده‌ها	**انتخاب داده‌ها** (دلیل انتخاب داده‌ها/ عدم انتخاب داده‌ها) **پاکسازی داده‌ها** (گزارش پاکسازی داده ها) **ساخت داده‌ها** (صفت‌های اشتقاقی، تولید رکورد) **یکپارچه سازی داده‌ها** (ادغام داده‌ها) **فرمت داده‌ها** (فرمت بندی مجدد داده‌ها) **مجموعه داده‌ها** (توصیف مجموعه داده‌ها)
مدلسازی	**انتخاب تکنیک مدلسازی** (تکنیک‌های مدلسازی، فرضیات مدلسازی) **ایجاد طرح آزمون** (طرح آزمون) **ساخت مدل** (تنظیم پارامترها، مدل‌ها، توصیف مدل‌ها) **ارزیابی مدل** (ارزیابی مدل، بازبینی تنظیمات پارامترها)
ارزیابی	**ارزیابی نتایج** (ارزیابی نتایج داده‌کاوی با توجه به معیارهای موفقیت تجاری، مدل‌های تاییدشده) **مرور فرآیند** (مرور فرآیند) **تعیین مراحل بعدی** (فهرست راهکارهای ممکن، تصمیم گیری)
گسترش و استقرار	**طراحی برای گسترش و استقرار** (طرح گسترش و استقرار) **طراحی برای نظارت و نگهداری** (طرح نظارت و نگهداری) **تهیه گزارش نهایی** (گزارش نهایی، ارائه نهایی) **بازبینی پروژه** (مستندسازی تجربیات)

شکل ۳: وظایف عمومی به همراه نتایج در مدل مرجع CRISP-DM

۱. شناخت و درک کسب و کار



شکل ۴: شناخت و درک کسب و کار

۱-۱ تعیین اهداف فعالیت تجاری

وظیفه: تعیین اهداف فعالیت تجاری

یـک تحلیلگر داده در نخسـتین گام باید بتواند با مد نظر قرار دادن چشـم انـداز فعالیت تجاری درک کاملی از خواسـته های واقعی مشـتری حاصل کند. مشـتری ها غالباً چندین هدف موازی و محدود دارند که باید در توازنی منطقی قرار گیرند. هدف تحلیلگر در آغاز آنسـت که عوامل مؤثر بر خروجی های پروژه را شناسایی کند. بی دقتی در این مرحله ممکن است ما را متحمل هزینه های زیادی کند که به خاطر اصرار ما در یافتن جوابی درست برای سؤالی نادرست ایجاد می‌شوند.

خروجی ها:

پس زمینه[۱۸]

ثبت اطلاعات مربوط به فعالیت های تجاری شرکت در آغاز پروژه

اهداف فعالیت تجاری

بـا توجـه به چشـم انـداز فعالیت تجاری هدف اصلی مشـتری را شـرح دهیـد. علاوه بر هـدف اصلی فعالیـت تجـاری سـؤالات دیگـری وجـود دارد کـه مشـتری مایـل اسـت مـورد بحـث قـرار دهـد. برای مثـال هـدف اصلـی یـک فعالیت تجـاری ممکن اسـت حفـظ مشـتری از طریق پیـش بینـی زمانی که او بـه همـکاری با شـرکت رقیـب تمایل پیـدا می‌کند باشـد. در ایـن صورت سـؤالات مربوطه مـی توانند اینگونـه طـرح شـوند:«چگونه کانـال اصلـی ارتباط با مشـتریان یـک بانـک (از قبیـل ATM، حضور در شـعبه یـا از طریـق اینترنـت) اثرگـذاری بیشـتری خواهد داشـت، چـه آن‌هـا بمانند یـا بروند». یـا اینکه «آیا با کاهش کارمزد سیستم ATM تعداد مشتریان ارزشمندی که ما را ترک خواهند کرد، به طور معنا داری کاهش خواهد یافت؟»

معیار موفقیت فعالیت تجاری

از منظر فعالیت تجاری، معیاری برای موفقیت خروجی ها تعیین کنید. این معیار می‌تواند کاملاً ملموس و واقعی باشد، مانند« کاهش مشتریانی که شما را ترک میکنند و به رقیب می پیوندند»، و هم می‌تواند کلی و ذهنی باشد، مثلاً« کسب یک بینش مفید از روابط». در مورد دوم باید مشخص شود ارزیابی ذهنی از جانب چه کسی صورت می‌گیرد.

---

۱۸ Background

۲-۱ ارزیابی وضعیت[۱۹]

وظیفه: ارزیابی وضعیت

این وظیفه جزئیات بیشتری از اطلاعات به دست آمده در مورد همه منابع، محدودیت ها و فرضیات است و همچنین شامل اطلاعات لازم از سایر عواملی که در تعیین هدف داده کاوی و طراحی پروژه باید مورد توجه قرار گیرند. در وظیفه قبلی هدف دستیابی هر چه سریع‌تر به صورت مسأله یک وضعیت بود. ولی هدف این وظیفه دقت هر چه بیشتر در جزئیات است.

خروجی ها:

فهرست منابع

این فهرست موارد زیر را در بر می گیرد:

پرسنل (متخصص کسب و کار، متخصص داده‌ها، پشتیبانی فنی، تیم داده کاوی)، داده‌ها (استخراج داده‌های ثبت شده، دسترسی به انبار داده‌های مؤثر یا داده‌های عملیاتی)، منابع محاسباتی (سخت افزارها) و نرم افزار (ابزارهای داده کاوی، دیگر نرم افزارهای مرتبط)

ملزومات، فرضیات و محدودیت ها

ملزومات پروژه عبارتند از: برنامه زمانی اجرا، قابل درک بودن، امنیت، کیفیت نتایج و همچنین رسیدگی حقوقی. به عنوان بخشی از خروجی مطمئن شوید اجازه استفاده از داده‌ها را دارید.

فهرستی از فرضیاتی که در حین پروژه ساخته می‌شوند تهیه کنید. این فهرست هم می‌تواند شامل فرضیاتی باشد که در حین داده کاوی رسیدگی می‌شوند هم شامل فرضیاتی غیر قابل بررسی در مورد فعالیت تجاری که در صورت مواجهه با آن‌ها پروژه متوقف می‌شود.

همچنین فهرستی در مورد محدودیت های پروژه تهیه کنید. این فهرست می تواند هم شامل محدودیت هایی در مورد در دسترس بودن منابع باشد هم شامل محدودیت های فنی. (مانند اندازه داده‌هایی که به صورت مؤثر در مدلسازی مورد استفاده قرار می گیرند)

پیشامدهای احتمالی

فهرستی از خطرات و حوادثی که ممکن است به تأخیر یا ناکارامدی در روند پروژه منجر شوند تهیه کنید. همچنین فهرستی از اقدامات مناسب برای مواجهه با هر کدام از آن پیشامدها تهیه کنید.

لغات فنی

واژه نامه ای از لغات فنی مربوط به پروژه تهیه کنید. این واژه نامه می تواند شامل دو بخش باشد:

۱) واژه نامه ای در مورد کلمات تخصصی فعالیت تجاری که تشکیل دهنده بخشی از درک فعالیت تجاری است و برای انجام پروژه مفید خواهد بود.

---

۱۹ Assess Situation

۲) واژه نامه ای در مورد لغات تخصصی داده کاوی که با مثال هایی مرتبط با فعالیت تجاری توضیح داده شده است.

هزینه ها و مزیت ها

برای پروژه یک تجزیه و تحلیل هزینه-فایده صورت دهید که هزینه های پروژه را با سودآوری آن (در صورت موفقیت فعالیت تجاری) مقایسه کند. مقایسه تا حد ممکن باید اختصاصی باشد، مثلاً استفاده از پیش بینی های مالی در یک وضعیت تجاری خاص.

۱-۳ تعیین اهداف داده کاوی

وظیفه: تعیین اهداف داده کاوی

اهداف فعالیت تجاری، هدف پروژه را با اصطلاحات خاص فعالیت تجاری و اهداف داده کاوی، هدف را به زبان فنی و تخصصی داده کاوی توضیح می‌دهد. به طور مثال یک هدف فعالیت تجاری می تواند «افزایش فروش کاتالوگ به مشتریان باشد». هدف داده کاوی متناظر می تواند «پیش بینی میزان خرید یک محصول از جانب مشتری، با توجه به خریدهای او در بیش از سه سال گذشته، اطلاعات جمعیت شناختی (سن، درآمد، شهر و غیره) و قیمت آن محصول» باشد.

خروجی ها:

اهداف داده کاوی

خروجی های منتخب پروژه که قابلیت دستیابی به اهداف فعالیت تجاری را دارند، شرح دهید.

معیار موفقیت داده کاوی

برای یک خروجی موفق، با بیانی فنی و تخصصی معیاری تعریف کنید. مثلاً سطحی معین برای صحت پیش بینی یا برآورد گرایش به خرید با میزان صعود[۲۰] معین. همچنین با داشتن یک معیار برای موفقیت تجاری ممکن است لازم باشد آن را با عباراتی کیفی توصیف کنیم تا قضاوت در مورد میزان موفقیت تسهیل شود.

۱-۴ ارائه طرح پروژه

وظیفه: ارائه طرح پروژه

طرح های منتخب را برای رسیدن به هدف داده کاوی و در نتیجه رسیدن به هدف فعالیت تجاری تبیین کنید. طرح باید مجموعه پیش بینی شده ای ازمراحل قابل اجرا در حین توقف پروژه به انضمام انتخاب اولیه ابزار ها و تکنیک ها باشد.

---

۲۰ Lift

خروجی ها:

طرح پروژه

فهرستی از مراحلی که در پروژه باید اجرا شود، به اضافه مدت زمان اجرای آن‌ها، منابع مورد نیاز، ورودی ها، خروجی ها و وابستگی ها تهیه کنید. در پیاده سازی فرآیند داده کاوی در موارد مورد نیاز، عملیات را به دفعات زیاد تکرار کنید. به عنوان مثال تکرار در فاز های مدلسازی و ارزیابی.

به عنوان بخشی از طرح پروژه، تجزیه و تحلیل وابستگی های میان برنامه زمانی و ریسکها از اهمیت بسزایی برخوردار است. تأثیر نتایج تجزیه و تحلیل را در طرح پروژه دقیقاً مشخص کنید. همچنین اقدامات وپیشنهادات مقتضی در مواجهه با هر ریسک را تعیین کنید.

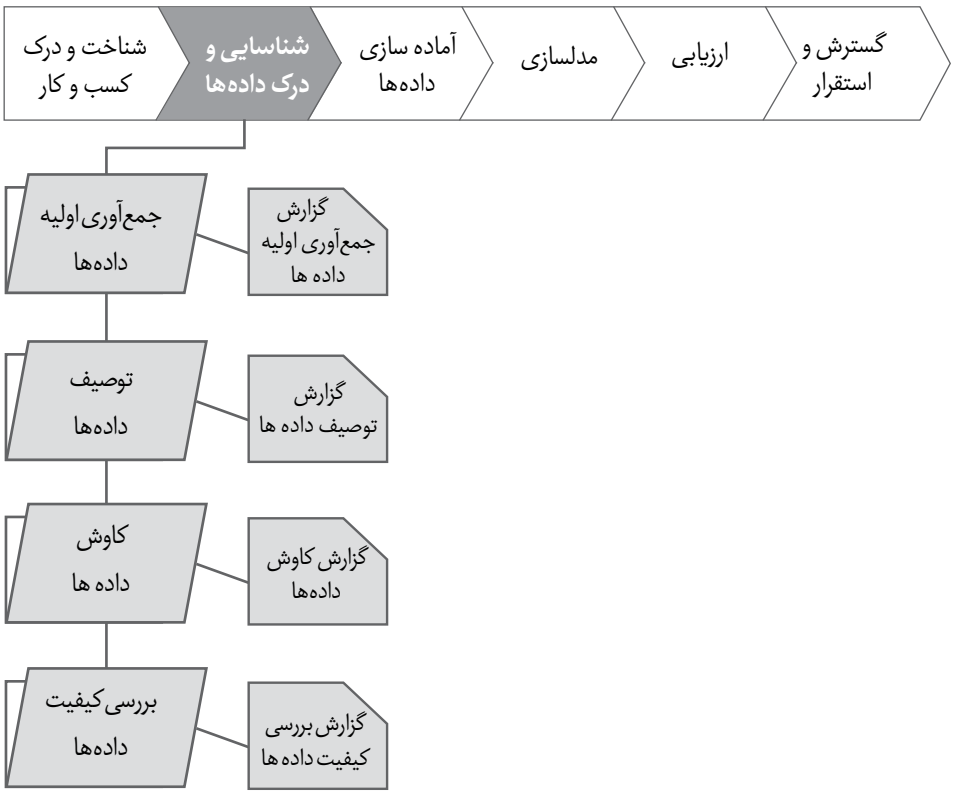
توجه: طرح پروژه شامل جزئیات طرح های هر فاز می‌باشد. به طور مثال در همین مرحله مشخص کنید قصد دارید از کدام استراتژی ارزیابی در فاز ارزیابی استفاده کنید.

طرح پروژه در حکم یک مستند پویاست که بررسی مجدد میزان پیشرفت و موفقیت آن در پایان هر فاز ضروریست و مطابق با پیشنهادات جدید باید بروز شود. همچنین زمان های بررسی مجدد را مشخص کنید چرا که این بررسی ها هم جزئی از طرح پروژه اند.

ارزیابی اولیه ابزارها و تکنیک ها

هم چنین در پایان فاز اول پروژه، باید یک ارزیابی اولیه از ابزارها و تکنیک ها انجام دهید. در این مرحله می‌بایستی ابزارهایی برای داده کاوی انتخاب کنید که کارایی لازم را در حین به کارگیری روشهای گوناگون برای مراحل مختلف داده کاوی داشته باشند. برای مثال، ارزیابی ابزارها و تکنیکها پیش از فرآیند، از اهمیت بسزایی برخوردار است چون انتخاب ابزارها و تکنیکها می تواند بر همه پروژه تأثیر گذار باشد.

۲. شناسایی و درک داده ها



شکل ۵: شناسایی و درک داده‌ها

۲-۱ جمع آوری اولیه داده ها

وظیفه: جمع آوری اولیه داده ها

دسترسی به داده‌ها در یک پروژه (یا دستیابی به داده‌ها)، در منابع پروژه لیست می‌شود. اگر برای فاز درک داده لازم باشد این جمع آوری اولیه شامل بارگذاری داده می‌شود. به عنوان مثال، در صورت استفاده از یک ابزار خاص برای درک داده، بارگذاری داده‌ها می تواند درک مناسبی از آن برای شما ایجاد نماید. این کار ممکن است منجر به مراحل اولیه آماده سازی داده‌ها شود.

تذکر: اگر چندین منبع داده به دست آوردید آنگاه بایستی یکپارچه سازی را نیز یا در این مرحله یا در مرحله بعد صورت دهید.

خروجی: گزارش جمع آوری اولیه داده ها

فهرستی از مجموعه داده (یا مجموعه داده‌های) قابل دستیابی، بعلاوه موقعیت آن‌ها در پروژه، روشهای دسترسی به آن‌ها در پروژه و مسائل و مشکلاتی که امکان رخ دادن دارند، تهیه کنید. مسائل پیش آمده و راه حل آن‌ها برای استفاده در دیگر مراحل همین پروژه (اگر باز هم تکرار شوند) یا پروژه‌های مشابه دیگر را یادداشت کنید.

۲-۲ توصیف داده ها

وظیفه: توصیف داده ها

ناخالصی و دیگر خواص ظاهری داده‌های به دست آمده را بررسی کنید و نتایج را گزارش دهید.

خروجی: گزارش توصیف داده ها

داده‌های به دست آمده را از نظر فرمت داده‌ها، مقدار داده‌ها (تعداد رکوردها و متغیرها در جدول)، ماهیت متغیرها و سایر ویژگی های ظاهری توصیف کنید. آیا داده‌های موجود نیازهای پروژه را برآورده می کنند؟

۲-۳ کاوش داده ها

وظیفه: کاوش داده ها

در این وظیفه پرسش های داده کاوی با استفاده از تحقیق، مجسم سازی و گزارش دهی طرح می‌شوند. این مرحله شامل توزیع فیلدهای کلیدی (مانند متغیر هدف یک عمل پیش بینی)، ارتباط بین دو یا چند صفت ، نتایج ترکیبات ساده، خواص معنی دار خرده جمعیت ها، و آنالیز آماری ساده است.

این تجزیه و تحلیل ها می‌تواند به طور مستقیم اهداف داده کاوی را مشخص کند و با استفاده از آن می توان توصیف داده‌ها وگزارش کیفیت ها را تصحیح کرد، و همچنین تبدیل ها و دیگر کارهای آماده سازی داده‌های مورد نیاز را برای تجزیه و تحلیل های ثانویه صورت داد.

خروجی ها: گزارش کاوش داده ها

نتایج این مرحله، شامل رابطه های کشف شده و فرضیات اولیه و تأثیر آن بر باقیمانده پروژه را بررسی کنید. در صورت لزوم از رسم نمودار که تعیین کننده خصوصیات داده‌هاست و موجد زیر مجموعه جالب توجهی از داده‌ها که امکان بررسی بیشتر را فراهم می‌کند استفاده کنید.

۲-۴ بررسی کیفیت داده ها

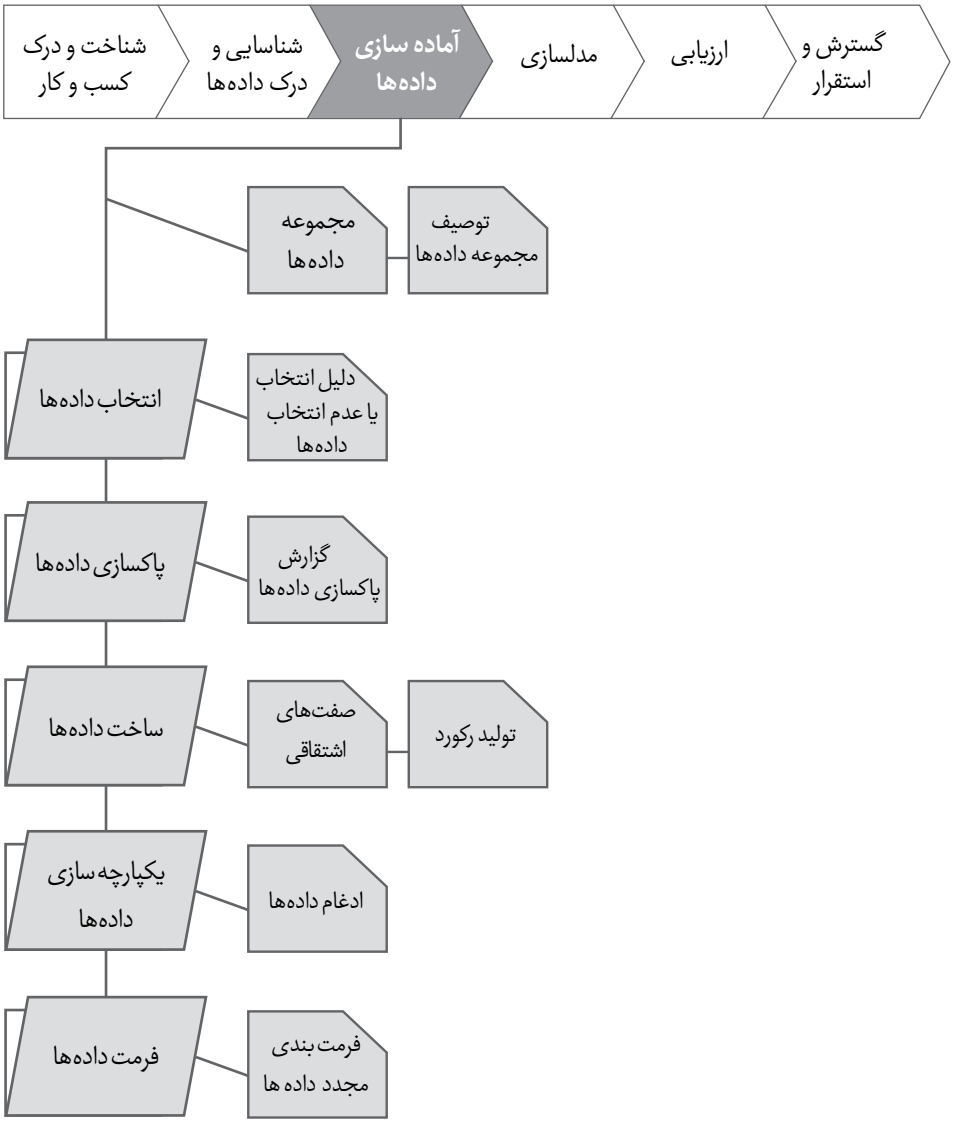
وظیفه: بررسی کیفیت داده ها

کیفیت داده‌ها را با سؤالاتی مانند : "آیا داده‌ها کامل اند؟ (آیا تمام موارد لازم را می پوشانند؟) آیا داده‌ها صحیح اند یا خطا دارند؟ و اگر خطا دارند، خطای آن‌ها چقدر شایع است؟ آیا در داده‌ها مقادیر گمشده وجود دارد؟ اگر هست چگونه نمایش داده می‌شوند، کجا اتفاق افتاده اند و چقدر شایع اند؟" بررسی نمائید.

خروجی: گزارش کیفیت داده ها

لیستی از نتایج بازرسی کیفیت داده‌ها تهیه کنید. اگر مشکل کیفی وجود دارد، راه حل های ممکن را لیست کنید. این راه حل ها وابستگی زیادی به اطلاعات داده کاوی و تجاری دارند.

۳. آماده سازی داده ها



شکل ۶: آماده سازی داده‌ها

خروجی ها:

مجموعه داده ها

مجموعه داده‌ها محصول فاز آماده سازی داده است، که برای مدلسازی یا تجزیه و تحلیل های عمده و اصلی پروژه استفاده خواهد شد.

توصیف مجموعه داده ها

مجموعه داده ای را که برای مدلسازی یا تجزیه و تحلیل های اصلی پروژه استفاده خواهند شد، تشریح کنید.

۳-۱ انتخاب داده ها

وظیفه: انتخاب داده ها

داده‌هایی را که برای تجزیه و تحلیل استفاده می‌شوند انتخاب کنید. معیارها شامل مرتبط بودن با اهداف داده کاوی، محدودیت های فنی و کیفی از قبیل محدودیت درحجم داده‌ها یا نوع داده‌هاست. توجه داشته باشید که انتخاب داده‌ها شامل انتخاب صفت ها (ستون ها) و همچنین انتخاب رکوردها (سطرها) در جدول است.

خروجی: دلیل انتخاب یا عدم انتخاب داده ها

فهرستی از داده‌هایی که تصمیم بر انتخاب یا عدم انتخاب آن‌ها گرفته اید تهیه کنید و دلایل تصمیم خود را نیز ذکر کنید.

۳-۲ پاکسازی داده ها

وظیفه: پاکسازی داده ها

کیفیت داده‌ها را تا سطح مورد نیاز تکنیک های تجزیه و تحلیل انتخاب شده افزایش دهید. این کار ممکن است شامل انتخاب زیرمجموعه ای بی نقص از داده‌ها، اضافه کردن مقادیر قراردادی مناسب یا تکنیک های جاه طلبانه تری مانند برآورد داده‌های گمشده با استفاده از مدلسازی باشد.

خروجی: گزارش پاکسازی داده ها

چگونگی اجرای اقدامات و تصمیماتی که تصمیم به انجام آن‌ها گرفته بودید را شرح دهید (این اقدامات و تصمیمات در گزارش مشکلات کیفی داده‌ها در جریان وظیفه بازرسی کیفیت داده‌ها در فاز شناسایی و درک داده‌ها ثبت شده بودند). همچنین تبدیلات روی داده‌ها، به منظور پاکسازی و بررسی تأثیرات آن روی نتایج تجزیه و تحلیل بایستی صورت گیرد.

۳-۳ ساخت داده ها

وظیفه: ساخت داده ها

این وظیفه شامل عملیات آماده سازی برای ایجاد داده‌های مفید می‌باشد، از قبیل ساخت صفت های اشتقاقی[۲۱]، ورود رکوردهای جدید دست نخورده یا تبدیل مقادیر صفت های موجود.

خروجی ها:

صفت های اشتقاقی

صفت های اشتقاقی صفت های جدیدی هستند که از یک یا تعداد بیشتری از صفت های موجود با رکوردهای یکسان ساخته می‌شوند. مثلاً مساحت یک صفت اشتقاقی است که از دو صفت طول و عرض به این صورت ساخته می‌شود: مساحت = طول × عرض

تولید رکورد

ایجاد رکوردهای جدید را به طور کامل تشریح کنید. مثال: ساخت رکورد برای مشتری هایی که در طی سال قبل خریدی نداشته اند. دلیلی برای وجود این نوع رکوردها در داده‌های اولیه نیست؛ اما برای مدلسازی، عمل صحیح آن است که مشتریان بدون خرید هم نمایش داده شوند.

۳-۴ یکپارچه سازی داده ها[۲۲]

وظیفه: یکپارچه سازی داده ها

روش هایی وجود دارد که به وسیله آن‌ها اطلاعات از چندین جدول یا منابع اطلاعاتی مختلف ترکیب می‌شوند و رکوردها یا مقادیر جدید می سازند.

خروجی: ادغام داده ها

ادغام جداول یعنی دو یا چند جدول را که حاوی اطلاعات متفاوتی در مورد موضوع واحدی هستند به هم الحاق کنیم. به طور مثال یک خرده فروش زنجیره ای یک جدول با اطلاعاتی در مورد مشخصه های عمومی فروشگاه (از قبیل سطح بنا، پیاده رو مجاور)، جدولی با اطلاعات خلاصه شده فروش (از قبیل سود و درصد تغییر در فروش از یک سال قبل) و جدولی با اطلاعات جمعیت شناختی درباره مناطق مجاور دارد. در هر کدام از این جداول یک سطر برای هر فروشگاه در نظر گرفته شده است. می توان این جداول را با هم ادغام کرد و یک جدول جدید ساخت که در آن برای هر فروشگاه یک سطر در نظر گرفته شده است و اطلاعات هر سطر ترکیبی از اطلاعات جداول قبل است.

همچنین داده‌های ادغام شده، تجمیع داده‌ها را تحت پوشش قرار می‌دهد. تجمیع داده‌ها عبارت

---

۲۱ Derived Attributes

۲۲ Integrate Data

اسـت از عملیاتـی که مقادیر جدید را به وسـیله خلاصه کـردن اطلاعات چندین رکورد یا جدول محاسـبه می‌کند. برای مثال، تبدیل یک جدول خرید مشتریان که در آن برای هر خرید یک رکورد ثبت شده است، به جدول جدیدی که برای هر مشـتری یک رکورد ثبت شده است و صفت های آن عبارتند از تعداد خرید، مقدار متوسط خرید، درصد سفارشات، هزینه کردن کارت اعتباری، درصد اقلام در حال گسترش و غیره.

وظیفه: فرمت داده ها

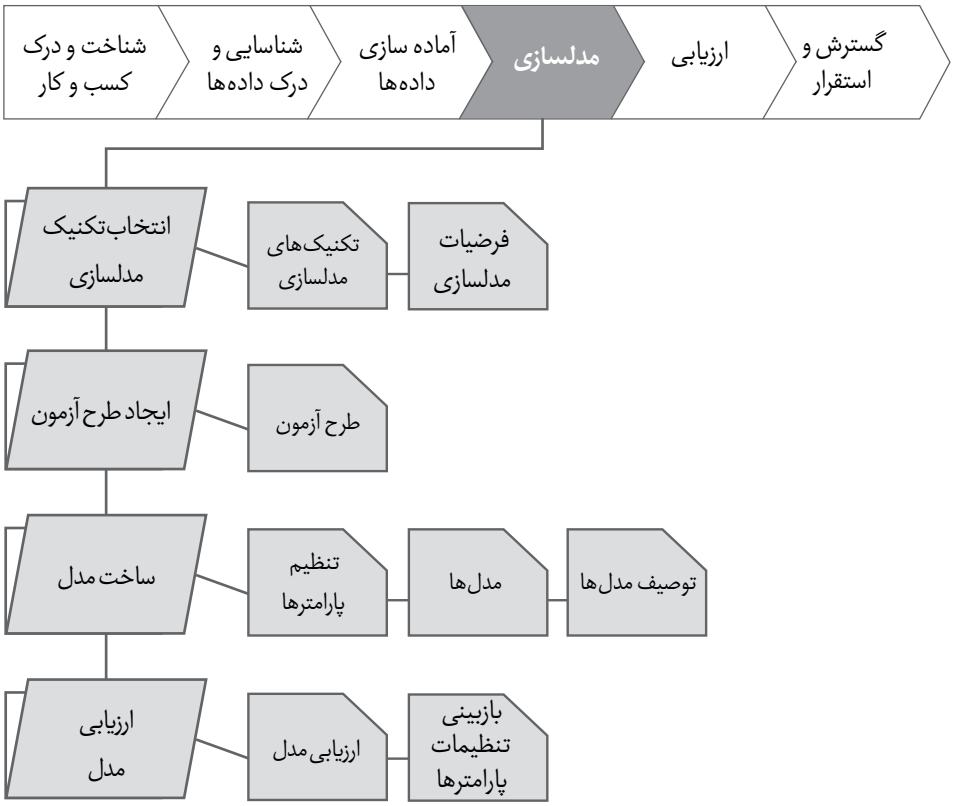
تغییر فرمت داده‌ها شامل اصلاحات نحوی است و در آن ارزش و مفهوم داده‌ها تغییر نمی‌کند. این کار در صورت لزوم برای ابزارهای مدلسازی استفاده می‌شود.

خروجی: فرمت بندی مجدد داده ها

بعضی نرم افزارها، شـرایطی بـرای متغیرها دارند، مثلا محتوای اولین متغیر، شناسـه های یکتا برای هر سـطر از داده‌ها باشـد، یا محتوای آخرین متغیر می‌بایستی متغیر خروجی باشد که به وسیله مدل پیش بینی می‌شـود. گاهی لازم اسـت ترتیب رکوردها را در مجموعه داده‌ها تغییر دهیم. ممکن اسـت ابزارهای مدلسازی به مرتب کردن رکوردها بر حسب مقادیر متغیرهای خروجی نیاز داشته باشند. معمولاً رکوردهای یک مجموعه داده، در ابتدا به صورتهای خاصی مرتب شـده اند، اما الگوریتم مدلسـازی بر حسـب لزوم، آن‌ها را به صورت کاملاً تصادفی مرتب می‌کند. برای مثال، هنگامی که از شـبکه های عصبی استفاده می کنیم، به طور کلی بهتر اسـت که رکوردها را با ترتیب تصادفی نمایش دهیم، اگر چه نرم افزارها، این کار را به صورت اتوماتیک و بدون دخالت مستقیم کاربر انجام می دهند.

بـه عـلاوه به طـور کلی تغییرات نحـوی دیگری نیز وجـود دارد که بـرای برآورده کـردن نیازهای نرم افزارهـای مدلسـازی خاصی به کار می روند. مثلاً حـذف کاما از درون فایل های متنی کـه در آن‌ها از جدا کننده کاما استفاده شده باشد، یا حذف تمام مقادیری که بیش از ۳۲ کاراکتر دارند.

۴. مدلسازی



شکل ۷: مدلسازی

۴-۱ انتخاب تکنیک مدلسازی

وظیفه: انتخاب تکنیک مدلسازی

نخستین گام در مدلسازی، انتخاب تکنیکی است که می خواهیم مورد استفاده قرار دهیم. از آنجا که احتمالاً نرم افزار مورد استفاده را پیش از این در فاز شناسایی و درک کسب و کار انتخاب نموده اید، این وظیفه شامل چند تکنیک خاص مدلسازی ست؛ تکنیک هایی مانند: ساخت درخت تصمیم با استفاده از C۴,۵ یا ایجاد شبکه عصبی به روش پس انتشار[۲۳]. اگر از چندین تکنیک استفاده شود برای هر یک از تکنیک ها این وظیفه را جداگانه مورد استفاده قرار دهید.

خروجی ها:

تکنیک های مدلسازی

تکنیک هایی را که برای مدلسازی از آن‌ها استفاده کرده اید یادداشت کنید.

فرضیات مدلسازی

در بسیاری از تکنیک های مدلسازی فرض هایی در مورد داده‌ها صورت می گیرد، مانند: همه صفت ها توزیع یکنواخت دارند، هیچ داده ای گم نشده است، صفت طبقه بندی کننده باید به صورت نمادی[۲۴] باشد. دیگر فرضهای صورت داده شده را نیز یادداشت کنید.

۴-۲ ایجاد طرح آزمون[۲۵]

وظیفه: ایجاد طرح آزمون

پیش از ساخت یک مدل، باید روشی برای سنجش کیفیت و اعتبار مدل ایجاد کنیم. به عنوان مثال در وظایف مربوط به داده کاوی نظارتی مانند طبقه بندی، از نرخ خطا برای سنجش کیفیت پیش بینی ها در مدل های داده کاوی استفاده می‌شود. بنابراین معمولاً داده‌ها را به مجموعه های آموزشی و آزمایشی تقسیم می کنیم و پس از ایجاد مدل روی داده‌های آموزشی، کیفیت آن را روی مجموعه جدا شده آزمایشی مورد بررسی قرار می دهیم.

خروجی: طرح آزمون

طرح انتخاب شده برای آموزش، آزمون و ارزیابی را تشریح کنید. یکی از بخش های ابتدایی طرح، نحوه تقسیم بندی مجموعه داده به داده‌های آموزش و آزمایشی، و ارزیابی این مجموعه هاست.

---

۲۳ Back Propagation

۲۴ Symbolic

۲۵ Test Design

۳-۴ ساخت مدل

وظیفه: ساخت مدل

با استفاده از نرم افزارهای داده کاوی، برای مجموعه داده‌های آماده شده یک یا چند مدل بسازید.

خروجی ها:

تنظیم پارامترها

در هر نرم افزار مدلسازی، عموماً پارامترهایی وجود دارند که باید تنظیم شوند. فهرستی از پارامترها و مقادیر انتخاب شده برای آن‌ها همراه با دلیل انتخاب آن‌ها، تهیه کنید.

مدل ها

مدل های واقعی که به وسیله نرم افزارهای مدلسازی ساخته شده اند.

توصیف مدل ها

مدل حاصله را تشریح کنید. گزارشی درباره تفسیر مدل ها تهیه کرده و اشکالات مدل و مفهوم آن‌ها را یادداشت کنید.

۴-۴ ارزیابی مدل

وظیفه: ارزیابی مدل

تفسیر یک مهندس داده کاو از مدل، بر اساس میزان اطلاعاتش از حوزه کاربرد مدل، معیارهای موفقیت داده کاوی و تست های طراحی شده شکل می گیرد. این وظیفه با فاز ارزیابی (بخش بعد) در تداخل است. به دلیل آنکه مهندس داده کاو مرجع قضاوت در مورد میزان موفقیت مدلسازی است و بر مبنای اصول فنی در جستجوی تکنیک های بیشتر است، باید در تماس دائم با تحلیلگران تجاری و متخصصین آن حوزه خاص باشد تا بتواند نتایج داده کاوی را در زمینه تجاری توضیح دهد. در این قسمت فاز ارزیابی مدل و دیگر نتایج حاصل در ضمن پروژه به یک اندازه مورد بررسی قرار می گیرند.

یک مهندس داده کاو باید جدولی برای رتبه بندی مدل ها ایجاد کند. این رتبه بندی با توجه به معیارهای ارزیابی صورت می گیرد و حتی الامکان معیارهای موفقیت تجاری و نیز مد نظر قرار دادن موارد واقعی تجارت را در بر می گیرد. در بیشتر پروژه‌های داده کاوی، مهندس داده کاو از یک تکنیک چند بار استفاده می‌کند یا نتایج داده کاوی را با تکنیک ها جایگزین به دست می آورد. در این بخش او همه نتایج را با توجه به معیارهای موفقیت با هم مقایسه میکند.

خروجی ها:

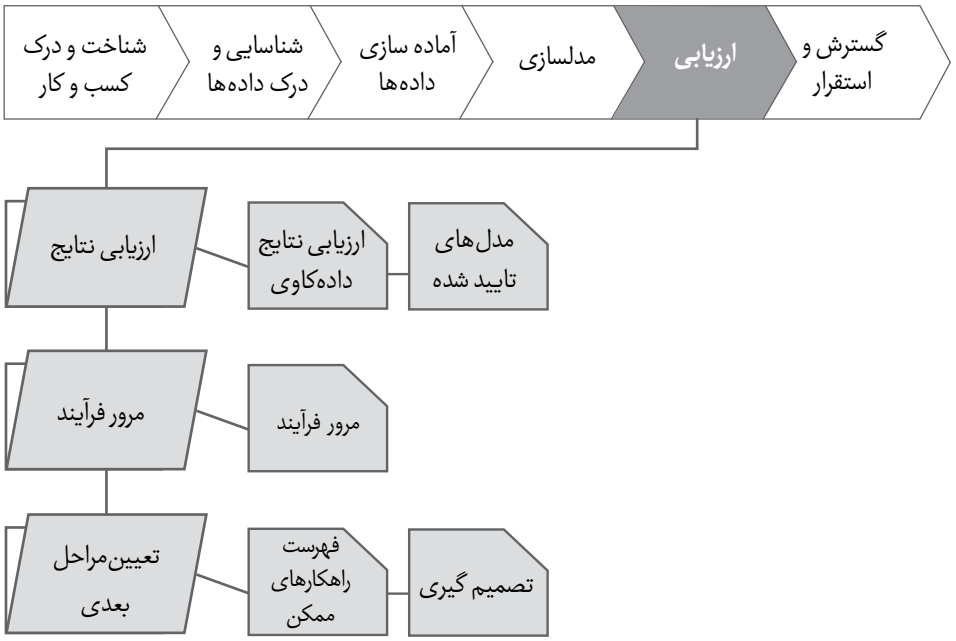
ارزیابی مدل

نتایج این بخش را خلاصه کنید. فهرستی از کیفیت مدل های تولید شده (مثلاً بر حسب دقت) تهیه کنید و آن را رتبه بندی کنید.

بازبینی تنظیمات پارامترها

با توجه به ارزیابی مدل، تنظیمات پارامترها بازبینی می‌شود و برای اجراهای آینده در بخش مدلسازی آن‌هـا را تنظیم می کنیم. مدلسـازی را تکرار میکنیم و مدل های تـازه را ارزیابی می کنیم تا کاملاً مطمئن شویم مدل حاصل بهترین مدل است. تمام بازنگری ها و ارزیابی ها را یادداشت می کنیم.

۵. ارزیابی



شکل ۸: ارزیابی

۵-۱ ارزیابی نتایج

وظیفه: ارزیابی نتایج

در مراحـل قبلی ارزیابی ویژگی هایی مانند صحت و جامعیت مدل ها مورد بررسـی قرار گرفت. در این مرحله میزان تناسـب مدل با اهداف فعالیت تجاری بررسـی می‌شود و در صورت وجود نقص در مدل، پاره ای از دلایل تجاری آن تشـخیص داده می‌شـوند. از دیگر خصیصه های فاز ارزیابی، آزمودن مدل در موارد واقعی است؛ البته اگر این امکان از نظر هزینه و زمان میسر باشد.

هم چنین دیگر نتایج ناشی از مدلسازی را ارزیابی کنید. نتایج داده کاوی می تواند هم پوشش دهنده مدلهایی که مرتبط با هدف اصلی داده کاوی هسـتند باشـد، هم دربرگیرنـده یافته های دیگری که لزوماً ارتباط چندانی به هدف اصلی داده کاوی نداشته باشند؛ اگر چه ممکن است حاوی اطلاعاتی باشند که در ادامه مورد استفاده قرار گیرد.

خروجی ها:

ارزیابی نتایج داده کاوی با توجه به معیارهای موفقیت تجاری

نتایـج ارزیابی را بـا توجه به معیارهای موفقیت تجاری خلاصه کنید و قضـاوت نهایی خود را در مورد تناسب پروژه با اهداف ابتدایی، به آن ضمیمه کنید.

مدل های تأیید شده

پـس از ارزیابی مدل با توجـه به معیارهای موفقیت تجاری، مدل های منطبق با معیارها، در لیسـت مدل های تأیید شده قرار می گیرند.

۵-۲ مرور فرآیند

وظیفه: مرور فرآیند

در این مرحله به مدل های رضایت بخشی دست یافته ایم که خواسته های تجاری مان را هم برآورده میکنـد. اکنون زمان مناسـبی اسـت که مروری کلی بر تعهدات داده کاوی داشـته باشـیم. آیـا عامل ها یا ویژگیهـای مهمی هسـت که مورد توجه قرار نگرفته باشـد؟ در این مرحلـه از داده کاوی، فرآیند بازنگری از یک بازنگری کیفی آغاز می‌شـود؛ آیا مدل را به درسـتی ساخته ایم؟ آیا فقط از متغیرهایی که مجاز بوده ایم استفاده کرده ایم و آیا آن‌ها برای تجزیه و تحلیل های بعدی هم در اختیار ما قرار خواهند داشت؟

خروجی: مرور فرآیند

فرآینـد بازنگری را خلاصه کنید و پیشـنهادهایی برای فعالیت هـای از قلم افتاده یا فعالیت هایی که مجدداً باید تکرار شوند، ارائه کنید.

۳-۵ تعیین مراحل بعدی

وظیفه: تعیین مراحل بعدی

با توجه به ارزیابی نتایج و بازبینی فرآیند باید در مورد نحوه ادامه پروژه تصمیم گیری شـود. اینکه آیا پروژه در این مرحله تمام شـده و گسـترش و استقرار آن شـروع گردد یا اینکه تکرارهای بعدی آغاز شده و یا اصلاً مقدمات انجام پروژه داده کاوی تازه ای فراهم شـود. این بخش شـامل بررسـی منابع باقی مانده و تأثیرگذاری بودجه بر نحوه تصمیم گیری خواهد بود.

خروجی ها:

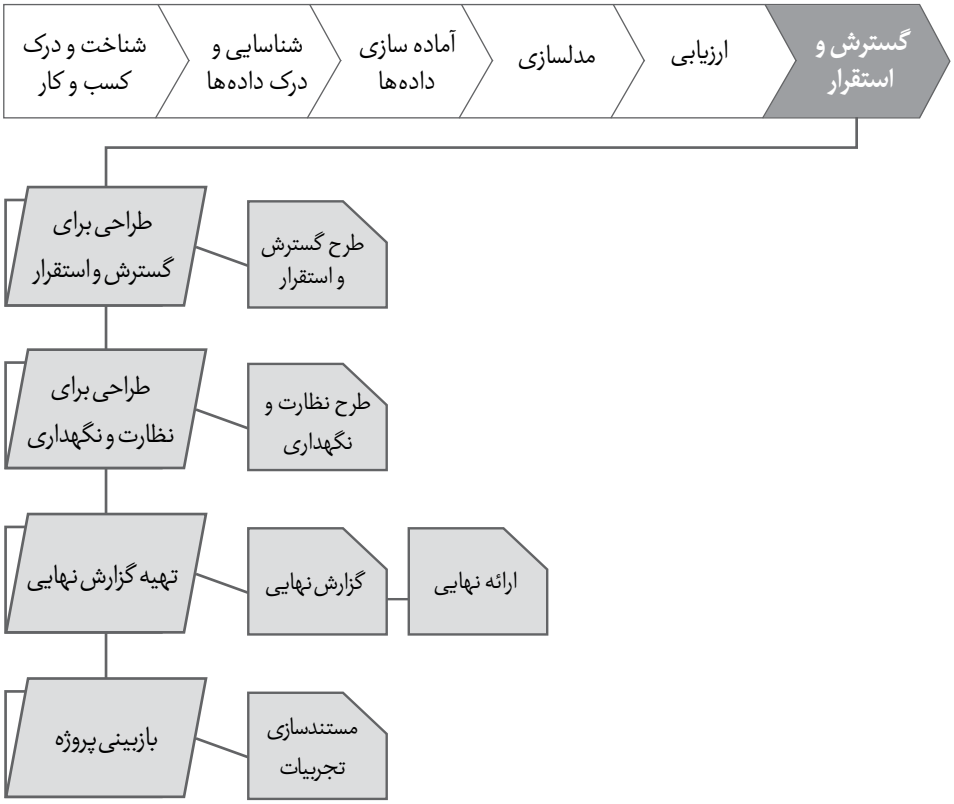
فهرست راهکارهای ممکن

فهرستی از راهکارهای ممکن بعدی، به همراه دلایلی در تأیید یا رد آن‌ها تهیه کنید.

تصمیم گیری

تصمیم خود در مورد نحوه پیگیری کار را به همراه دلایل آن شرح دهید.

۶. گسترش و استقرار



شکل ۹: گسترش و استقرار و استقرار

۶-۱ طراحی برای گسترش و استقرار

وظیفه: طرح گسترش و استقرار

این فاز بر اساس نتایج ارزیابی، یک استراتژی مناسب برای گسترش و استقرار نتایج تعیین می‌کند. اگر برای ساخت مدل روال عمومی مناسبی به دست آورده اید پرونده ای شامل مشروح آن برای توسعه های بعدی تهیه کنید.

خروجی: طرح گسترش و استقرار

استراتژی گسترش و استقرار را همراه با گام های لازم و چگونگی انجام آن‌ها به طور خلاصه بیان کنید.

۶-۲ طراحی برای نظارت و نگهداری[۲۶]

وظیفه: طرح نظارت و نگهداری

اگر نتایج داده کاوی از اطلاعات روزانه تجاری و محیط آن حاصل شـده باشـد، نظارت و نگهداری آن بسـیار حائز اهمیت است. یک آماده سـازی با دقت در استراتژی نگهداری منجر به عدم استفاده نادرست در بلند مدت از نتایج داده کاوی خواهد شد. با هدف نظارت بر گسترش و استقرار نتایج داده کاوی، پروژه بـه یک طرح تفصیلـی از فرآیند نظارت نیاز دارد. این طرح نیازمند برخی گسـترش و اسـتقرار های خاص می‌باشد.

خروجی: طرح نظارت و نگهداری

استراتژی نظارت و نگهداری را همراه با گام های لازم و چگونگی انجام آن‌ها به طور خلاصه بیان کنید.

۶-۳ تهیه گزارش نهایی

وظیفه: تهیه گزارش نهایی

در پایان کار، سرپرسـتان پروژه و تیم آن‌ها یک گزارش نهایی تهیه می کنند. موضوع گزارش بسـتگی به طرح گسـترش و اسـتقرار دارد، این گزارش هم می تواند خلاصه ای از وضعیت پروژه و نتایج تجربی آن باشد هم می تواند گزارشی نهایی برای آشنایی با نتایج داده کاوی باشد.

خروجی ها:

گزارش نهایی

این آخرین گزارش تعهد شده در پروژه داده کاوی است. این گزارش شامل همه موارد ذکر شده قبلی و همچنین نتایج خلاصه سازی شده و سازماندهی شده می‌باشد.

---

۲۶ Monitoring and Maintenance

ارائه نهایی

این امر در صورت برگزاری نشستی در پایان پروژه برای معرفی شفاهی نتایج به مشتری انجام می‌شود.

۶-۴ بازبینی پروژه

وظیفه: بازبینی پروژه

این امر شامل ارزیابی صحت امور انجام شده و بهبود مواردی که نیازمند اصلاحند می‌باشد.

خروجی: مستندسازی تجربیات

خلاصه ای از تجربیات مهم کسب شده در طی پروژه تهیه کنید. به عنوان مثال ریسک ها، موارد گمراه کننده یا راهنمایی هایی برای انتخاب مناسب ترین تکنیک های داده کاوی در شرایط مشابه می تواند بخشی از این خلاصه سازی را به خود اختصاص دهد. در یک پروژه ایده‌آل، پرونده تجارب، شامل گزارش های شخصی همه اعضا از قسمت های مختلف پروژه می‌باشد.

فصل سوم
راهنمای کاربر
CRISP-DM

۱. شناخت و درک کسب و کار

۱-۱ تعیین اهداف فعالیت تجاری

وظیفه: تعیین اهداف فعالیت تجاری

یـک تحلیلگر داده در نخسـتین گام باید بتواند با مد نظر قراردادن چشـم انـداز فعالیت تجاری درک کاملی از خواسـته های واقعی مشـتری حاصل کند. مشـتری ها غالباً چندین هدف موازی و محدود دارند که باید در توازنی منطقی قرار گیرند. هدف تحلیلگر در آغاز آنسـت که عوامل مؤثر بر خروجی های پروژه را شناسایی کند. بیدقتی در این مرحله ممکن است ما را متحمل هزینه های زیادی کند که به خاطر اصرار ما در یافتن جوابی درست برای سؤالی نادرست ایجاد می‌شود.

خروجی: پس زمینه

جمع آوری و تطبیق اطلاعات اولیه درباره وضعیت کسب و کار شرکت در شروع پروژه انجام می پذیرد. این جزئیات علاوه بر کمک به شناخت دقیق اهداف تجاری مورد نظر، برای شناخت دقیقتر منابع انسانی و غیر انسانی که درخلال پروژه مورد استفاده قرار می گیرند نیز سودمند است.

فعالیت ها:

سازماندهی

- ایجاد چارت سـازمانی برای بخشـها و گـروه های در گیر پـروژه. چارت تنها باید شـامل نام مدیران و مسئولیت هایشان باشد.
- شناسایی افراد کلیدی و نقش آن‌ها در فعالیت تجاری
- شناخت یک حامی داخلی (که هم حامی مالی باشد، هم یک کاربر اصلی و متخصص در حوزه مربوط)
- بررسی وجود یک هیأت هدایت کننده و شناخت اعضای آن
- شـناخت واحد های تجاری که با پروژه داده کاوی دارای اثر متقابل اند. (بازاریابی، فروش، قسـمت مالی)

محدوده مسئله

- شناخت محدوده مسئله (مانند بازاریابی، توجه به مشتری، توسعه تجارت و...)
- توضیح مسئله در حالت کلی
- بررسـی شـرایط فعلی پروژه (بررسـی اینکه آیا واحدی که برای آن داده کاوی انجام می‌شـود شناخت کافـی از داده کاوی دارد یـا اینکـه داده کاوی بـه عنوان یـک تکنولوژی کلیدی باید به آن‌ها شناسـانده شود؟)
- تعییـن پیـش نیازهـای پروژه (انگیزه داده کاوی چیسـت؟ آیا داده کاوی کاربردی در کسـب و

کار کنونـی دارد؟)

- معرفی داده کاوی در فعالیت تجاری در صورت لزوم
- شـناخت گروه هدف برای نتایج پروژه (آیا قصد شـما تهیه یک گزارش برای مدیریت عالی است یا راه اندازی سیستمی برای کاربرانی بی اطلاع)
- شناخت انتظارات و خواسته های کاربران.

راه حل های کنونی

- بیان تمام راه حل هایی که هم اکنون برای این مسئله مورد استفاده قرار می گیرند.
- بیان معایب و مزایای راه حل های کنونی و سطح مقبولیت آن‌ها نزد کاربران.

خروجی: اهداف فعالیت تجاری

با توجه به چشم انداز فعالیت تجاری هدف اصلی مشتری را شرح دهید. علاوه بر هدف اصلی فعالیت تجاری سؤالات دیگری وجود دارد که مشتری مایل است مورد بحث قرار دهد. برای مثال هدف اصلی یک فعالیت تجاری ممکن اسـت حفظ مشـتری از طریق پیش بینی زمانی که او به همکاری با شـرکت رقیب تمایل پیدا می‌کند باشد. در حالی که هدف دوم پروژه ممکن است پی بردن به این سؤال باشد آیا کاهش حق الزحمه، فقط روی بخشی از مشتریان تأثیرگذار است یا روی همه آن‌ها.

فعالیت ها:

- توضیح غیر رسمی مسأله ای که باید با داده کاوی حل شود.
- آماده کردن سؤالات در چهارچوب فعالیت تجاری مورد بررسی با دقت هر چه بیشتر.
- آماده کردن سایر نیازمندی ها. (در تجارت هیچ کدام از مشتری ها نباید از دست داده شوند)
- تعیین منافع مورد انتظار در چهارچوب فعالیت تجاری.

توجه!

اهداف غیر قابل دستیابی را با اهدافی واقعی و قابل دسترس جایگزین کنید.

خروجی: معیار های موفقیت فعالیت تجاری

معیار های مفید بودن و موفقیت آمیز بودن نتایج یک پروژه را از دیدگاه تجاری شرح دهید. این معیار ها ممکن سـت کاملاً معلوم و سـهل الوصول باشـند، مانند کاهش مشـتری های در گردش از یک سـطح معین؛ و یا کلی و غیر قابل لمس باشند، مانند «ایجاد نگاهی نو در روابط». در مورد دوم باید مشخص شود قضاوت درباره موفقیت آمیز بودن پروژه از جانب چه کسی صورت خواهد گرفت.

فعالیت ها:

- تعیین معیارهای تجارت موفق. (مثلاً بهبود ده درصدی سرعت پاسخ دهی در تبادلات و افزایش بیست درصدی سرعت ثبت نام)
- تعیین ارزیابی کننده معیارها.

یاد آوری!

هـر کـدام از معیارهای موفقیت باید حداقل با یکی از اهداف تعیین شـده فعالیت تجاری در ارتباط باشـد.

پیشنهاد!

پیـش از ارزیابـی وضعیـت (گام بعـدی) می‌توانید تجربه های پیشـین در مورد حل این مسـأله که با CRISP-DM یا روش های دیگر انجام شده اند را ملاحظه کنید.

۱-۲ ارزیابی وضعیت

وظیفه: ارزیابی وضعیت

این وظیفه جزئیات بیشـتری از اطلاعات به دست آمده در مورد همه منابع، محدودیت ها و فرضیات است و همچنین سایر عواملی که در تعیین هدف داده کاوی و طراحی پروژه باید مورد توجه قرار گیرند.

خروجی: فهرست منابع

فهرستی از منابع موجود پروژه تهیه کنید. این فهرست شامل پرسنل (کارشناس داده، پشتیبان فنی، تیـم داده کاوی)، داده (اسـتخراج داده‌های ثبت شـده، دسترسـی به پایگاه داده‌های فعـال یا غیر قابل اسـتفاده)، منابع محاسباتی (سـخت افزارها) و نرم افزار (ابزارهای داده کاوی ودیگر نرم افزارهای مربوط) خواهد بود.

فعالیت ها:

منابع سخت افزاری

- شناخت سخت افزار پایه
- فراهم کردن سخت افزارهای پایه ای مورد نیاز برای پروژه داده کاوی
- چک کردن برنامه نگهداری سخت افزار در صورت تعارض با دیگر سخت افزارهای موجود برای پروژه
- شناسـایی سـخت افزارهای موجود برای ابزار داده کاوی مورد اسـتفاده (در صورت شناخته شده بودن ابزار این مرحله)

منابع داده و دانش

- شناسایی منابع داده
- شناسایی انواع منابع داده (منابع آنلاین، منابع خاص، اسناد نوشتاری و...)
- شناسایی منابع دانش
- شناسایی انواع منابع دانش (منابع آنلاین، خاص، اسناد نوشتاری و...)
- بررسی ابزار و تکنیک های موجود
- تشریح زمینه های دانشی مربوطه (رسمی و غیررسمی)

منابع پرسنلی

- شناسایی حامیان پروژه (در صورت تمایز از حامیان داخلی مطرح شده در قسمت ۱-۱-۱)
- شناسایی مدیران سیستم، مدیر پایگاه داده و تیم پشتیبان فنی برای سؤالات بعدی.
- شناسایی تحلیلگر بازار، کارشناسان داده کاوی و آمار و بررسی وجود این افراد.
- بررسی وجود متخصص در زمینه های مرتبط به فازهای بعدی

یادآوری!

از یاد نبرید ممکن است در قسمت هایی از پروژه به کارمندان تیم پشتیبانی فنی احتیاج داشته باشید، مثلاً در مرحله تبدیل داده.

خروجی: نیازها، فرضیات و محدودیت ها[۲۷]

فهرستی از همه نیازهای پروژه همراه با برنامه زمان بندی تکمیل آن‌ها، کیفیت و قابلیت درک نتایج، تضمین ها و همچنین پی آمد های قانونی ممکن تهیه کنید. به عنوان قسمتی از خروجی مطمئن شوید اجازه استفاده از داده‌ها را دارید.

فهرستی از فرضیات حاصل از پروژه تهیه کنید. این فهرست هم می‌تواند شامل فرضیاتی باشد که در حین داده کاوی رسیدگی می‌شوند یا فرضیاتی غیر قابل بررسی که بستگی به شرایط تجاری داشته باشد. اهمیت این مسئله زمانی بیشتر می‌شود که این فرضیات بر درستی نتایج تاثیرگذار باشد.

همچنین فهرستی از محدودیت های ایجاد شده به وسیله پروژه تهیه کنید. این محدودیت ها ممکن است شامل کمبود منابع برای انجام وظایف پروژه یا محدودیت های قانونی و اخلاقی برای استفاده از خود داده‌ها یا راه حل های داده کاوی باشد.

---

۲۷ Requirements, Assumptions and Constraints

فعالیت ها:

نیازمندی ها

- نمایه سازی[۲۸] گروه هدف
- ثبت نیازمندی ها با برنامه زمانی
- ثبت نیازمندی ها به صورتی که قابل درک، دقیق و قابل توسعه باشند و تکرار پذیری پروژه داده کاوی و نتیجه دهی مدل ها را امکان پذیر کند.
- ثبت نیازمندی ها با ذکر تضمین ها، موانع قانونی، موارد محرمانه، اطلاع رسانی ها و زمان بندی های در نظر گرفته شده برای پروژه.

فرضیات

- توضیح همه فرض ها (شامل موارد نا مشخص) و شفاف سازی آن‌ها. (به عنوان مثال، با توجه به سؤالات تجاری، حداقل به چه تعداد مشتری با سن حدود ۵۰ سال نیاز داریم؟).
- تهیه فهرستی از فرضیات با توجه به وضعیت کیفی داده‌ها. (دقت، میزان دسترسی).
- تهیه فهرستی از فرضیات با توجه به عوامل خارجی. (نتایج اقتصادی، تولیدات رقابتی و پیشرفت های تکنیکی)
- تشریح فرضیاتی که منجر به تولید برآوردهایی گشته اند. (بطور مثال قیمت نرم افزار خاصی کمتر از ۱۰۰۰ دلار برآورد شده است)
- لیست کردن همه فرضیات با توجه به قابل فهم بودن و توضیح آن‌ها یا توضیح مدل. (مدل و نتایج آن از چه طریقی باید به مدیران و حامیان معرفی شود)

محدودیت ها

- چک کردن محدودیت های عمومی (پی آمدهای قانونی، بودجه، زمان بندی، و منابع)
- چک کردن دسترسی درست به منابع داده‌ها (محدردیت های دسترسی، رمزهای مورد نیاز)
- چک کردن قواعد فنی دسترسی به داده‌ها (سیستم های عملیاتی، سیستم مدیریت داده‌ها، فرمت داده‌ها) چک کردن دسترسی به اطلاعات مورد نیاز
- چک کردن محدودیت های بودجه (هزینه های ثابت، هزینه های جاری و...)

یادآوری!

لیست فرضیات شامل فرضیات مرتبط با شروع پروژه نیز می‌باشد، مانند نقطه شروع پروژه.

---

۲۸ Profiling

خروجی: ریسک ها و پیشامدها

ریسـکها یا پیشامدهایی که ممکن است رخ دهد، برنامه زمانبندی فشرده، هزینه ها و نتایج را لیست کنید. همچنین طرح مقتضی[۲۹] خود را نیز مشخص کنید؛ برای کم کردن فشار و رهایی از پیامدهای منفی چه اقداماتی باید انجام داد؟

فعالیت ها:

شناسایی ریسکها

- شناسایی ریسکهای کسب و کار (مثال: رقیب با نتایج بهتر رشد بالاتری دارد)
- شناسایی ریسکهای سازمانی (مثال: واحد سفارش دهنده پروژه قادر به تامین هزینه های پروژه نباشد)
- شناسـایی ریسـک های مالی (مثال: تامین مالی پروژه در فازهای مختلف آن به نتایج اولیه پروژه داده کاوی بستگی خواهد داشت)
- شناسایی ریسک های فنی
- شناسـایی ریسـک هایی که بـه داده‌ها و پایـگاه داده‌ها بسـتگی دارد. (مثال: ضعف هـای کیفیتی و پوششی داده‌ها)

توسعه طرح های مقتضی

- شرایط ممکن برای رخ دادن هر ریسک را تعیین کنید.
- طرح های مقتضی را توسعه دهید.

خروجی: مجموعه لغات

واژه نامه ای مرتبط با مفاهیم بکار رفته در پروژه تهیه کنید. این واژه نامه حداقل شامل دو بخش زیر باید باشد:

۱) واژه نامه لغات جاری مربوطه که شامل مفاهیم تجاری موجود در پروژه است.

۲) واژه نامه لغات داده کاوی که از طریق مثالی مرتبط با مسأله تجاری، توضیح داده شده باشد.

فعالیت ها

- واژه نامـه های قبلی را بررسـی کنید. در صورت عدم وجود چنیـن واژه نامه ای برای آماده کردن پیش نویس یک واژه نامه تازه دست به کار شوید.
- برای درک بهتر لغات از متخصصان آن بخش کمک بگیرید.
- با اصطلاحات تجاری آشنا شوید.

خروجی: هزینه و فایده

---

۲۹ Contingency Plan

یک تجزیه و تحلیل سود و زیان برای پروژه تهیه کنید. این تجزیه و تحلیل باید هزینه های پروژه را با پتانسیل های سود دهی آن (در صورت موفقیت) مقایسه کند.

پیشنهاد!

مقایسه باید حتی الامکان اختصاصی باشد. در این صورت شما قادر به بهبود فعالیت تجاری خود خواهید بود.

فعالیت ها

- هزینه جمع آوری داده را برآورد کنید.
- هزینه تهیه و اجرای یک راه حل را برآورد کنید.
- مزایای استقرار یک راه حل را شناسایی کنید. (افزایش رضایت مشتری، بازدهی سرمایه، و افزایش سود)
- هزینه عملیاتی شدن طرح را برآورد کنید.

توجه!

برآورد هزینه های پنهان از قبیل استخراج داده‌های تکراری و آماده سازی آن‌ها، تغییییر در جریان کاری و زمان آموزش برای یادگیری را فراموش نکنید.

۱-۳ تعیین اهداف داده کاوی

وظیفه: تعیین اهداف داده کاوی

اهداف فعالیت تجاری، هدف پروژه را با اصطلاحات خاص فعالیت تجاری و اهداف داده کاوی، هدف را به زبان فنی و تخصصی داده کاوی توضیح می‌دهد. برای مثال یک هدف تجاری ممکن است «افزایش فروش به مشتری های موجود» باشد؛ در حالی که یک هدف داده کاوی ممکن است «پیش بینی تعداد جنسی که یک مشتری خرید خواهد کرد»، «برآورد درآمد یک سال خاص» یا «ارتباط آمار نفوس و قیمت یک کالا» باشد.

خروجی: اهداف داده کاوی

خروجی های قابل توسعه پروژه را که قابلیت دستیابی به اهداف پروژه دارند مشخص نمائید. به کار بستن این تکنیک در خروجی ها رایج است.

فعالیت ها

- مسئله تجاری را به اهداف داده کاوی تبدیل کنید. (بطور مثال یک کمپین بازاریابی نیاز به تقسیم

بندی مشتریان دارد تا درباره افرادی که از آن کمپین استقبال می نمایند تصمیم گیری نماید؛ ضمن آنکه اندازه بخش های تقسیم بندی شده باید مشخص باشد)

- نوع مسئله داده کاوی را تعیین کنید. (به عنوان مثال آیا مسئله طبقه بندی است یا پیش بینی یا خوشه بندی) برای مطالعه بیشتر در این زمینه به پیوست ۲ مراجعه کنید.

پیشنهاد!

در مواقع خاصی بهتر است تعریف دوباره ای از مسأله ارائه دهیم. به عنوان مثال، مدلسازی بقای محصول نسبت به بقای مشتری (زمانی که هدف گذاری بقای مشتری به علت زمان بر بودن مشاهده اثر آن بر نتایج تأثیر اندکی بر خروجی داشته باشد)، در اولویت قرار دارد.

خروجی: معیارهای موفقیت داده کاوی

با بیانی فنی معیارهایی برای موفقیت پروژه تعریف کنید؛ برای مثال سطح معینی برای دقت پیش بینی، یا درجه معینی از صعود[۳۰] در نمودار تمایل به خرید. گاهی اوقات علاوه بر یک معیار تجارت موفق توضیح شرایط غیرعینی هم لازم است تا امکان قضاوت ذهنی فراهم شود.

فعالیت ها

- معیارهایی برای ارزشیابی مدل تعیین کنید. (بطور مثال: دقت مدل، کارایی و پیچیدگی).
- مبنایی برای معیارهای ارزشیابی معین کنید.
- معیاری برای ارزشیابی های ذهنی تعیین کنید. (بطور مثال: توانایی تفسیرپذیری مدل و قابلیت آن در درک مسائل بازاریابی)

توجه!

فراموش نکنید معیارهای موفقیت داده کاوی با معیارهای موفقیت تجارت متفاوت است. همچنین توصیه می‌شود طرح گسترش و استقرار پروژه از همان ابتدا تهیه شود.

۳۰ Lift

۴-۱ ارائه طرح پروژه

وظیفه: ارائه طرح پروژه

طرح مورد نظر برای رسیدن به اهداف داده کاوی و به تبع آن رسیدن به اهداف تجاری را تشریح کنید

خروجی: طرح پروژه

فهرستی از مراحل مختلف پروژه همراه با مدت زمان لازم، منابع مورد نیاز، ورودی ها، خروجی ها و وابستگی های آن‌ها تهیه کنید. تا حد امکان، لزوم تکرارها در حین انجام پروژه را تصریح کنید. به عنوان مثال تکرار در فاز های مدلسازی و ارزیابی. قسمت مهمی از طرح پروژه، آنالیز زمانبندی پروژه و ریسک هاست. در طرح پروژه نتایج پروژه را همراه با پیشنهادی برای عملکرد مناسب در هنگام بروز ریسک صراحتاً مشخص کنید.

اگر چه این بخش تنها قسمتی است که در آن مستقیماً از عنوان طرح پروژه یاد می‌شود، با این وجود پروژه بایستی دائماً مورد بازبینی قرار گیرد و درباره آن مشورت شود. این کار باید در شروع هر قسمت یا شروع هر اقدامی صورت گیرد.

فعالیت ها

- قسمت ابتدایی طرح پروژه را مشخص کنید و توضیح دهید آیا این قسمت با درگیر کردن تمام کارکنان قابلیت اجرایی دارد؟
- همه اهداف انتخاب شده را وارد کنید و تکنیک ها و شیوه ای انتخاب کنید که مسئله را حل کند و با معیارهای تجارت موفق مطابقت داشته باشد.
- نیرو و منابع مورد نیاز را برای دستیابی به راه حل و توسعه آن برآورد کنید. (این امر می تواند برآورد مقیاس زمانی پروژه داده کاوی را برای دیگر متخصصان تسهیل کند. برای مثال، اغلب ۵۰ تا ۷۰ درصد از وقت و نیرو را برای مرحله آماده سازی داده‌ها، ۲۰ تا ۳۰ درصد را برای مرحله درک داده، ۱۰ تا ۲۰ درصد برای مدلسازی،ارزیابی وفهم تجاری و ۵تا ۱۰ درصد را مرحله گسترش و استقرار دربر می گیرد).
- مراحل حساس را مشخص کنید.
- نقاط تصمیم گیری را معلوم کنید.
- نقاط تجدید نظر را معلوم کنید.
- تکرارهای عمده را مشخص کنید.

خروجی: ارزیابی اولیه از ابزارها و تکنیک ها

درپایان مرحله اول، فقط یک ارزشیابی اولیه از ابزار و تکنیک ها شکل گرفته است. در اینجا یک ابزار داده کاوی که برای روش های متنوع مراحل مختلف داده کاوی قابل استفاده باشد انتخاب کنید. ارزیابی ابزارها و تکنیک ها از همان لحظه انتخاب آن‌ها مهم است و می تواند در کل پروژه تأثیرگذار باشد.

فعالیت ها:

- فهرستی از معیارهای انتخاب ابزار و تکنیک ها تهیه کنید (یا از یک لیست قبلاً تهیه شده استفاده کنید)
- ابزارها و تکنیک های بالقوه را انتخاب کنید.
- تناسب تکنیک های موجود را ارزیابی کنید.
- تکنیک های اجرایی را با توجه به ارزیابی راه حلها بررسی و رتبه بندی کنید.

۲. شناسایی و درک داده ها

۲-۱ جمع آوری اولیه داده

وظیفه: جمع آوری اولیه داده

به دست آوردن داده‌ها در یک پروژه (یا دستیابی به داده‌ها)، در منابع پروژه لیست می‌شود. اگر برای فاز درک داده لازم باشد این جمع آوری اولیه شامل بارگذاری[۳۱] داده می‌شود. به عنوان مثال، در صورت استفاده از یک ابزار خاص برای درک داده، بهتر است داده‌هایتان را در این ابزار بارگذاری کنید.

خروجی: گزارش جمع آوری اولیه داده ها

فهرستی از مجموعه داده (یا مجموعه داده‌های) قابل دستیابی در پروژه به همراه جزئیات مورد نیاز برای این کار تهیه کنید. این گزارش بایستی مشخص کننده متغیرهایی که دارای اهمیت بیشتری هستند باشد. به خاطر داشته باشید که ارزیابی کیفی مجموعه داده‌ها نباید بر اساس منابع جداگانه باشد. الحاق مجموعه داده‌ها به یکدیگر باعث کشف اشکالاتی خواهد شد که در منابع اصلی داده‌ها نیست و این امر ناشی از تفاوت بین منابع است.

فعالیت ها:

برنامه ریزی برای داده های مورد نیاز

- برنامه ریزی برای اطلاعاتی که مورد نیاز است. (بطور مثال: آیا صفت[۳۲] های فعلی کافیست، یا اطلاعات اضافی دیگر هم مورد نیاز است؟)

---

۳۱ Loading

۳۲ Attribute

- بررسی میزان دسترس پذیری اطلاعات مورد نیاز.

معیارهای انتخاب

- مشخص کردن معیارهای انتخاب (کدام صفت ها برای اهداف تعیین شده داده کاوی مورد نیاز است؟ کدام صفت ها بی ارتباط شناخته شده اند؟ با توجه به تکنیک های انتخاب شده چند ویژگی (صفت) نیاز ما را برطرف می‌کند؟)
- انتخاب جداول و فایل های قابل توجه
- انتخاب داده‌ها از جداول و فایل ها
- بررسی اینکه تحلیل چه مدت زمانی از داده‌ها مورد نیاز است؟ (بطور مثال: داده‌های ۱۸ ماه موجود است اما با داده‌های ۱۲ ماه نیاز ما برطرف می‌شود.)

توجه!

هنگام جمع آوری داده‌ها از منابع مختلف، ممکن است در الحاق آن‌ها با مشکل مواجه شویم (ناسازگاری فرمت ها، بی اعتباری داده‌ها و ...)

الحاق داده ها

- اگر داده‌ها دارای متن هستند آیا برای مدل نیاز به تبدیل کد وجود دارد یا بایستی آن‌ها را گروه بندی نمود؟
- چگونه می توان متغیرهای گم شده را به دست آورد؟
- چگونه می توان داده‌ها را استخراج کرد؟

پیشنهاد!

برخی داده‌ها از منابع غیر الکترونیکی به دست آمده اند. مانند دفاتر ثبتی و ... .

علاوه بر آن گاهی اوقات داده‌ها باید پیش پردازش شوند. (سری های زمانی، میانگین وزنی و ...)

۲-۲ توصیف داده‌ها

وظیفه: توصیف داده‌ها

پس از بررسی های لازم، ناخالصی داده‌های به دست آمده را گزارش کنید.

خروجی: گزارش توصیف داده‌ها

داده‌های بدست آمده را توصیف کنید. این توصیف شامل فرمت داده‌ها، ابعاد داده‌ها (تعداد رکوردها و فیلدهای هر جدول) به همراه معرفی و شناسایی هر یک از فیلدها می‌باشد.

فعالیت ها:

سنجش حجمی داده‌ها

- شناسایی داده‌ها و روشهای جمع آوری
- دستیابی به پایگاه داده
- استفاده از آنالیز آماری (در صورت نیاز)
- توضیح جداول و روابط حاکم بر آن‌ها
- بررسی حجم داده‌ها، تعداد ابعاد و نوع ترکیب ها
- و بالاخره اینکه آیا داده‌ها شامل متن هستند یا خیر؟

نوع صفت ها و مقادیر آن‌ها

- بررسی صفت های موجود و میزان دسترسی به آن‌ها
- بررسی نوع صفت ها (عددی، اسمی، ترتیبی و ...)
- بررسی دامنه صفت ها
- تحلیل همبستگی صفت ها
- درک مفهوم و ماهیت هر صفت و مقادیر آن در حوزه کسب و کار
- محاسبات آماری پایه برای هر صفت (محاسبه توزیع داده‌ها، میانگین، ماکزیمم، مینیمم، انحراف معیار، واریانس، مد، چولگی و...)
- آنالیزهای آماری پایه ای و بیان مفهوم نتایج آن در حوزه کسب و کار
- بررسی مناسبت هر صفت با هدف خاص داده کاوی
- آیا ماهیت هر یک از صفت ها با هدف پروژه سازگار می باشند؟
- بررسی نظرات متخصصین مربوطه در مورد نقش هریک از صفت ها
- آیا داده‌ها نیازمند متوازن سازی[۳۳] می باشند؟ (بسته به مدل مورد استفاده)

کلیدها

- تجزیه و تحلیل روابط کلیدی
- بررسی میزان اشتراک بین صفت های کلیدی در جدول

بررسی فرضیات و اهداف

- درصورت نیاز، بروزرسانی لیست فرضیات انجام شود.

---

۳۳ Balance

۲-۳ کاوش داده ها

وظیفه: کاوش داده ها

این وظیفه شـامل ابزارهای پرس و جو، مجسـم سـازی روابط و گزارش دهی می‌باشـد. این وظایف بـه واسـطه توضیحاتی در ارتباط با داده‌ها، گـزارش های کیفی، تبدیل ها و تکنیک های آماده سـازی و تصحیح آن‌ها، نقش مهمی در نزدیک شدن به اهداف داده کاوی دارد.

خروجی: گزارش کاوش داده ها

نتایج این مرحله، شامل رابطه های کشف شده و فرضیات اولیه و تأثیر آن بر باقیمانده پروژه را بررسی کنید. در صورت لزوم از رسـم نمودار که تعیین کننده خصوصیات داده‌هاسـت و موجد زیر مجموعه جالب توجهی از داده‌ها که امکان بررسی بیشتر را فراهم می‌کند استفاده کنید.

فعالیت ها:

کاوش داده ها

- تجزیه و تحلیل ویژگی های جالب صفت ها با جزئیات کامل (بطور مثال: آماره های مقدماتی، زیر گروه های قابل توجه)
- شناسایی خواص زیر گروه ها

شکل دهی فرضیات برای آنالیزهای بعدی

- ارزیابی اطلاعات و کاوش در گزارش تشریحی داده‌ها
- شکل دهی فرضیات اولیه و تعیین عملیات مورد نیاز
- (در صورت امکان) تبدیل فرضیات به یک هدف داده کاوی
- اهداف داده کاوی را به صورت دقیق و روشن تنظیم کنید. جستجوی کورکورانه لزوماً بی فایده نیست اما یک جستجوی هدایت شده به سمت اهداف داده کاوی ارجحیت دارد.
- اجرای آنالیزهای مقدماتی برای اعتبارسنجی فرضیات

۲-۴ بررسی کیفیت داده ها

وظیفه: بررسی کیفیت داده ها

کیفیت داده‌ها با را با سـؤالات زیر بررسـی کنید: آیا داده‌ها کامل هسـتند؟ (همه موارد مورد نیاز تحت پوشش قرار میگیرند؟) آیا آن‌ها صحیح اند یا خطا دارند؟ و اگر خطا دارند، خطای آن‌ها چقدر شایع است؟ آیا در داده‌ها مقادیر گمشـده وجود دارد؟ اگر هسـت چگونه نمایش داده می‌شوند، کجا اتفاق افتاده اند و چقدر شایع اند؟

خروجی: گزارش کیفیت داده ها

فهرستی از نتایج بررسی کیفیت داده‌ها تهیه کنید؛ اگر مشکلی وجود داشت فهرستی هم برای راه حل های ممکن تهیه کنید.

فعالیت ها:

- مقادیر خاص و متمایز را شناسایی کرده و مفهوم آن‌ها را مشخص نمائید.

مرور کلیدها و متغیرها

- بررسی شمولیت داده‌ها (آیا تمامی مقادیر ممکن در مجموعه داده‌ها وجود دارد؟)
- بررسی کلیدها
- آیا مفهوم هریک از صفت ها و مقادیر آن‌ها همخوانی دارند؟
- شناسایی صفت های گمشده و خالی
- درک مفهوم داده‌های گمشده
- بررسی صفت هایی که مفهوم یکسانی دارند ولی دارای مقادیر مختلفی می باشند. (به طور مثال: رژیم غذایی و میزان چربی غذا)
- بررسی املای مقادیر (ممکن است مقادیر یکی باشد اما حرف شروع آن‌ها یکسان نباشد، مثلاً با حروف کوچک شروع شده باشد یا حروف بزرگ)
- بررسی انحرافات و تصمیم گیری در مورد اینکه انحراف موجود، نشان دهنده اختلال است یا نشانه یک پدیده قابل توجه!
- بررسی قابل قبول بودن مقادیر فیلدها (به عنوان مثال: تمام فیلدها مقادیر همسان یا نزدیک به هم داشته باشند)

پیشنهاد!

پس از بررسی صفت ها، معقول بودن مقادیر آن‌ها را بررسی کنید. (مثلاً یک نوجوان با درآمد بالا). با استفاده از نمودارها، هیستوگرام و... ناسازگاری ها را آشکار کنید.

کیفیت داده ها در فایل های Flat

- جداکننده[۳۴] استفاده شده در فایل را بررسی کنید. آیا بین همه صفت ها از جداکننده یکسان استفاده شده است؟
- تعداد صفت ها را در هر رکورد بررسی کنید. آیا تناقضی وجود دارد؟

---

۳۴ Delimiter

اختلال[۳۵] و ناسازگاری بین منابع

- میزان سازگاری و همچنین وجود اطلاعات تکراری بین منابع مختلف را بررسی کنید.
- طرحی برای مواجهه با اختلال ها ترتیب دهید.
- نوع اختلال و همچنین متغیرهای تأثیر گرفته از آن را پیدا کنید.

پیشنهاد!

گاهی لازم است داده‌هایی که رفتارهای مشخصی ندارند از سایر داده‌ها جدا شوند و مورد استفاده قرار نگیرند. (بطور مثال: برای بررسی رفتار وام گیرندگان، حذف اطلاعات کسانی که وامی دریافت نکرده اند و تمام کسانی که هنوز زمان سررسید وامشان نرسیده است، نوعی جداسازی داده‌های با رفتارهای نامشخص است)

قابل قبول بودن فرضیات و همچنین قابلیت آن‌ها برای به دست آوردن اطلاعات کافی را بررسی کنید.

۳. آماده سازی داده ها

خروجی: مجموعه داده

مجموعه داده‌ها محصول فاز آماده سازی داده است، که برای مدلسازی یا سایر تجزیه و تحلیل ها استفاده خواهد شد.

خروجی: توصیف مجموعه داده

مجموعه داده ای را که برای مدلسازی یا تجزیه و تحلیل های اصلی پروژه استفاده خواهند شد، تشریح کنید.

۳-۱ انتخاب داده ها

وظیفه: انتخاب داده ها

داده‌هایی را که برای تجزیه و تحلیل استفاده می‌شوند انتخاب کنید. معیارها شامل مرتبط بودن با اهداف داده کاوی، محدودیت های فنی و کیفی از قبیل محدودیت درحجم داده‌ها یا نوع داده‌هاست.

خروجی: دلیل انتخاب یا عدم انتخاب داده ها

فهرستی از داده‌های انتخاب شده یا حذف شده را به همراه دلیل انتخاب یا عدم انتخاب هر یک از آن‌ها تهیه کنید.

---

۳۵ Noise

فعالیت ها:

- داده‌های مناسب بیشتری را جمع آوری کنید. (از سایر منابع – داخلی و خارجی)
- برای انتخاب فیلدهای مناسب، تست های همبستگی و معنا داری را انجام دهید.
- معیارهای انتخاب داده را، با توجه به تجربه مرحله کیفیت داده‌ها، بازبینی کنید (وظیفه ۲-۱ را ببینید؛ چون ممکن است بخواهید حذف یا اضافه ای در داده‌ها اعمال کنید.)
- معیارهای انتخاب داده را با توجه به نتایج مدلسازی بازبینی کنید. (وظیفه ۲-۱ را ببینید؛ ممکن است ارزیابی مدل نشان دهنده نیاز به مجموعه دیگری از داده‌ها باشد.)
- زیـر مجموعه هـای مختلفی از داده‌ها انتخاب کنیـد. (بطور مثال: صفت هـای مختلف، داده‌های با شرایط خاص)
- استفاده از تکنیک های نمونه گیری را در نظر داشته باشید (بطور مثال: یک راه حل سریع ممکن است شامل جداسازی مجموعه داده‌ها به دو بخش آموزشی و آزمایشی می‌باشد، و یا در مواقعی که نرم افزار قادر به استفاده از تمامی داده‌ها نیست، می توان کاهش حجم داده‌های آزمایشی را مد نظر داشت. همچنین ممکن است نمونه گیری بصورت وزنی برای متغیرهایی با درجه اهمیت متفاوت انجام گردد و یا در داخل متغیرهای همسان، مقادیر دارای وزن های متفاوت باشند.)
- دلایل حذف و اضافه را ثبت کنید.
- تکنیک های موجود نمونه گیری را بررسی کنید.

پیشنهاد!

بر اساس معیارهای انتخاب داده، مهم‌ترین متغیر یا متغیرها را شناسایی کنید و با قرار دادن آن متغیر به عنوان مبنا، سایر متغیرها را وزن دهی کنید.

۳-۲ پاکسازی داده ها

وظیفه: پاکسازی داده ها

کیفیت داده‌ها را تا سـطح مورد نیاز تکنیک های تجزیه و تحلیل انتخاب شده افزایش دهید. این کار ممکن اسـت شامل انتخاب زیرمجموعه ای بی نقص از داده‌ها، اضافه کردن مقادیر قراردادی مناسب یا تکنیک های جاه طلبانه تری مانند برآورد مقادیر گمشده با استفاده از مدلسازی باشد.

خروجی: گزارش پاکسازی داده ها

چگونگـی اجـرای اقدامات و تصمیماتی که تصمیـم به انجام آن‌ها گرفته بودید را شـرح دهید. (این اقدامـات و تصمیمـات در گزارش مشـکلات کیفی داده‌ها در جریان وظیفه بازرسـی کیفیـت داده‌ها در فاز

شناسـایی و درک داده‌هـا ثبت شـده بودند) گزارش تهیه شـده بایسـتی به شـاخص هـای اصلی کیفیت داده‌های مورد استفاده در پروژه داده کاوی اشاره نموده و تاثیرات احتمالی آن بر نتایج را بیان نماید.

فعالیت ها:

- عملکرد خود در مواجهه با انواع اختلال ها را بازبینی کنید.
- در مواجهه با اختلال ها یکی از این سه راهکار را بر گزینید: تصحیح، حذف، یا نادیده گرفتن
- در خصـوص مواجهـه با مقادیر خاص و مبهـم تصمیم گیری کنید. دامنه این مقادیـر می تواند باعث نتایج عجیبی شود و باید به دقت آن‌ها را امتحان کرد. مقادیر خاص می تواند ناشی از نتایج بررسی هایی باشـد که در آن بعضی از سـؤالات یا پرسیده نشده باشـند یا جوابی نگرفته باشند. این مسئله ممکن است ناشـی از ثبت مقدار ۹۹ برای داده‌های ناشـناخته باشـد. برای مثال ۹۹ را برای وضعیـت تأهل یا گرایش سیاسـی ناشـناخته در نظر بگیریم. همچنین مقادیر خاص می تواند برخاسـته از داده‌های ناقص و بریده شـده باشـد، برای مثال «۰۰» برای افراد صد سـاله یا برای همه ماشـین هایی که بیش از ۱۰۰۰۰۰ کیلومتر مسافت طی نموده اند.
- معیارهـای انتخـاب داده را با توجه به تجربه مرحله پاک سـازی داده‌ها، بازبینی کنید. (ممکن اسـت مجموعه داده ای را اضافه یا حذف کنیم)

پیشنهاد!

گاهی اوقات برخی صفت ها ارتباط چندانی با اهداف داده کاوی ندارند. در این صورت اختلال در آن‌ها بی معنی است. اما در صورت تغییر شرایط، این اختلال ها نیز بایستی لحاظ شوند.

۳-۳ ساخت داده ها

وظیفه: ساخت داده ها

این وظیفه شـامل عملیات آماده سـازی برای ایجاد داده‌های مفید می‌باشـد، از قبیل سـاخت صفت های اشتقاقی، ورود رکوردهای جدید دست نخورده یا تبدیل مقادیر صفت های موجود.

فعالیت ها:

- مکانیزم ترکیب های قابل حصول را با توجه به لیست ابزارهای[۳۶] پیشنهاد شده در پروژه بررسی کنید.
- در مورد انجام ترکیب در داخل ابزار یا خارج از آن تصمیم گیری کنید. (کدام یک مؤثرتر است؟)
- معیارهـای انتخـاب داده را با توجه به تجربیـات مرحله ترکیب داده‌ها، بازنگری کنید. (ممکن اسـت بخواهیم مجموعه داده ای را اضافه یا حذف کنیم)

---

۳۶ Tools

خروجی: صفت های اشتقاقی

صفت های اشـتقاقی صفت های جدیدی هسـتند که از یک یا تعداد بیشتری از صفت های موجود با رکوردهای یکسان ساخته می‌شوند: مثلاً مساحت=طول×عرض.

چرا نیاز به ایجاد صفت های ترکیبی در طی پروژه داده کاوی وجود دارد؟ بایستی توجه نمود که برای ساخت مدل ضرورتی برای استفاده بدون تغییر از داده‌های خام اولیه وجود ندارد. دلایل ایجاد صفت های ترکیبی، از این قرارند:

- تجربه نشان داده است در بسیاری از مواقع، فاکتورهای بسیار مهمی وجود دارد که بایستی در ساخت مدل وجود داشته باشد، ولی در لیست داده‌های اولیه نیستند.
- الگوریتم های مورد اسـتفاده انواع خاصی از داده‌ها را اسـتفاده می کنند؛ برای مثال هنگام اسـتفاده از رگرسیون خطی به داده‌های خاصی نیاز داریم که باید با ترکیب سایر داده‌ها ایجاد شوند.
- ممکن است خروجی مرحله مدلسازی نشان دهد فاکتورهای مهمی در نظر گرفته نشده است.

فعالیت ها:

صفت های اشتقاقی

- در مورد متغیرهایی که بایستی نرمالیزه شوند، تصمیم گیری کنید. (مثلاً وقتی از الگوریتم خوشه بندی برای سن و درآمد استفاده می‌شود، درآمد غالب می گردد)
- نکات لازم در مورد اهمیت ارتباط متغیرها با متغیرهای جدید اضافه شده، را ذکر کنید. (بطور مثال: وزن ها و وزن های استاندارد شده)
- چگونه می توان صفت های گمشده را بسازیم یا وارد کنیم؟ (از کدام نوع ترکیب استفاده کنیم: تجمیع، میانگین، استقرا)
- اضافه کردن صفت های جدید برای دستیابی به داده‌ها.

پیشنهاد!

پیـش از اضافـه کردن صفـت ترکیبی، صفت هایی که پروسـه مدل کردن را راحت تـر می کنند، یا به الگوریتم ها کمک می کنند را مشـخص کنید. شـاید اسـتفاده از صفت «درآمد سرانه» نسبت به استفاده از صفت «درآمد خانوار» بهتر یا راحت تر باشد.

از دیگـر انواع صفت های ترکیبی، تبدیلات تک متغیری اسـت، که معمولاً بـرای تأمین نیازهای ابزار انتخاب شده انجام می گیرد.

فعالیت ها:

تبدیلات تک متغیری

- تعییـن مراحل تبدیـلات مورد نیاز با توجه به امکان پذیری آن‌ها. (بطور مثال: تغییر متغیرهای عددی به گسسته)
- انجام مراحل تعیین شده.

راهنمایی!

تبدیـل ها می توانند شـامل تغییـر داده‌های عددی به مقادیرنمادین[۳۷] باشـند (مثلاً تبدیل سـن به مقاطع سنی) و یا تبدیل فیلدهای نمادین (عالی، خوب، متوسط، بد) به مقادیر عددی. ابزار مدلسازی یا الگوریتم ها معمولاً به این موارد نیاز دارند.

خروجی: تولید رکورد

رکوردهـای تولیـدی رکوردهـای کاملاً جدیدی هسـتند، کـه اطلاعات جدیـدی را اضافه مـی کنند یا داده‌های جدیدی که به صورت دیگری قابل شناسایی نیستند با آن‌ها نمایش داده می‌شود. (برای مثال در داده‌های بخش بندی شـده، تولید و اسـتفاده از رکوردهای جدیدی که اشاره به نمونه های اولیه[۳۸] هر بخش داشته باشد، در پردازش های آتی می تواند مفید باشد.)

فعالیت:

تکنیک های موجود را در صورت نیاز بررسـی کنید. (مثلاً مکانیسـم ساخت نمونه های اولیه برای هر بخش از داده‌های بخش بندی شده)

۳-۴ یکپارچه سازی داده ها

وظیفه: یکپارچه سازی داده ها

روش هایی وجود دارد که به وسیله آن‌ها اطلاعات از چندین جدول یا منابع اطلاعاتی مختلف ترکیب می‌شوند و رکوردها یا مقادیر جدید می سازند.

خروجی: ادغام داده ها

ادغام جدول یعنی دو یا چند جدول را که حاوی اطلاعات متفاوتی در مورد موضوع واحدی هستند به هم الحاق کنیم. در این مرحله ممکن اسـت لازم باشـد رکورد جدیدی تولید کنیم. همچنین ممکن است

---

۳۷ Symbolic

۳۸ prototypes

تولید مقادیر تجمیعی[۳۹] نیز در دستور کار قرار گیرد.

تجمیع داده‌ها عبارت است از عملیاتی که مقادیر جدید را به وسیله خلاصه کردن اطلاعات چندین رکورد یا جدول محاسبه می‌کند.

فعالیت ها:

- در صورت نیاز امکان یکپارچه سازی و موثر بودن الحاق منابع ورودی را بررسی کنید.
- منابع داده ای مختلف را یکپارچه کنید.
- معیارهای انتخاب داده را، با توجه به مرحله یکپارچه سازی داده‌ها، بازبینی کنید. (ممکن است مجموعه داده ای را اضافه یا حذف کنیم)

پیشنهاد!

به خاطر داشته باشید برخی اطلاعات دارای فرمت غیر الکترونیکی می باشند.

۳-۵ فرمت داده ها

وظیفه: فرمت داده ها

تغییر فرمت داده‌ها شامل اصلاحات نحوی است و در آن ارزش و مفهوم داده‌ها تغییر نمی‌کند. این کار در صورت لزوم برای ابزارهای مدلسازی استفاده می‌شود.

خروجی: فرمت بندی مجدد داده ها

بعضی نرم افزارها، شرایطی برای متغیرها دارند، مثلا محتوای اولین فیلد، شناسه های یکتا برای هر سطر از داده‌ها باشد، یا محتوای آخرین فیلد می‌بایستی متغیر خروجی (هدف) باشد که به وسیله مدل پیش بینی می‌شود.

فعالیت ها:

بازآرایی فیلدها

بعضی نرم افزارها، شرایطی برای متغیرها دارند، مثلا محتوای اولین فیلد، شناسه های یکتا برای هر سطر از داده‌ها باشد، یا محتوای آخرین فیلد می‌بایستی متغیر خروجی (هدف) باشد که به وسیله مدل پیش بینی می‌شود.

بازآرایی رکوردها

بعضی مواقع تغییر در ترتیب رکوردها در مجموعه داده‌ها مهم می‌باشد. ممکن است ابزار مدلسازی به مرتب سازی رکوردها بر اساس مقادیر خروجی نیاز داشته باشد.

---

۳۹ Aggregate

فرمت بندی مجدد مقادیر داخلی

- واضح است که هر تغییر خاص، نیازمندی های ابزار مدلسازی خاصی را تأمین می‌کند.
- معیارهای انتخاب داده را، با توجه به مرحله پاک سـازی داده‌ها بازبینی کنید. (ممکن است مجموعه داده ای را اضافه یا حذف کنیم)

۴. مدلسازی

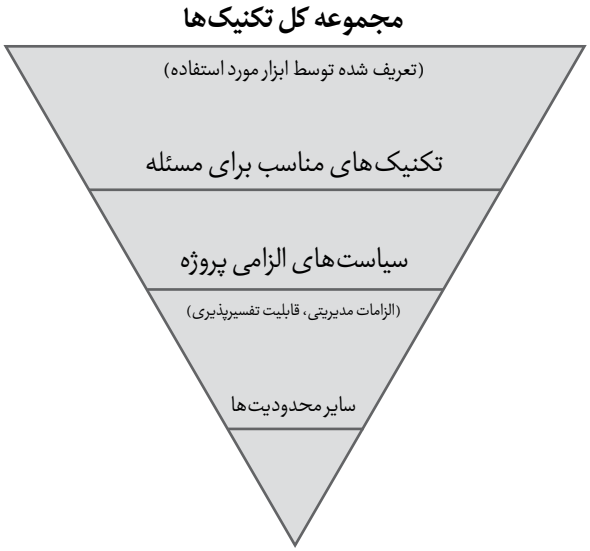
۴-۱ انتخاب تکنیک مدلسازی

وظیفه: انتخاب تکنیک مدلسازی

نخسـتین گام در مدلسـازی، انتخاب تکنیکی اسـت که می خواهیم مورد اسـتفاده قرار دهیم. اگر از چندین تکنیک استفاده شود برای هر یک از تکنیک ها این وظیفه را جداگانه مورد استفاده قرار دهید.

فراموش نکنید همه تکنیک ها و ابزارها در همه قسـمت ها مورد اسـتفاده نیسـتند. برای هر مسـئله خاص فقط تعداد محدودی از تکنیک ها مورد محاسبه قرار می گیرد. (ضمیمه ۲ به بررسی تکنیک های مناسـب برای انواع خاص مسـائل داده کاوی می پردازد) سیاسـت های[۴۰] موجود در پروژه (شامل الزامات مدیریتی، قابلیت تفسـیرپذیری نتایج و ...) به همراه سـایر محدودیت ها نیز با تنگتر نمودن این حلقه، امکان انتخاب درست را به تحلیلگر می دهند. این امکان وجود خواهد داشت تا فقط یک ابزار یا تکنیک برای حل مسئله در اختیار باشد و یا حتی ابزار قابل دسترس دارای بهترین نتیجه از نقطه نظر تکنیک های مورد نیاز نباشد.

۴۰ Political Requirements



شکل ۱۰: روند انتخاب تکنیک های مناسب

خروجی: تکنیک های مدلسازی

مدل هایی که تاکنون مورد استفاده قرار گرفته اند را ثبت کنید.

فعالیت:

تکنیک هایی را که برای مدلسازی از آن‌ها استفاده کرده اید یادداشت کنید.

خروجی: فرضیات مدلسازی

در بسیاری از تکنیک های مدلسازی فرض هایی در مورد داده‌ها صورت می گیرد.

فعالیت ها:

- فرضیاتی که برای ساخت مدل در خصوص داده‌ها در نظر گرفته شده است را تعریف نمائید. (کیفیت، فرمت و توزیع داده‌ها)
- فرضیات مورد استفاده را با فرض هایی که در گزارش شرح داده‌ها آمده، مقایسه کنید.
- از برقراری فرضیات مورد استفاده در داده‌ها اطمینان حاصل کنید و در صورت لزوم به مرحله آماده سازی داده‌ها بازگردید.

۴-۲ ایجاد طرح آزمون

وظیفه: ایجاد طرح آزمون

پیش از ساختن یک مدل، باید روشی برای سنجش کیفیت و اعتبار مدل ایجاد کنیم. به عنوان

مثال در وظایف مربوط به داده کاوی نظارتی، مانند طبقه بندی، از نرخ خطا برای سنجش کیفیت پیش بینی ها در مدل های داده کاوی استفاده می‌شود. بنابراین معمولاً داده‌ها را به مجموعه های آموزشی و آزمایشی تقسیم می کنیم و پس از ایجاد مدل روی داده‌های آموزشی، کیفیت آن را روی مجموعه جدا شده آزمایشی مورد بررسی قرار می دهیم.

خروجی: طرح آزمون

طرح انتخاب شده برای آموزش، آزمون و ارزیابی را تشریح کنید. یکی از بخش های ابتدایی طرح، نحوه تقسیم بندی مجموعه داده به داده‌های آموزش و آزمایشی، و ارزیابی این مجموعه هاست.

فعالیت ها:

- طرح های آزمون موجود را به صورت جداگانه برای هر هدف داده کاوی بررسی کنید.
- برای مراحل مورد نیاز تصمیم گیری کنید. (تعداد تکرارها، تعداد فیلدها و ...)
- تغییرات مورد نیاز داده‌ها را برای آزموده شدن مهیا کنید.

۳-۴ ساخت مدل

وظیفه: ساخت مدل

با استفاده از نرم افزارهای داده کاوی، برای مجموعه داده‌های آماده شده یک یا چند مدل بسازید.

خروجی: تنظیم پارامترها

در هر نرم افزار مدلسازی، عموماً پارامترهایی وجود دارند که باید تنظیم شوند. فهرستی از پارامترها و مقادیر انتخاب شده برای آن‌ها همراه با دلیل انتخاب صورت گرفته، تهیه کنید.

فعالیت ها:

- پارامترهای اولیه را تنظیم کنید.
- دلایل انتخاب مقادیر را مستند کنید.

خروجی: مدل ها

با استفاده از نرم افزارهای داده کاوی، برای مجموعه داده‌های آماده شده یک یا چند مدل بسازید.

فعالیت ها:

- تکنیک های انتخاب شده را روی داده‌های ورودی اجرا کنید و مدل بسازید.
- نتایج داده کاوی را پردازش کنید. (بطور مثال: ویرایش قوانین و نمایش درختی)

خروجی: توصیف مدل

مدل حاصله را تشریح کنید. گزارشی درباره تفسیر مدل ها تهیه کنید و اشکالات مدل و مفهوم آن‌ها را یادداشت کنید.

فعالیت ها:

- ویژگی های مدل موجود را شرح دهید، چون ممکن است در آینده مورد استفاده قرار گیرد.
- تنظیمات مربوط به هریک از پارامترهای مدل را ثبت کنید.
- جزئیات مدل و هریک از ویژگی خاص آن را به صورت دقیق تشریح کنید.
- فهرستی برای مدل های قاعده مند تهیه کنید که شامل قواعد مدل و ارزیابی هریک از قوانین بر اساس میزان پوشش و صحت باشد.
- برای مدل های بی قاعده، هر گونه اطلاعات تکنیکی در مورد مدل (مانند توپولوژی شبکه عصبی) و دیگر ویژگی های ایجاد شده به وسیله فرآیند مدلسازی (مانند صحت و حساسیت) را فهرست کنید.
- رفتارهای مدل و تفسیر آن‌ها را شرح دهید.
- نتیجه مربوط به الگوها را بیان کنید (در صورت وجود)، گاهی اوقات مدل بدون انجام پروسه ارزیابی، ویژگی های مهمی در مورد داده‌ها بیان می‌کند. (بطور مثال: خروجی مدل، با یکی از ورودی ها برابر یا همسان است)

۴-۴ ارزیابی مدل

وظیفه: ارزیابی مدل

ارزیابی به معنای بررسی تناسب مدل با معیار های موفقیت داده کاوی است. این بررسی یک ارزیابی کاملاً تکنیکی است که بر مبنای خروجی های وظیفه مدلسازی صورت می گیرد.

خروجی: ارزیابی مدل

نتایج این بخش را خلاصه کنید. فهرستی از کیفیت مدل های تولید شده (مثلاً بر حسب دقت) تهیه کنید و آن را رتبه بندی کنید.

فعالیت ها:

- نتایج را با توجه به معیارهای ارزیابی بررسی کنید.
- با یک استراتژی مناسب نتایج را آزمایش کنید. (بطور مثال: ایجاد مجموعه داده آموزشی و آزمایشی، ارزیابی متقاطع[۴۱]، بوت استراپ[۴۲] و ...)
- نتایج ارزیابی و تفسیر آن‌ها را مقایسه کنید.
- با توجه به معیارهای ارزیابی و موفقیت نتایج را رتبه بندی کنید.
- بهترین مدل را انتخاب کنید.

---

۴۱ Cross-Validation

۴۲ Bootstrapping

- نتایج را از نگاه تجاری تفسیر کنید. (تا حدی که در این مرحله ممکن است)
- به کمک متخصصین حوزه کسب و کار و داده‌ها، توضیحاتی به مدل ضمیمه کنید.
- معقول بودن مدل را بررسی کنید.
- هم جهت بودن آن با اهداف داده کاوی را بررسی کنید.
- بدیع و مفید بودن اطلاعات اکتشافی حاصل از مدل را با توجه به دانش موجود بررسی نمائید.
- قابلیت اطمینان مدل را بررسی کنید.
- قابلیت گسترش و استقرار هریک از نتایج را تجزیه و تحلیل کنید.
- در صورت وجود توضیحات شفاهی در مورد ایجاد مدل (ارزیابی قواعد، منطقی بودن آن‌ها، عملی بودن آن‌ها، تعداد آن‌ها،و...) آن‌ها را بیان کنید.
- نتایج را بطور کلی ارزیابی کنید.
- در خصوص چرایی تاثیر مثبت یا منفی یک تکنیک یا پارامتر بر نتایج مدل، بینش مناسبی ایجاد نمائید.

پیشنهاد!

برای تعیین میزان خوب بودن قدرت مدل ها در پیش بینی، ایجاد "جداول صعود"[۴۳] یا "جداول بهره"[۴۴] می تواند کمک کننده باشد.

خروجی: بازبینی پارامترها

با توجه به ارزیابی مدل، تنظیمات پارامترها بازبینی می‌شود و برای اجراهای آینده در بخش مدلسازی آن‌ها را تنظیم می کنیم. مدلسازی را تکرار میکنیم و مدل های تازه را ارزیابی می کنیم تا کاملاً مطمئن شویم مدل حاصل بهترین مدل است.

فعالیت:

پارامترها را تنظیم کنید تا به مدل بهتری دست پیدا کنید.

۵. ارزیابی

در مراحل قبلی ارزیابی ویژگی هایی مانند صحت و جامعیت مدل ها مورد بررسی قرار گرفت. در این مرحله میزان تناسب مدل با اهداف فعالیت تجاری بررسی می‌شود و در صورت وجود نقص در مدل، پاره

۴۳ Lift Tables

۴۴ Gain Tables

ای از دلایل تجاری آن تشخیص داده می‌شوند. در این مرحله نتایج کسب شده با معیارهای ارزیابی مشخص شده در ابتدای پروژه مقایسه می‌شوند.

یک راه مناسب برای تعیین مجموع خروجی های یک پروژه داده کاوی استفاده از تساوی زیر است:

نتایج = مدل ها + یافته ها

این تساوی نشان می‌دهد که مدل ها، همه خروجی های یک پروژه داده کاوی نیستند (اگر چه مدل بخش مهمی از آن است)، بلکه یافته ها نیز بخش مهمی از خروجی اند و نشان دهنده مسائل مهمی در مسیر نیل به اهداف داده کاوی می باشند. یافته ها همچنین به سؤالات تازه و برخی اثرات جانبی (مانند مسئله کیفیت داده‌ها که در داده کاوی تحت پوشش قرار نمی گیرد) اشاره می کنند.

تذکر: اگر چه مدل مستقیماً با مسائل تجاری در رابطه است، اما یافته ها لزوماً مرتبط به مسأله یا هدف خاصی نیستند و اهمیت آن‌ها در مراحل ابتدایی پروژه است.

۵-۱ ارزیابی نتایج

وظیفه: ارزیابی نتایج

این مرحله میزان تطابق مدل با اهداف تجاری را بررسی نموده و در صورت مشاهده نقص در نتایج مدل سعی در یافتن دلائل تجاری برای آن می‌باشد. از دیگر اختیارات فاز ارزیابی، آزمودن مدل در موارد کاربردی و واقعی ست؛ البته اگر محدودیت های زمان و بودجه اجازه این امر را بدهد.

علاوه بر این، فاز ارزیابی سایر نتایج تولید شده به وسیله داده کاوی را نیز مورد سنجش قرار می‌دهد. نتایج داده کاوی شامل مدل های مرتبط با اهداف اصلی کسب و کار و سایر یافته ها می‌باشد. بعضی از آن‌ها می تواند در ارتباط با اهداف تجاری بوده و برخی نیز شامل اطلاعات و راهنمایی هایی باشند که در ادامه سودمند واقع شوند.

خروجی: ارزیابی نتایج داده کاوی با توجه به معیارهای موفقیت تجاری

نتایج ارزیابی را با توجه به معیارهای موفقیت تجاری خلاصه کنید و قضاوت نهایی خود را در مورد تناسب پروژه با اهداف ابتدایی، به آن ضمیمه کنید.

فعالیت ها:

- سعی کنید نتایج بدست آمده را فهمیده و درک نمائید.
- نتایج را بر حسب کاربردشان تفسیر کنید.
- اثر آن‌ها را بر اهداف داده کاوی بررسی نمائید.
- بدیع بودن و سودمندی نتایج داده کاوی را با توجه به پایگاه دانش موجود بررسی کنید.
- نتایج را با توجه به معیارهای موفقیت تجاری ارزیابی کنید. (آیا پروژه به اهداف اصلی کسب و کار

دست یافته است؟)

- نتایج ارزیابی و تفاسیر آن را با یکدیگر مقایسه کنید.
- نتایج را با توجه به معیارهای موفقیت تجاری رتبه بندی کنید.
- تأثیر نتایج را بر هدف کاربردی اولیه، بررسی کنید.
- آیا نشانی از اهداف تجاری جدید در این پروژه یا پروژه‌های بعدی هست؟
- نتایج را برای پروژه‌های بعدی داده کاوی شرح دهید.

خروجی: مدل های تأیید شده

بعد از ارزیابی مدل با توجه به معیارهای موفقیت تجاری، مدل های منطبق با معیارها، در لیست مدل های تأیید شده قرار می گیرند.

۵-۲ مرور فرآیند

وظیفه: مرور فرآیند

در این مرحله از کار به مدل های رضایت بخشی دست یافته ایم که خواسته های تجاری مان را هم برآورده میکند. اکنون زمان مناسبی ست که مروری کلی بر تعهدات داده کاوی داشته باشیم ، تا اگر عامل مهمی از نظر دور مانده است، لحاظ شود. فرآیند بازنگری در این مرحله از داده کاوی شامل بازنگری کیفی است.

خروجی: مرور فرآیند

فرآیند بازنگری را خلاصه کنید و پیشنهادهایی برای فعالیت های از قلم افتاده یا فعالیت هایی که مجدداً باید تکرار شوند، ارائه کنید.

فعالیت ها:

- خلاصه ای از فرآیند داده کاوی صورت گرفته ارائه کنید.
- فرآیند داده کاوی را مورد تجزیه و تحلیل قرار دهید.

برای هر مرحله فرآیند به این سوالات پاسخ دهید:

- آیا لزومی برای انجام آن فعالیت در آن مرحله خاص هست؟
- آیا آن فعالیت به صورت بهینه اجرا شده است؟
- فرآیند از چه طریقی می تواند بهبود پیدا کند؟
- نواقص موجود را شناسایی کنید.
- مراحل گمراه کننده را شناسایی کنید.
- راهکارهای جایگزین و مراحل غیر قابل انتظار را شناسایی کنید.

- با توجه به معیارهای موفقیت تجاری نتایج داده کاوی را بازبینی کنید.

۵-۳ تعیین مراحل بعدی

وظیفه: تعیین مراحل بعدی

بـا توجه به ارزیابی نتایـج و بازبینی فرآیند باید در مورد نحوه ادامه پروژه تصمیم گیری کنیم. این که آیا پروژه را در این مرحله تمام کنیم و گسـترش و اسـتقرار آن را شروع کنیم یا اینکه تکرارهای بعدی را آغاز کنیم و یا اصلاً مقدمات انجام پروژه داده کاوی تازه ای را فراهم کنیم.

خروجی: فهرست راهکار های ممکن

فهرستی از راهکارهای ممکن، به همراه دلایلی در تأیید یا رد آن‌ها تهیه کنید.

فعالیت ها:

- قابلیت های هر نتیجه را برای مورد گسترش قرار گرفتن بررسی کنید.
- قابلیت های فرآیند حاضر، برای بهبود یافتن را برآورد کنید.
- منابع باقیمانده را بررسـی کنید تا امکان تکرارهای بیشـتر فرآیند مشـخص شود. (همچنین می توانید امکان دستیابی به منابع بیشتر را نیز بررسی کنید).
- یک مسیر جایگزین پیشنهاد کنید.
- اصلاحات لازم را در طرح فرآیند ایجاد کنید.

خروجی: تصمیم گیری

تصمیم خود در مورد نحوه پیگیری کار را به همراه دلایل آن شرح دهید.

فعالیت ها:

- راهکارهای ممکن را رتبه بندی کنید.
- یکی از راهکارهای ممکن را انتخاب کنید.
- دلایل انتخاب خود را یادداشت کنید.

۶ . گسترش و استقرار

۶-۱ طراحی برای گسترش و استقرار

وظیفه: طرح گسترش و استقرار

این فاز بر اساس نتایج ارزیابی، یک استراتژی مناسب برای گسترش و استقرار نتایج تعیین می‌کند.

خروجی: طرح گسترش و استقرار

استراتژی گسترش و استقرار را همراه با گام های لازم و چگونگی انجام آن‌ها به طور خلاصه بیان کنید.

فعالیت ها:

- نتایج قابل گسترش و استقرار را خلاصه کنید.
- طرح های دیگری برای گسترش و استقرار ارائه دهید و آن‌ها را ارزیابی کنید.
- برای هر یک از اطلاعات و دانش بدست آمده از نتایج تصمیم گیری کنید.
- چگونگی انتشار اطلاعات و دانش بدست آمده برای کاربران را تعیین نمائید.
- تعیین کنید چگونه می توان بر استفاده از نتایج نظارت کرد و مزایای آن را محاسبه نمود؟ (در چه مکان هایی قابل اجراست؟)
- درباره نتایج نرم افزاری و مدل های قابل گسترش و استقرار تصمیم گیری کنید.
- چگونه یک مدل یا خروجی نرم افزاری از طریق یک نظام سازمانی قابل گسترش و استقرار است؟
- در یک نظام سازمانی چگونه می توان بر استفاده از نتایج نظارت کرد و مزایای آن را محاسبه کرد؟ (در چه مکان هایی قابل اجراست؟)
- مسائل و مشکلاتی که می توانند هنگام گسترش و استقرار نتایج داده کاوی رخ دهند تعیین کنید.

۶-۲ طراحی برای نظارت و نگهداری

وظیفه: طرح نظارت و نگهداری

اگر نتایج داده کاوی از اطلاعات روزانه تجاری و محیط آن حاصل شده باشد، نظارت و نگهداری آن اهمیت دارد. یک آماده سازی با دقت در استراتژی نگهداری منجر به عدم استفاده نادرست در بلند مدت از نتایج داده کاوی خواهد شد. با هدف نظارت بر گسترش و استقرار نتایج داه کاوی، پروژه به یک طرح تفصیلی از فرآیند نظارت نیاز دارد. این طرح به برخی گسترش و استقرار های خاص نیاز دارد.

خروجی: طرح نظارت و نگهداری

استراتژی نظارت و نگهداری را همراه با گام های لازم و چگونگی انجام آن‌ها بطور خلاصه بیان کنید.

فعالیت ها:

- حالات پویا را بررسی کنید. (چه چیزهایی در حیطه کاری تغییر می‌کند؟)
- چگونه می توان بر صحت نتایج نظارت کرد؟
- در چه مواقعی نباید از خروجی های داده کاوی و یا سایر مدل های بدست آمده استفاده کرد؟ معیارهایی برای این موضوع تعریف کنید. (بطور مثال: میزان درستی، مرز صحت قابل قبول، داده‌های جدید، تغییر در دامنه کاربردهای مدل و...) همچنین مشخص نمائید در صورتیکه از یک زمانی قادر به استفاده

مناسب از مدل نباشیم چه اتفاقی خواهد افتاد؟ (بروز رسانی مدل، شروع پروژه داده کاوی جدید و...)

- آیا هدف تجاری مرتبط با مدل، با گذشت زمان تغییر می‌کند؟ مسئله اولیه ای را که مدل به قصد حل آن ساخته شد را به صورت کامل ثبت کنید.
- طرح نظارت و نگهداری را گسترش دهید.

۶-۳ تهیه گزارش نهایی

وظیفه: تهیه گزارش نهایی

در پایان کار، سرپرستان پروژه و تیم آن‌ها یک گزارش نهایی تهیه می کنند. موضوع گزارش بستگی به طرح گسترش و استقرار دارد، این گزارش هم می تواند خلاصه ای از وضعیت پروژه و نتایج تجربی آن باشد هم می تواند گزارشی نهایی برای آشنایی با نتایج داده کاوی باشد.

خروجی: گزارش نهایی

در پایان پروژه حداقل یک گزارش نهایی وجود دارد که شامل همه جزئیات خواهد بود. در این گزارش باید شرحی از نتایج به دست آمده، شرح فرآیند و هزینه های صرف شده، هر گونه انحراف از طرح اصلی، شرح طرح اجرایی و پیشنهاداتی برای پروژه‌های بعدی گنجانده شود.

محتوای جزئیات واقعی پروژه تا حد زیادی بستگی به مخاطبان گزارش ویژه دارد.

فعالیت ها:

- مشخص کنید چه نوع گزارشی لازم است. (معرفی با اسلاید، خلاصه مدیریت، جزئیات یافته ها، توضیح مدل و...)
- بررسی کنید چقدر به اهداف اولیه داده کاوی پرداخته شده است؟
- گروه هدف را برای گزارش معین کنید.
- ساختار و محتویات گزارش را نشان دهید.
- یافته هایی را که گزارش شامل آن‌ها خواهد بود، انتخاب کنید.
- یک گزارش بنویسید.

خروجی: ارائه نهایی

یک گزارش نهایی ممکن است نیاز به ارائه نهایی به مدیران حامی داشته باشد تا در جریان خلاصه پروژه قرار گیرند. یک ارائه معمولی شامل زیر مجموعه ای از اطلاعات موجود در گزارش است، اما به طرق مختلفی ساختاربندی می‌شود.

فعالیت ها:

- برای ارائه نهایی، گروه هدف خود را انتخاب کنید. (آیا گزارش نهایی تاکنون به آن‌ها رسیده است؟)
- بخش هایی از گزارش نهایی را که ارائه نهایی شامل آن‌ها می‌شود انتخاب کنید.

۶-۴ بازبینی پروژه

وظیفه: بازبینی پروژه

این امر شامل ارزیابی صحت امور انجام شده و بهبود مواردی که نیازمند اصلاحند می‌باشد.

خروجی: مستندسازی تجربیات

خلاصه ای از تجربیات مهم انجام شده در طی پروژه تهیه کنید. به عنوان مثال ریسک ها، موارد گمراه کننده یا راهنمایی هایی برای انتخاب مناسب ترین تکنیک های داده کاوی در شرایط مشابه می تواند بخشی از این خلاصه سازی را به خود اختصاص دهد. در یک پروژه ایده آل، پرونده تجارب، شامل گزارش های شخصی همه اعضا از قسمت های مختلف پروژه می‌باشد.

فعالیت ها:

- با همه افرادی که نقشی در پروژه داشته اند مصاحبه کنید و از آن‌ها در مورد تجربیاتشان در طول پروژه بپرسید.
- در صورت مصاحبه با کاربران اصلی که پروژه برای آن‌ها انجام شده است سؤالات زیر را از آن‌ها بپرسید: آیا راضی هستند؟ چه کارهایی می توانست به صورت بهتری انجام شود؟ آیا نیاز به خدمات بیشتر احساس می‌شود؟
- خلاصه ای از بازخوردها تهیه کنید و تجربیات کسب شده را مستند کنید.
- فرآیندهای خاص داده کاوی را یادداشت کنید. (چگونه نتایج و تجربیات حاصل از بکارگیری مدل می توانند در فرآیند بازخورد داشته باشند؟)
- چکیده تجربیات مفید خود را برای پروژه‌های بعدی یادداشت کنید.

فصل چهارم
خروجی‌های CRISP-DM

این فصل شامل توضیح مختصری در مورد اهداف و محتوای گزارش دهی است. گزارش هایی که در این فصل توضیح داده می‌شوند برای انتقال نتایج یک فاز به افرادی که درگیر مراحل آن فاز و یا حتی کل پروژه نبوده اند، استفاده می‌شود. این گزارش ها لزوماً همان خروجی هایی که در مدل مرجع و راهنمای کاربر توضیح داده شده اند نیستند. هدف خروجی ها عموماً مستند سازی نتایج در حین انجام پروژه است.

۱. شناخت و درک کسب و کار

نتایج این فاز را می توان در یک گزارش خلاصه کرد. برای شروع قسمت های زیر پیشنهاد می‌شوند.

پیش زمینه

پیش زمینه یک نمای کلی از پروژه فراهم می‌کند که زمینه های فعالیت پروژه را مشخص می‌کند و نشان می‌دهد چه مشکلاتی برای انجام پروژه شناخته شده، و چرا داده کاوی به عنوان راه حل این مشکلات انتخاب شده است.

اهداف کسب و کار و معیارهای موفقیت

اهداف کسب وکار، اهداف انجام پروژه را به زبان تجاری بیان می‌کند. برای هر هدف باید معیارهای موفقیت کسب و کار تعریف شوند که این معیارها، شاخص های روشنی برای تشخیص موفقیت یا عدم موفقیت یک پروژه اند.

موجودی منابع

هدف این بخش تشخیص پرسنل، منابع داده ای، تسهیلات تکنیکی و سایر منابع مفید برای انجام پروژه است.

نیازها، مفروضات و محدودیت ها

این خروجی، نیازهای کلی پروژه در مورد نحوه اجرای آن، نتایج پروژه، فرضیاتی که در مورد طبیعت مسئله صورت می گیرد، داده‌های مورد استفاده، و محدودیت های اعمال شده بر پروژه را لیست می‌کند.

ریسک ها و احتمالات

این خروجی مشکلاتی را که احتمال دارد در حین پروژه رخ دهند و عملیاتی را که در شرایط مختلف برای به حداقل رسانی آثار مخرب این مشکلات می توان انجام داد، شرح می‌دهد.

لغات تخصصی

این بخش امکان یک آشنایی نسبی با موضوع پروژه را در اختیار افراد ناآشنا با پروژه قرار می‌دهد.

هزینه و فایده

این بخش هزینه های پروژه و سود تجاری پیش بینی شده (در صورت موفقیت پروژه) را شرح

می‌دهد. مزایای دیگری نیز که کمتر محسوس اند (مانند رضایت مشتری) باید در نظر گرفته شوند.

اهداف داده کاوی و معیارهای موفقیت

اهداف داده کاوی بیانگر نتایجی از پروژه است که اهداف کسب و کار را برآورده می کنند. همچنین در این بخش راه حل ها و رهیافت های محتمل و معیارهای موفقیت برای رسیدن به نتایج مذکور، با توجه به ادبیات داده کاوی لیست می‌شوند.

طرح و برنامه پروژه

این قسمت شامل مراحل مختلف اجرای پروژه همراه با زمان و منابع لازم، ورودی ها، خروجی ها و وابستگی هاست. همچنین چگونگی تکرارهای اصلی فرآیند داده کاوی باید به صورت واضح بیان شود؛ مثلاً تکرار فازهای مدلسازی و ارزیابی.

ارزیابی مقدماتی ابزار و تکنیک ها

این قسمت دید اولیه ای از ابزارها و تکنیک هایی که برای استفاده مناسب به نظر می رسند و چگونگی استفاده از آن‌ها را به دست می‌دهد. این خروجی، ابزارها و تکنیک های در دسترس ومورد نیاز را لیست کرده و آن‌ها را با نیازمندی ها هماهنگ می‌کند.

۲. شناسایی و درک داده ها

خروجی های فاز درک داده در گزارش های مختلفی که باید در حین انجام پروژه و وظایف مربوطه تهیه شوند، مستندسازی می‌شود. این گزارش ها داده‌هایی را که در زمان انجام فرآیند داده کاوی کاوش می‌شوند توصیف می‌کند. برای گزارش نهایی خلاصه ای از قسمت های مرتبط و مهم کافی خواهد بود.

گزارش جمع آوری اولیه داده‌ها

این گزارش چگونگی تشخیص و استخراج انواع پایگاه داده را بیان می‌کند.

مباحثی که باید ذکر شوند عبارتند از:

- پس زمینه داده‌ها
- لیست پایگاه داده‌ها و حوزه و دامنه ای که هر یک شامل می‌شوند.
- روش استخراج یا دسترسی به داده‌ها در هر پایگاه داده
- مشکلاتی که در مرحله استخراج یا دسترسی داده‌ها پیش آمده است.

گزارش توصیف داده ها

در این بخش هر یک از مجموعه داده‌های به دست آمده توصیف می‌شوند.

مباحثی که باید مطرح شوند عبارتند از:

- توصیف جزئیات هر یک از منابع داده ای
- لیست جداول و دیگر موضوعات مرتبط با هر پایگاه داده
- توصیف هر فیلد شامل واحدها، سیستم کدگذاری مورد استفاده و...

گزارش کاوش داده ها

این گزارش، جستجوی داده و نتایج آن را توضیح می‌دهد.

مباحثی که باید مطرح شوند عبارتند از:

- پس زمینه، شامل اهداف کلان کاوش داده‌ها.

برای هر نوع از کاوش داده موارد زیر مشخص گردد:

- قواعد و الگوهای قابل انتظار
- روش های تشخیص
- قواعد و الگوهای یافت شده (قابل انتظار و غیر قابل انتظار)
- هر موضوع غیر قابل انتظار دیگر
- نتایج تبدیل داده، پاکسازی و سایر مراحل پیش پردازش
- نتایج مربوط به اهداف داده کاوی و دورنماهای کسب و کار
- خلاصه کلیه نتایج بدست آمده

گزارش کیفیت داده ها

این گزارش کلیت و صحت داده‌ها را توصیف می‌کند.

مباحثی که باید ذکر شوند عبارتند از:

- پس زمینه، شامل کیفیت مورد انتظار داده‌ها

برای هر مجموعه داده:

- راهکار مورد استفاده برای ارزیابی کیفیت داده‌ها
- نتایج ارزیابی کیفیت داده‌ها
- خلاصه ای از نتایج کیفیت داده‌ها

۳. آماده سازی داده ها

گزارش های مرحله آماده سازی داده، بیشتر بر مراحل پیش پردازش داده به منظور آماده ساختن آن‌ها برای ساخت مدل های داده کاوی، تمرکز می کنند.

گزارش توصیف مجموعه داده

این گزارش شامل توصیف یک مجموعه داده (پس از پیش پردازش) و فرآیند تولید آن می‌شود.

مباحثی که باید مطرح شوند عبارتند از:

- پس زمینه، شامل اهداف کلان و طرح کلی پیش پردازش
- ارائه توضیحات لازم در خصوص انتخاب یا حذف مجموعه داده‌ها.

برای هر مجموعه داده:

- توصیف مراحل پیش پردازش، شامل عملیات لازم برای اطمینان از دارا بودن کیفیت مطلوب داده‌ها.
- توصیف مفصل مجموعه داده‌های حاصل، بصورت جدول به جدول و فیلد به فیلد.
- ارائه توضیحات لازم در خصوص انتخاب یا حذف صفات و ویژگی ها.
- ارائه هر یافته ای که در طول پیش پردازش بدست آمده و کلیه نکاتی که برای پیشرفت کار مناسب به نظر می رسد.
- نتیجه گیری و خلاصه سازی.

۴. مدلسازی

خروجی های تولید شده در طول فاز مدلسازی می توانند در قالب یک گزارش جمع آوری شوند. بخش های زیر برای این گزارش پیشنهاد می گردد:

مفروضات مدلسازی

این قسمت تمام مفروضاتی را که درباره داده‌ها در نظر گرفته شده، یا به صورت ضمنی در تکنیک های مدلسازی وارد شده اند، به وضوح تعریف می‌کند.

طرح آزمون

این قسمت نحوه ایجاد مدل و چگونگی مورد ارزیابی قرار گرفتن آن‌ها را توضیح می‌دهد.

مباحثی که باید مطرح شوند عبارتند از:

- پس زمینه، شرح مدل های استفاده شده و ارتباط آن‌ها با اهداف داده کاوی به صورت مختصر.

برای هریک از وظایف مدلسازی:

- شرح مفصل نوع مدل و داده‌های آموزشی استفاده شده
- توضیح چگونگی آزمون و ارزیابی مدل
- توصیف داده‌های مورد نیاز برای آزمون
- برنامه تولید داده‌های آزمون در صورت وجود
- در صورت وجود، ارائه طرح چگونگی تولید مجموعه داده آزمون
- توصیف هرگونه طرح ارزیابی مدل ها توسط متخصصین حوزه کسب و کار
- خلاصه طرح آزمون

توصیف مدل

این گزارش مدل های نهایی و خلاصه ای از فرآیند ایجاد آن‌ها را بیان می‌کند.

مباحثی که باید در این گزارش مطرح شوند عبارتند از:

- خلاصه مدل های تولید شده

برای هر مدل:

- نوع مدل و رابطه آن با اهداف داده کاوی
- تنظیمات بکار رفته برای تولید مدل
- توصیف جزئیات مدل و همه ویژگی های خاص آن

برای مثال:

- برای مدل های قانون مند، همه قواعد شکل گرفته، همچنین ارزیابی های انجام شده برای هر قانون و مدل کلی از لحاظ دقت و پوشش
- برای مدل های بی قاعده، همه اطلاعات تکنیکی در مورد مدل (از قبیل توپولوژی شبکه های عصبی و...) و نیز لیست توصیفات رفتاری.
- توصیف رفتار مدل و تفسیر آن.
- ثبت نتایج مربوط به الگوهای موجود در داده‌ها (در صورت وجود)؛ گاهی اوقات مدل واقعیات مهمی را در مورد داده‌ها آشکار می‌کند، بدون آنکه ارزیابی جداگانه ای بر روی داده‌ها انجام شود. (مثلاً خروجی عیناً با یکی از ورودی ها برابر شده است)
- خلاصه ای از نتایج

ارزیابی مدل ها

این قسمت نتایج آزمایش مدل را بر اساس طرح آزمون شرح می‌دهد.
مباحثی که باید ذکر شوند عبارتند از:

- چکیده ای از مراحل فرآیند و نتایج ارزیابی و نیز هر تخلفی از طرح آزمون

برای هر مدل:

- ارزیابی مفصل مدل شامل اندازه گیری ها (مثل دقت مدل) و تفسیر نتایج و رفتار مدل
- ثبت هر گونه نظری در مورد مدل که متخصصین داده و متخصصین حوزه تجاری مربوط اظهار کرده باشند.
- خلاصه ارزیابی مدل
- ملاحظات و نکاتی در مورد اینکه چرا یک تکنیک مدلسازی خاص به همراه حالت خاصی از تنظیم پارامترها به نتایج خوب یا بدی منجر شده است.
- خلاصه ارزیابی مجموعه کامل مدل ها

۵. ارزیابی

ارزیابی نتایج داده کاوی با توجه به معیارهای موفقیت کسب وکار.
این گزارش اهداف داده کاوی را با اهداف کسب و کار و معیارهای موفقیت آن مقایسه می‌کند.
مباحثی که باید مطرح شوند عبارتند از:

- مروری بر اهداف کسب و کار و معیارهای موفقیت آن (که ممکن است در طول دوره زمانی داده کاوی و یا پس از حصول نتایج تغییر کرده باشند)

برای هر معیار موفقیت کسب و کار

- مقایسه دقیق معیارهای موفقیت و نتایج به دست آمده داده کاوی
- نتیجه گیری در مورد میزان دستیابی به موفقیت و میزان مناسب بودن فرآیند داده کاوی
- مروری بر موفقیت پروژه؛
- آیا پروژه به اهداف اصلی کسب وکار دست یافته است؟
- آیا خواسته ها و اهداف جدیدی در کسب و کار به وجود آمده اند که در ادامه پروژه یا در پروژه‌های بعدی باید مورد توجه قرار گیرند؟
- نتیجه گیری در خصوص پروژه‌های داده کاوی بعدی.

مرور فرآیند

این قسمت کارایی پروژه را ارزیابی می‌کند و هر عاملی را که بایستی در صورت تکرار پروژه در نظر گرفته شود مد نظر قرار می‌دهد.

لیست فعالیت های ممکن

این قسمت پیشنهاداتی در مورد مراحل بعدی پروژه در اختیار می گذارد.

۶ . گسترش و استقرار

طرح گسترش و استقرار

این قسمت نحوه به کارگیری نتایج داده کاوی را مشخص می‌کند.

مباحثی که باید ذکر شوند عبارتند از:

- خلاصه نتایج قابل گسترش و استقرار (این خلاصه از گزارش مراحل بعدی به دست می آید)
- شرح طرح گسترش و استقرار

طرح نظارت و نگهداری

این طرح چگونگی نگهداری نتایج گسترش و استقرار یافته را مشخص می‌کند.

مباحثی که باید مطرح شوند عبارتند از:

- خلاصه نتایج گسترش و استقرار و شناسایی نتایجی که ممکن است به بروز رسانی نیاز داشته باشند. (با ذکر دلیل)

برای هر نتیجه گسترش و استقرار یافته:

- شرح چگونگی آغاز بروز رسانی (بروز رسانی عادی، بروز رسانی پس از یک رخداد یا نظارت اجرایی)
- شرح چگونگی انجام بروز رسانی
- خلاصه نتایج فرآیند بروز رسانی

گزارش نهایی

کزارش نهایی برای خلاصه سازی پروژه و نتایج آن استفاده می‌شود.

مندرجات این گزارش عبارتند از:

- خلاصه شناخت و درک کسب و کار: پس زمینه، اهداف و معیارهای موفقیت
- خلاصه فرآیند داده کاوی
- خلاصه نتایج داده کاوی

- خلاصه نتایج ارزیابی
- خلاصه طرح های گسترش و استقرار و نگهداری
- تجزیه و تحلیل هزینه و فایده
- نتیجه گیری برای کسب و کار
- نتیجه گیری برای داده کاوی های آینده

۷. خلاصه وابستگی ها

جـدول زیـر ورودی های اصلی را برای هر یک از محتواهـای خروجی خلاصه می‌کند. این به این معنا نیسـت که تنها ورودی های لیسـت شـده باید مد نظر قرار گیرند. به عنوان مثال اهداف کسـب و کار باید درمورد همه خروجی ها لحاظ شـوند. ضمن آن که مسـتندات خروجی باید پوشـش دهنده نکات خاصی باشند که از جانب ورودی ها تصریح شده اند.

ارتباط نزدیک با	اشاره به	محتوای خروجی	فازها
		پس زمینه	شـناخت و درک کسب و کار
لغات فنی	پس زمینه	اهداف فعالیت تجاری	
	اهداف فعالیت تجاری	معیار موفقیت فعالیت تجاری	
		فهرست منابع	
	اهداف فعالیت تجاری	ملزومات فرضیات و محدودیت ها	
	اهـداف فعالیـت تجـاری : معیـار موفقیت فعالیت تجاری	پسامدهای احتمالی	
اهداف فعالیت تجاری	پس زمینه	لغات فنی	
طرح پروژه	اهداف فعالیت تجاری	هزینه ها و مزیت ها	
	اهـداف فعالیت تجـاری ملزومات فرضیات و محدودیت ها	اهداف داده کاوی	
هزینه ها و مزیت ها	اهداف ، فعالیت تجاری، فهرست منابـع، ملزومـات ، فرضیـات و محدودیـت هـا ، پیشـامدهای احتمالی	طرح پروژه	

	اهـداف فعالیت تجاری، فهرسـت منابع ، اهداف داده کاوی	گزارش جمع آوری اولیه داده ها	شـنایایی و درک داده ها
گزارش کمیت داده ها	اهـداف فعالیـت تجـاری ، گزارش جمع آوری اولیه داده ها	گزارش توصیف داده ها	
	اهـداف فعالیـت تجـاری، گزارش جمع آوری اولیه داده ها	گزارش کیفیت داده ها	
	اهـداف فعالیـت تجـاری ، گزارش جمع آوری اولیه داده ها	گزارش کاوش داده ها	
	اهـداف فعالیـت تجـاری، اهداف داده کاوی،گـزارش توصیـف داده ها ، گزارش کمیت داده ها ،گزارش کاوش داده ها	مجموعه داده ها و توصیف آن	آماده سازی داده ها
	اهداف داده کاوی ،معیار موفقیت داده کاوی	طرح آزمون	مدلسازی
تنظیم پارامترها	اهداف داده کاوی	مدل ها	ارزیابی
	اهداف داده کاوی	تنظیم پارامترها	
	مـدل ها ، تنظیـم پارامترها ، طرح آزمون	توصیف مدل ها	
	معیار موفقیت ، داده کاوی، طرح آزمون مدل ها	ارزیابی	
	طرح پروژه ارزیابی نتایج بر اساس معیار موفقیت فعالیت تجاری	مراحل بعدی	
طرح نگهداری	اهداف فعالیت ، تجاری، ملزومات ، فرضیات و محدودیت ها	طرح گسترش و استقرار	گسترش و استقرار
طرح گسترش و استمرار	اهـداف فعالیـت تجـاری ، لغـات فنی ، ارزیابی نتایج بر اساس معیار موفقیت فعالیت تجاری	گزارش و ارائه نهایی	
	طرح پروژه بازبینی فرآیند	مستند سازی تجربیات	

فصل پنجم
پیوست

۱. واژه نامه و شرح اصطلاحات تخصصی

فعالیت

بخشی از وظیفه در راهنمای کاربر که اعمال لازم برای انجام وظایف را شرح می‌دهد.

روش شناسی CRISP-DM

عبارتی جامع برای همه مفاهیم شکل گرفته در CRISP-DM.

چهارچوب داده کاوی[۴۵]

مجموعه ای از محدودیت ها و مفروضات. (مانند نوع مسئله، ابزارها و تکنیک ها و دامنه کاربرد)

نوع مسئله داده کاوی[۴۶]

طبقه بندی ویژه مسائل داده کاوی (مانند توصیف و خلاصه سازی داده‌ها، خوشه بندی، توصیف مفاهیم، طبقه بندی، پیش بینی و تحلیل وابستگی)

وظیفه عمومی[۴۷]

وظیفه ای که در همه پروژه‌های داده کاوی وجود دارد.

مدل

مجموعه اعمالی که با اجرای آن‌ها روی مجموعه داده، توانایی پیش بینی متغیر هدف میسر می‌شود.

خروجی

نتایج محسوسی که از انجام یک وظیفه حاصل می‌شوند.

فاز[۴۸]

عبارتی است برای بخش سطح بالایی از مدل فرآیند CRISP-DM که شامل وظایفی خاص می‌باشد.

---

۴۵ Data Mining Context

۴۶ Data Mining Problem Type

۴۷ Generic Task

۴۸ Phase

نمونه فرآیند[۴۹]

پروژه ای خاص که در قالب مدل فرآیند توصیف می‌شود.

مدل فرآیند[۵۰]

ساختار پروژه داده کاوی را تعریف می‌کند و راهنمایی هایی برای اجرای آن فراهم می نماید و شامل مدل مرجع و راهنمای کاربر است.

مدل مرجع

تقسیم بندی پروژه داده کاوی به فازها، وظایف و خروجی ها.

وظیفه خصوصی سازی[۵۱]

وظیفه ای که مفروضات خاصی را در چهارچوب خاصی از داده کاوی ایجاد می‌کند.

وظیفه

قسمتی از فاز که مجموعه ای از فعالیت ها را برای تولید یک یا چند خروجی مشخص می‌کند.

راهنمای کاربر

راهنمای تخصصی در مورد نحوه انجام پروژه داده کاوی.

۲. انواع مسئله داده کاوی

یک پروژه داده کاوی معمولاً با ترکیبی از انواع مسائل داده کاوی سروکار دارد که از حل شدن مجموع آن‌ها، مسئله کسب و کار حل می‌شود.

۲-۱ توصیف و خلاصه سازی داده‌ها

هدف از توصیف و خلاصه سازی داده‌ها، توصیف دقیق خصوصیات داده به صورت ابتدایی و جمع بندی شده است. این توصیف دیدی کلی از ساختار داده در اختیار کاربر قرار می‌دهد. گاهی اوقات توصیف و خلاصه سازی داده می تواند به تنهایی هدف پروژه داده کاوی باشد.

به عنوان مثال ممکن است یک خرده فروش علاقمند به تفکیک حجم معاملات مالی بازار بر اساس اقلام خاصی باشد. در این مورد تغییرات و تفاوت ها نسبت به دوره قبل می تواند خلاصه و مورد توجه قرار

---

۴۹ Process Instance

۵۰ Process Model

۵۱ Specialized Task

گیرد. این نوع مسائل در پایین ترین سطح از مقیاس مسائل داده کاوی قرار می گیرند.

اگـر چه تقریباً در تمام مسـائل داده کاوی، توصیف و خلاصه سـازی داده یـک هدف میانی در فرآیند اسـت، امـا کاربـر اغلب در زمان شـروع داده کاوی نه از هدف دقیق تحلیل مطلع اسـت نـه از مقدار دقیق داده هـا. تحلیـل اولیه داده ها مـی تواند به درک طبیعـت داده ها کمک کند و مفروضـات بالقوه ای برای اطلاعات ناآشـکار پیدا کند. توصیف ساده آماری و تکنیک های مجسم سازی، بینش اولیه ای نسبت به داده ها ایجاد می‌کند. برای مثال توزیع سن مشتریان و محل زندگی آنان نکات مفیدی در مورد مشتریانی که برای استراتژی های آینده بازار باید مورد توجه بیشتری قرار گیرند در اختیار می نهد.

توصیـف و خلاصـه سـازی داده عموماً همراه با سـایر مسـائل داده کاوی انجام می‌شـود. برای مثال توصیـف داده ممکن اسـت به شـکل گیری فرضیاتی در مـورد بخش های جالب توجهـی از داده ها منجر شود. توصیف و خلاصه سازی داده ها زمانی مفید است که بخش های مختلف شناسایی و تعریف شوند. توصیه می‌شود توصیف وخلاصه سازی پیش از هر نوع مسئله داده کاوی دیگر انجام شود. این امر باعث شده است در این کتاب، بخش توصیف و خلاصه سازی داده به عنوان یکی از وظایف فاز درک داده مطرح شود.

خلاصه سـازی از طرفی نقش مهمی در تهیه و نمایش نتایج نهایی پروژه خواهد داشت. خروجی سایر مسـائل داده کاوی (مانند توصیف مفاهیم و مدل های پیش بینی کننده) نیز ممکن اسـت نوعی خلاصه سازی داده تلقی شوند که در سطح مفهومی بالاتری ارائه شده اند.

بسیاری از سیستم های گزارش گیری، بسته های آماری و سیستم های EIS[۵۲] و OLAP نیز می توانند عملیات توصیف و خلاصه سازی داده را انجام دهند. اما هیچ روشی برای انجام مدلسازی های پیشرفته فراهم نمی کنند. بنابراین اگر توصیف و خلاصه سـازی به تنهایی به عنوان نوع مسـئله تلقی شود و هیچ نیازی به مدلسازی نباشد این ابزار برای انجام مسئله داده کاوی مناسب خواهد بود.

۲-۲ خوشه بندی

خوشه بندی نوعی مسئله داده کاوی است که داده ها را به خوشه ها یا زیر گروه های با معنی تقسیم می‌کند. تمام اعضای یک زیر گروه خصوصیات مشـترکی دارند. برای مثال در تحلیل سـبد خرید می توان خوشه های سبد را بر اساس کالاهایی که در آن وجود دارد تعریف کرد. (یعنی خصوصیت مشترک اعضای زیر گروه یا خوشه، نوع کالای خریداری شده باشد)

خوشه بندی می تواند به صورت دستی یا نیمه- خودکار انجام شود. تحلیلگر می تواند زیر گروه های

۵۲ Executive Information System

خاصی را مرتبط با سؤالات کسب و کار فرض کند که در نتیجه دانسته های قبلی او و یا نتایج مرحله توصیف و خلاصه سازی داده است. با این حال تکنیک های خوشه بندی خودکار نیز وجود دارند که می توانند ساختارهایی را در داده‌ها تشخیص دهند که قبلاً نامشخص و پنهان بوده اند و خوشه بندی را به این صورت انجام دهند.

خوشه بندی می تواند به خودی خود یکی از انواع مسائل داده کاوی باشد که در این صورت تشخیص خوشه ها هدف اصلی داده کاوی خواهد بود. برای مثال ممکن است برای ارسال آگهی های بیمه پرستاری، بخواهیم آدرس ها و کدهای پستی تمام افرادی را که از لحاظ سن و درآمد بالاتر از میانگین جامعه هستند به دست آوریم.

با این وجود معمولاً خوشه بندی یک مرحله موقت در جهت حل سایر مسائل داده کاوی ست و در این موارد هدف ممکن است تقسیم مجموعه داده‌ها به قسمت های قابل مدیریت و کنترل پذیر بوده و یا یافتن زیرمجموعه های همگن تری از داده‌ها برای ساده نمودن تحلیل باشد. معمولاً در مجموعه داده‌های بزرگ تأثیر انواع داده‌های مختلف بر روی هم باعث می‌شود تا در نهایت الگوهای موجود و مطلوب در داده‌ها مبهم و غیر قابل تشخیص شوند. بنابراین خوشه بندی مناسب می تواند به آسانتر شدن مراحل دیگر داده کاوی کمک کند. برای مثال تحلیل وابستگی های بین کالاهای موجود درمیلیونها سبد خرید بسیار مشکل است در حالی که تشخیص وابستگی ها در خوشه های مطلوب سبد خرید بسیار آسانتر و منطقی تر خواهد بود. تعیین خوشه های مرتبط با سبدهای گران قیمت، سبدهای لوازم دم دستی و راحتی و یا سبدهای مربوط به روز یا یک زمان خاص.

تذکر: گاهی میان مفهوم واژه خوشه بندی و سایر مفاهیم داده کاوی مانند طبقه بندی، تداخل به وجود می آید. این تداخل به این دلیل است که برخی افراد از این کلمه برای اشاره به ایجاد کلاس ها و زیر گروه ها استفاده می کنند. منظور برخی دیگر ایجاد مدل هایی برای پیش بینی کلاس های از پیش شناخته شده، روی داده‌های جدید است. اما در این کتاب واژه طبقه بندی را به مفهوم دوم اختصاص داده و کلمه خوشه بندی را برای اشاره به مفهوم اول استفاده می کنیم. با این وجود تکنیک های طبقه بندی می توانند برای فهمیدن توصیف خوشه های یافت شده استفاده شوند.

تکنیک های مناسب عبارتند از:

- تکنیک های خوشه بندی
- شبکه های عصبی
- مجسم سازی

مثال:

یک شرکت اتومبیل سازی معمولاً اطلاعات مربوط به مشتریانش را در رابطه با ویژگی های اجتماعی و اقتصادی آن‌ها جمع آوری می‌کند. خصوصیاتی مانند میزان درآمد، جنسیت، سن، شغل و... . با استفاده از تحلیل خوشه ای، شرکت می تواند مشتری های خود را به زیر گروه های قابل فهم تر و معنی دارتر تقسیم کند و ساختار هر گروه را به صورت جداگانه تحلیل کند و استراتژی های بازاریابی خاصی برای هر یک از زیرگروه ها اعمال کند.

۲-۳ توصیف مفاهیم

هدف از توصیف مفاهیم، ارائه توصیفی قابل فهم از مفاهیم یا کلاس هاست. هدف، ساخت مدل های کامل با دقت پیش بینی بالا نیست بلکه هدف کسب بینش و درک بهتر است. برای مثال ممکن است شرکتی بخواهد در مورد مشتری های وفادار و مشتری های گذرای خود بیشتر بداند و بر اساس توصیف مفاهیم «وفادار» و «گذرا» نتیجه بگیرد چه کارهایی می تواند انجام دهد تا مشتریان وفادار، بمانند، یا مشتریان گذرا، وفادار شوند.

توصیف مفاهیم هم با خوشه بندی و هم با طبقه بندی رابطه نزدیکی دارد. خوشه بندی ممکن است بدون این که هیچ توصیف قابل فهمی در اختیار قرار دهد به صورت بندی اشیاء متعلق به یک کلاس منجر شود. اصولاً خوشه بندی قبل از توصیف مفاهیم انجام می‌شود اما برخی تکنیک ها مانند خوشه بندی مفهومی، خوشه بندی و توصیف مفاهیم را به طور همزمان انجام می دهند.

توصیف مفاهیم برای اهداف طبقه بندی نیز می تواند استفاده شود. از سوی دیگر برخی از تکنیک های طبقه بندی مدل های قابل فهمی ایجاد می کنند که می تواند به عنوان توصیف مفاهیم تلقی شود. تفاوت مهم این دو این است که طبقه بندی باید کامل باشد زیرا نتایج آن باید روی تمام اعضای جامعه مورد مطالعه اعمال شود. اما از سوی دیگر، توصیف مفاهیم نیازی به کامل بودن ندارد و فقط کافیست قسمت های مهم مفاهیم و یا کلاس ها را شرح دهد. در مثال بالا توصیف مفهوم مشتریان وفادار، با توجه به ویژگی های مشتریانی که وفاداری آن‌ها واضح است، کافی خواهد بود.

تکنیک های مناسب عبارتند از:

- روش های استنتاج قانونی
- خوشه بندی مفهومی

مثال:

یک شرکت می تواند با استفاده از داده‌های مربوط به خریداران اتومبیل های جدید و تکنیک استنتاج قانون، قوانینی ایجاد کند که مشتریان وفادار و گذرایش را توصیف کند. قوانین زیر نمونه هایی از قوانین تدوین شده شرکت مذکورند:

اگر جنس : مرد ، سن ❯ ۵۱ ← مشتری: وفادار
اگر جنس : زن ، سن ❯ ۲۱ ← مشتری: وفادار
اگر شغل : مدیر ، سن ❮ ۵۱ ← مشتری: گذرا
اگر وضعیت تأهل: مجرد ، سن ❮ ۵۱ ← مشتری: گذرا

۲-٤ طبقه بندی

طبقه بندی بر این فرض استوار است که مجموعه ای از اشیا به واسطه چندین ویژگی از یکدیگر متمایز می‌شوند و متعلق به کلاس های متفاوتی هستند. برچسب کلاس یک مقدار گسسته (اسمی) است و برای هر شیء مقدار مشخصی است. هدف از ساخت مدل های طبقه بندی کننده آن است که مدل بتواند به اشیائی که قبلاً دیده و برچسب گذاری نشده اند، برچسب صحیحی نسبت دهد. مدل های طبقه بندی کننده غالباً برای مدلسازی پیشبینانه استفاده می‌شوند.

برچسب کلاس می تواند به وسیله کاربر یا با استفاده از خوشه بندی تعیین گردد. طبقه بندی یکی از مهم‌ترین انواع مسائل داده کاوی است که گستره کاربرد وسیعی دارد. بسیاری از مسائل داده کاوی قابل تبدیل به مسائل طبقه بندی هستند. به عنوان مثال، تعیین اعتبار سعی می‌کند تا ریسک اعتباردهی به یک مشتری جدید را ارزیابی کند. این موضوع می تواند به یک مسئله طبقه بندی تبدیل شود به این صورت که دو کلاس برای مشتریان در نظر گرفته می‌شود. مشتریان قابل اعتماد و غیر قابل اعتماد. یک مدل طبقه بندی کننده می تواند با توجه به اطلاعات موجود مشتری و رفتار اعتباری او ساخته شود. سپس مدل ساخته شده می تواند برای طبقه بندی مشتریان جدید به یکی از دو کلاس فوق استفاده شود و بر اساس آن، مشتری مورد نظر پذیرفته یا رد گردد.

طبقه بندی تقریباً با همه انواع مسائل داده کاوی در ارتباط است. مسائل پیش بینی با گسسته سازی مقادیر پیوسته فیلد هدف به مسائل طبقه بندی تبدیل می‌شوند. «گسسته سازی» امکان تبدیل طیف های پیوسته به بازه های گسسته را فراهم می‌کند. این مقادیر گسسته سپس به جای مقادیر عددی دقیق، به عنوان برچسب کلاس ها استفاده می‌شود و مسئله را به یک مسئله طبقه بندی تبدیل می‌کند. برخی از تکنیک های طبقه بندی، توصیفی از مفاهیم و یا کلاس های قابل فهم ایجاد می کنند. طبقه بندی با تحلیل وابستگی ها نیز مرتبط است زیرا مدل های طبقه بندی کننده معمولاً وابستگی های بین ویژگی ها را استخراج می کنند.

خوشه بندی می تواند برای به دست آوردن برچسب اولیه کلاس ها مورد استفاده قرار گیرد و یا مجموعه داده‌ها را طوری محدود می‌کند که مدل طبقه بندی کننده بتواند به طور بهتر و کاراتری ساخته شود. تجزیه و تحلیل انحرافات قبل از ساخت مدل طبقه بندی کننده، بسیار مفید است زیرا وجود

انحرافات و مقادیر پرت ممکن است باعث شود که الگوهایی که در مجموعه داده وجود دارد مبهم و غیر قابل تشخیص شوند و در نتیجه به دست آوردن یک مدل طبقه بندی کننده کارا، ممکن نباشد. از سوی دیگر یک مدل طبقه بندی کننده می تواند برای تشـخیص داده‌های پرت، انحرافات و سـایر مشـکلات موجود در داده‌ها نیز استفاده شود.

تکنیک های مناسب عبارتند از:

- تحلیل ممیزی
- روش استنتاج قانون
- یادگیری درخت تصمیم
- شبکه های عصبی
- KNN (k- نزدیک ترین همسایه)
- استدلال مبتنی بر موارد
- الگوریتم ژنتیک

مثال:

بانـک هـا عموماً اطلاعاتی در زمینه رفتار اعتباری متقاضیان وام خـود دارند. اگر این اطلاعات مالی را با سـایر اطلاعات شخصی مانند جنس، سن و درآمد ترکیب کنیم، آنگاه امکان توسعه سیستم برای طبقه بندی مشتریان جدید به کلاس های قابل اعتماد و غیر قابل اعتماد به وجود خواهد آمد. (این طبقه بندی مشخص کننده میزان ریسکِ اعتباری در پذیرفتن یک مشتری خواهد بود)

۲-۵ پیش بینی

مسئله مهم دیگری که کاربرد وسیعی دارد پیش بینی است. پیش بینی شباهت زیادی با طبقه بندی دارد. تنها تفاوت آن در این است که در پیش بینی، متغیر هدف (کلاس) یک ویژگی کیفی گسسته نیست، بلکه مقداری پیوسته است. هدف پیش بینی یافتن مقادیر عددی برای نمونه های مشاهده نشده است. این مسـئله گاهی رگرسـیون نامیده می‌شـود و در صورتی که پیش بینی با داده‌های سـری زمانی سـروکار داشته باشد، پیش گویی[۵۳] نامیده می‌شود.

تکنیک های مناسب عبارتند از:

- تحلیل رگرسیون

۵۳ Forecasting

- درخت های رگرسیون
- شبکه های عصبی
- KNN (k- نزدیک ترین همسایه)
- روش های Box-Jenkins
- الگوریتم ژنتیک

مثال:

سود سالانه یک شرکت بین المللی با ویژگی هایی مانند تبلیغات، نرخ تبدیل، نرخ تورم و... مرتبط است. در این صورت شرکت با در دست داشتن این مقادیر (یا تخمین قابل اعتمادی از آن‌ها) برای سال بعد می تواند سود مورد انتظار خود را در سال آینده پیش بینی کند.

۲-۶ تحلیل وابستگی

تحلیل وابستگی عبارت است از یافتن مدلی که وابستگی بین داده‌ها یا رخداد ها را شرح دهد. یکی از کاربرد های تحلیل وابستگی، پیش بینی مقدار یک عنصر داده ای، با در اختیار داشتن سایر عناصر است. اگر چه وابستگی ها در مدلسازی پیشبینانه هم قابل استفاده اند اما آن‌ها اغلب برای توصیف روابط در بین داده‌ها استفاده می‌شوند. وابستگی ها می توانند اکید یا احتمالی باشند.

«روابط انجمنی[۵۴]» حالت خاصی از وابستگی است که اخیراً بسیار رایج شده است. روابط انجمنی، رابطه های موجود بین داده‌ها را شرح می دهند. (یعنی داده‌ها یا رخدادهایی که مکرراً با هم رخ می دهند را کشف می کنند). یکی از کاربردهای رایج روابط انجمنی، تحلیل سبدهای خرید است. یک مثال رایج برای روابط انجمنی از این قرار است: «در ۳۰ درصد از کل خریدها نوشیدنی و بادام زمینی با هم خریداری می‌شوند».

الگوریتم هایی که روابط انجمنی را تشخیص می دهند بسیار سریع هستند و تعداد زیادی از این نوع رابطه ها ایجاد می کنند. اما نکته مهم، یافتن مطلوب ترین آن‌هاست.

تحلیل وابستگی رابطه نزدیکی با پیش بینی و طبقه بندی دارد. مثلاً برای ساخت مدل های پیشبینانه، تلویحاً از تحلیل وابستگی ها استفاده می‌شود. همچنین این تحلیل وقتی که ارتباطات خاصی را برجسته می سازد با مسئله توصیف مفاهیم نیز در ارتباط است.

در کاربردها تحلیل وابستگی معمولاً همراه با خوشه بندی انجام می‌شود. در مجموعه داده‌های بزرگ به علت تاثیر متقابل داده‌ها بر روی هم وابستگی ها به ندرت مشخص و برجسته می‌شوند. در این مورد پیشنهاد می‌شود که تحلیل وابستگی به جای کل مجموعه داده روی خوشه های همگن انجام شود.

---

۵۴ Association

«تحلیل توالی» حالت خاصی از تحلیل وابستگی می‌باشد که تمرکز آن بر روی ترتیب زمانی رخدادها هست. بطور مثال در تحلیل سبد خرید، روابط انجمنی، وابستگی بین اقلامی که در یک زمان باهم خرید می‌شوند را مشخص می نماید. اما تحلیل توالی، به توصیف الگوی خرید یک مشتری خاص یا گروهی از مشتریان در طول زمان می پردازد.

تکنیک های مناسب عبارتند از:

- تحلیل همبستگی
- تحلیل رگرسیون
- قوانین انجمنی
- شبکه های بیزین
- برنامه نویسی منطقی استنتاجی
- تکنیک های مجسم سازی

مثال ۱:

با استفاده از تحلیل رگرسیونی، یک تحلیل گر کسب و کار در می یابد که وابستگی های مهمی بین فروش کل یک محصول، قیمت و میزان تبلیغات آن وجود دارد. به محض به دست آوردن این اطلاعات او می تواند با تغییر دادن قیمت یا تبلیغات، فروش خود را به میزان مناسبی برساند.

مثال ۲:

یک شرکت سازنده اتومبیل با اعمال الگوریتم های قوانین انجمنی روی داده‌های مربوط به لوازم جانبی اتومبیل، دریافته است که در ۹۵ درصد از موارد، درصورت سفارش سیستم صوتی پیشرفته، جعبه دنده اتوماتیک نیز سفارش داده می‌شود. بر اساس دریافت این وابستگی، شرکت سازنده اتومبیل تصمیم می گیرد که این محصولات را با هم و به صورت یک مجموعه ارائه دهد و در نتیجه هزینه های مربوط را به نحو قابل توجهی کاهش دهد.

CRISP-DM

Step-by-Step Data Mining Guide

"در مدت فعالیت من، بعنوان پژوهشگر ارشد و رییس دپارتمان داده‌کاوی در شرکت بنز آلمان، این دپارتمان در چند پروژه داده‌کاوی نیز شرکت داشت که توسط بازار مشترک اروپا پشتیبانی مالی می‌شد. CRISP-DM بنظر من مهمترین آنها بود. امروزه که بیش از چهارده سال از پایان این پروژه میگذرد، بدون اغراق می توان گفت که CRISP-DM مهمترین استاندارد موجود جهت انجام پروژه های کاربردی داده کاوی در جهان می باشد."

غلامرضا نخعی زاده – استاد دانشگاه کالسروهه آلمان و مدیر سابق دپارتمان داده کاوی شرکت بنز

انتشارات صانعی
www.sanatbook.com