فصل اول

مقدمه ای بر

CRISP-DM

۱- متدلوژی CRISP-DM

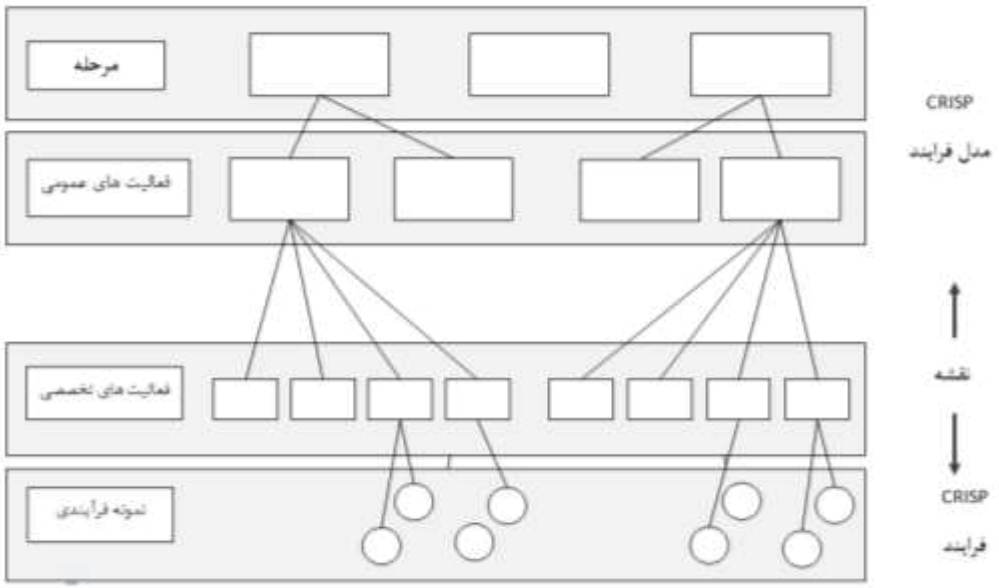
تقسیم بندی سلسله مراتبی[1]

روش داده کاوی CRISP-DM یک مدل فرایندی سلسله مراتبی است که متشکل از مجموعه ای از فعالیت شرح داده شده در چهار سطح از انتزاع[2] می باشد (از عمومی تا تخصصی): مرحله، فعالیت های عمومی، فعالیت های تخصصی و نمونه فرایندی (شکل ۱ را ببینید). در سطح بالا، فرایند داده کاوی به تعدادی از مراحل سازماندهی می شود. هر مرحله شامل چندین فعالیت عمومی سطح دوم است. سطح دوم، عمومی نامیده می شود، چون برای پوشش دادن تمام موقعیت های داده کاوی ممکن، کافی است. فعالیت عمومی تا حد ممکن کامل و پایدار در نظر گرفته می شوند. ابزار های کامل پوشش دهنده کل فرایند داده کاوی و تمام کاربردهای داده کاوی امکان پذیر هستند. پایدار بودن بدان معنی است که مدل باید برای تحولات هنوز پیش بینی نشده مانند تکنیک های مدل سازی جدید نیز معتبر باشد.

سطح سوم، سطح و فعالیت های تخصصی است، و جایی برای توضیح این است که اقدامات در فعالیت عمومی چگونه باید در موقعیت های خاص انجام شوند. به عنوان مثال، در سطح دوم ممکن است یک فعالیت عمومی به نام پاک سازی داده ها[3] وجود داشته باشد. سطح سوم توضیح می دهد چگونه این کار در موقعیت های مختلف متفاوت است، مانند پاک سازی مقادیر عددی در مقابل پاک سازی مقادیر قطعی و یا اینکه آیا نوع مسئله خوشه سازی[4] است و یا مدل سازی پیش گویانه است. توصیف مراحل و فعالیت ها به عنوان مراحل گسسته در یک نظم خاص نشان دهنده توالی ایده آلی از رویدادها هستند. در عمل، بسیاری از فعالیت ها می توانند در یک جهت متفاوت انجام شوند و اغلب لازم است که بارها و بارها به فعالیت های قبلی برگردید و اقدامات خاصی را تکرار کنید. مدل فرایند ما تلاش نمی کند تا تمام این مسیرهای ممکن را از طریق فرایند داده کاوی بدست آورد چون این موضوع نیاز به یک مدل فرآیندی بیش از حد پیچیده دارد. سطح چهارم، نمونه فرایندی، یک رکورد از اقدامات، تصمیمات و نتایج حاصل از تعامل واقعی داده کاوی است. یک نمونه فرآیندی با توجه به فعالیت های تعریف شده در سطوح بالاتر سازمان دهی می شود، اما نشان می دهد آنچه را که در واقع در یک تعامل خاص اتفاق افتاده است، و نه آنچه که به طور کلی اتفاق می افتد.

---

[1] Haeirarchical
[2] Abstraction
[3] clean data
[4] Clustering



شکل ۱ : ۴ سطح تقسیم بندی متدلوژی CRISP-DM

۲-۱ مدل مرجع و راهنمای استفاده

بصورت افقی، روش CRISP-DM بین مدل مرجع و راهنمای استفاده تمایز قائل می شود. مدل مرجع ارائه دهنده یک دید کلی از مراحل، فعالیت و خروجی هاست و توصیف می کند چه کاری در یک پروژه داده کاوی انجام می شود. راهنمای استفاده راهنمایی ها و نکات بیشتری برای هر مرحله و هر فعالیت در یک مرحله ارائه می دهد و به تصویر می کشد که چگونه یک پروژه داده کاوی به انجام می رسد. این سند هر دو مدل مرجع و راهنمای استفاده را در سطح عمومی پوشش می دهد.

۲- نقشه مدل های عمومی به مدل های تخصصی

زمینه داده کاوی،نقشه بین سطح عمومی و سطح تخصصی در CRISP-DM را هدایت می کند. در حال حاضر، ما بین چهار بعد مختلف داده کاوی تمایز قائل می شویم:

- دامنه کاربرد حوزه خاصی است که در آن پروژه داده کاوی اتفاق می افتد.
- نوع مسئله داده کاوی، کلاس خاصی (ES) از اهداف را توصیف می کند که پروژه داده کاوی با آن در ارتباط است (همچنین نگاه کنید به پیوست V.2).

- جنبه های فنی مسائل خاصی در داده کاوی را پوشش می دهد که چالش های مختلفی را توصیف می کند که معمولا در داده کاوی رخ می دهد.
- سپس ابزار و روش مشخص می کند که کدام ابزارها و تکنیک های داده کاوی در طول پروژه داده کاوی استفاده می شوند.

جدول ۱ این ابعادِ در زمینه های داده کاوی را خلاصه می کند و نمونه های خاصی برای هر یک از ابعاد را نشان می دهد.

مفاهیم داده کاوی				
ابعاد	**دامنه کاربرد**	**نوع مسئله داده کاوی**	**جنبه های فنی**	**ابزار و روش**
مثال ها	پاسخ مدل سازی	توصیف و تلخیص[1]	داده های ناقص	کلمنتاین[2]
	پیش بینی	بخش بندی	پرت	مجموعه داده
		توصیف مفاهیم		درخت تصمیم
		دسته بندی		
		پیش بینی		
		تحلیل وابستگی ها		

جدول۱: ابعاد زمینه های داده کاوی و نمونه ها

یک زمینه داده کاوی خاص یک مقدار واقعی برای یک یا چند مورد از این ابعاد است.

به عنوان مثال، یک پروژه داده کاوی در ارتباط با یک مسئله طبقه بندی[3] در پیش بینی به منزله یک زمینه خاص است. هر چه مقادیر بیشتری برای ابعاد زمینه مختلف ثابت باشند، زمینه داده کاوی گسترده تر است.

نقشه با مفاهیم

ما بین دو نوع مختلف از نقشه ها بین سطح عمومی و تخصصی در CRISP-DM تمایز قائل می شویم:

نقشه در حال حاضر:

[1] summarization
[2] Clementine
[3] Classification

اگر ما تنها مدل فرایند عمومی را برای انجام یک پروژه داده کاوی واحد بکار ببریم و تلاش کنیم تا فعالیت عمومی و توصیف های آن ها را به پروژه های خاص مورد نیاز نقشه کنیم، ما در مورد یک نقشه واحد تنها برای (احتمالا) یک راهنما[1] صحبت می کنیم.

نقشه برای آینده:

اگر ما به طور سیستماتیک مدل فرایند کلی را با توجه به یک زمینه از پیش – تعریف شده شناسایی کنیم (و یا به طور مشابه و بطور سیستماتیک تجزیه و تحلیل کنیم و تجارب یک پروژه واحد را به سمت یک مدل فرایند تخصصی برای کاربری های آینده در زمینه های قابل مقایسه جمع آوری کنیم)، ما به صراحت در مورد نوشتن یک مدل فرایند تخصصی از نظر CRISP-DM صحبت می کنیم. این که کدام نوع از نقشه برای اهداف شما مناسب است وابسته به زمینه داده کاوی خاص شما و نیازهای سازمان شما است.

چگونه نقشه را ترسیم کنیم؟

استراتژی اساسی برای نقشه مدل فرایند کلی به سطح تخصصی برای هر دو نوع نگاشت یکسان است:

- زمینه[2] خاص خود را تجزیه و تحلیل کنید.
- هر گونه جزئیات غیر قابل اجرا در به زمینه خود را حذف کنید.
- هر گونه اطلاعات خاص را به زمینه خود اضافه کنید.
- مطالب عمومی را با توجه به ویژگی های درون مایه ی زمینه خود ایجاد کنید.
- محتویات عمومی برای ارائه معانی صریح و روشن تر را در زمینه خود تغییر نام دهید.

۳- توصیف بخش ها

۳-۱ مضامین

مدل فرایند CRISP-DM (این سند) به پنج بخش مختلف سازمان دهی می شود:

- قسمت اول مقدمه روش CRISP-DM است و برخی از دستورالعمل های عمومی برای نقشه از مدل فرایند عمومی به مدل های فرایند تخصصی را فراهم می کند.
- بخش دوم توصیف کننده مدل مرجع CRISP-DM، مراحل آن، وظایف عمومی و خروجی است.
- بخش سوم ارائه دهنده راهنمای کاربر CRISP-DM که فراتر از شرح ناب مراحل، وظایف

[1] Usage

[2] Context

عمومی و خروجی است و شامل مشاوره مفصل تر درباره ی چگونگی انجام پروژه های داده کاوی است از جمله آن چک لیستی می باشد.

- بخش چهارم بر گزارش های تولید شده در طول و بعد از یک پروژه تمرکز می کند و خروجی های این پروژه ها را نشان می دهد. همچنین ارجاعات متقابل میان خروجی و وظایف را نشان می دهد.
- در نهایت، بخش V پیوست است، که واژه نامه اصطلاحات مهم را پوشش می دهد و همچنین یک مشخصه از نوع مسئله داده کاوی است.

۲-۳ اهداف

کاربران و خوانندگان این سند باید از دستورالعمل های زیر آگاه باشند:

- اگر شما در حال خواندن این مدل فرایند CRISP-DM برای اولین بار هستید با بخش ۱ آغاز کنید، مقدمه، به منظور درک روش CRISP-DM و تمام مفاهیم آن و چگونگی ارتباط مفاهیم مختلف با یکدیگراست.
- اگر شما نیاز به دسترسی سریع به یک مرور کلی از مدل فرایند CRISP-DM دارید، به بخش دوم مرجعه کنید، مدل مرجع CRISP-DM ، برای شروع سریع یک پروژه داده کاوی و یا برای دست یابی به یک معرفی از راهنمای استفاده CRISP-DM است.
- اگر شما نیاز به مشاوره مفصل در انجام پروژه داده کاوی خود دارید، قسمت سوم، راهنمای کاربر CRISPDM ، با ارزش ترین بخش این سند است. توجه داشته باشید ، اگر شما مقدمه و یا مدل مرجع را نخوانده باشید، به عقب برگردید و شروع به خواندن این دو قسمت کنید.
- اگر شما در مرحله داده کاوی هستند وقتی که شما گزارشات خود را ارسال می کنید، به بخش IV بروید.
- اگر شما ترجیح می دهید توصیفات قابل ارائه ی خود را در طول پروژه تولید کنید، به بخش های III و IV بروید.
- در نهایت، پیوست سند به عنوان اطلاعات زمینه ی اضافی در مورد داده کاوی و استاندارد کرسیپ بسیار مفید است. اگر شما هنوز یک متخصص در این زمینه نیستید از پیوست برای جستجو در شرایط مختلف استفاده کنید.