

۱۲۵۵۶۱۲

عنوان: فرایند استاندارد فراصنعتی داده‌کاوی
نویسنده: دکتر محمد روزبه
تاریخ: ۱۳۹۵/۰۲/۰۱
موضوع: داده‌کاوی

فرایند استاندارد فراصنعتی داده‌کاوی

راهنمای گام به گام داده‌کاوی
مترجم: محمد روزبه

WWW.KOOLAB-IR

عنوان و نام پدیدآور:	راهنمای گام به گام پیاده سازی داده کاوی / مترجم محمد روزبه.
مشخصات نشر: تهران:	تهران: صانعی شه میرزادی، ۱۳۹۳.
مشخصات ظاهری:	۹۶ ص.: مصور، جدول، نمودار.
شابک:	۴۴-۶-۹۷۸۹۳۳۳-۹۶۴-
وضعیت فهرست نویسی:	فیبا
موضوع:	داده کاوی
موضوع:	مدیریت تولید — داده پردازي
شناسه افزوده:	روزبه، محمد، ۱۳۶۳-
رده بندی کنگره:	۱۳۹۲ ۷۶/۹/د۲، ۱۳۹۲ QA
رده بندی دیویی:	۰۰۶/۳۱۲



راهنمای گام به گام پیاده سازی داده کاوی

مترجم: محمد روزبه

ناشر: صانعی شه میرزادی

ناظر فنی و مدیر اجرایی: مریم صدر

لیتوگرافی صاحب

طراحی جلد و صفحه آرایی: بیتا امین پور، نعیمه وجدانی فخر

چاپ اول فروردین ۹۳

شمارگان: ۱۰۰۰ نسخه

قیمت: ۶۵۰۰ تومان

کلیه حقوق برای ناشر محفوظ است.

آدرس فروشگاه: تهران، خیابان انقلاب، بین ۱۲ فروردین و

اردیبهشت، فروشگاه انتشارات صانعی، پلاک ۱۳۱۲

تلفن: ۶۶۴۰۹۹۲۴ - ۶۶۴۰۵۳۸۵ فکس: ۶۶۴۶۲۱۷۷

آدرس دفتر: تهران، خیابان انقلاب، خیابان ۱۲ فروردین

خیابان وحید نظری، پلاک ۱۰۵، واحد ۱

تلفن: ۶۶۴۱۲۳۹۵ - ۶۶۴۰۸۱۶۵

در مدت فعالیت من، بعنوان پژوهش گرا رشد و رئیس دپارتمان داده کاوی در شرکت بنز آلمان، این دپارتمان در چند پروژه داده کاوی نیز شرکت داشت که توسط بازار مشترک اروپا پشتیبانی مالی می شد. CRISP-DM بنظر من مهم ترین آن ها بود. ایده این پروژه که در انجام آن علاوه بر دپارتمان داده کاوی شرکت بنز سه شرکت دیگر از جمله شرکت SPSS نیز سهیم بودند در سال ۱۹۹۶ میلادی شکل گرفت و انجام آن سه سال بطول انجامید. این شرکت که چند سال پیش توسط IBM خریداری شد هماهنگی پروژه را نیز بعهده داشت.

امروزه که بیش از چهارده سال از پایان این پروژه می گذرد، بدون اغراق می توان گفت که CRISP-DM مهم ترین استاندارد موجود جهت انجام پروژه های کاربردی داده کاوی در جهان می باشد.

بسیاری از پروژه های کاربردی داده کاوی که در مدت حضور من در دپارتمان داده کاوی شرکت بنز انجام گرفت، با استفاده از این استاندارد بود. افزون بر این، تاجایی که من اطلاع دارم، در ایران نیز تاکنون این استاندارد در بسیاری از رساله های کارشناسی ارشد که در زمینه داده کاوی نوشته شده اند، مورد استفاده قرار گرفته است. همچنین مدت زیادی است که این استاندارد در ابزار داده کاوی SPSS Modeler نیز تلفیق گردیده و در دسترس کاربران این ابزار قرار دارد.

برای من بعنوان شخصی که نقش کوچکی در بوجود آمدن این استاندارد داشته، بسیار مسرت بخش است که اکنون ترجمه توصیف این استاندارد به زبان فارسی، به همت گروه داده کاوی دایکه و بویژه با کوشش مستمر آقای محمد روزبه، در اختیار پژوهش گران و کاربران فارسی زبان داده کاوی قرار می گیرد. ضمن ابراز قدردانی از زحمات ایشان برای تمام خوانندگان و بویژه برای دانشجویان عزیز آرزوی بهره گیری فراوان از این اثر را دارم.

غلامرضا نخعی زاده

استاد دانشگاه کارلسروهه آلمان

مدیر اسبق دپارتمان داده کاوی شرکت بنز

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 ترجمه‌ای 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 سال‌های بعد متن اولیه بارها دستخوش تغییر شد و با توجه به تجربیات روزافزون که در قالب آموزش، مشاوره و اجرای پروژه‌های متعدد بدست آمده بود مورد بازنگری قرار می‌گرفت. ولی گذر زمان و مشغله‌های کاری باعث کم‌رنگ شدن تصمیم اولیه برای چاپ شده بود؛ تا اینکه برگزاری مستمر دوره‌های آموزشی «پیاده‌سازی فرآیند داده‌کاوی در ابزارهای تخصصی داده‌کاوی» از جمله (Clementine SPSS Modeler) در دو سال اخیر و مشاوره‌های متعددی که برای پیاده‌سازی و اجرای پروژه‌های داده‌کاوی به مدیران و کارشناسان شرکت‌ها، سازمان‌ها و همچنین دانشجویان مقاطع تکمیلی ارائه می‌شد، ضرورت چاپ این استاندارد را بیش از پیش نمایان ساخت.

با توجه به ماهیت استاندارد و همچنین طیف مخاطبان آن، سعی شده تا در کنار پایداری 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 نموده‌اند که صمیمانه از همه آن‌ها تشکر می‌کنم. جناب پروفسور نخعی زاده که راهنمایی‌ها و همراهی‌های ایشان مسلماً برای من فراموش‌نشدنیست. همکاران خویم در هسته اولیه دپارتمان داده‌کاوی، خانم زهرا زجاجی و آقای عباسعلی ناطقی که در تهیه نسخه اولیه این ترجمه از اعضای اصلی بوده‌اند. همچنین دوستان باارزش و عزیزم خانم زهرا ذوالقدر و آقایان احسان اژدری، خشایار مقدم، صادق طولابی فر، پویان رضائی و سایر دوستانی که در بازبینی متن، ویرایش و همچنین حمایت‌هایشان برای تسریع در اتمام این کار نقش بسزایی داشتند.

با تمام این صحبت‌ها معتقدم متن حاضر خالی از اشکال نیست و بهبود کیفیت آن مستلزم ارائه پیشنهادات و انتقادات خوانندگان عزیز می‌باشد.

محمد روزبه

مرداد ۱۳۹۲

فهرست مطالب

۸	پیشگفتار
۱۱	فصل اول: مقدمه
۱۲	روش شناسی CRISP-DM
	جداسازی سلسله مراتبی
	مدل مرجع و راهنمای کاربر
۱۳	نگاشت مدل های عمومی به مدل های اختصاصی
	چهارچوب داده کاوی
۱۵	نگاشت با چهارچوب ها
	چگونگی نگاشت
۱۵	توضیح فصل ها
	مطالب
	اهداف
۱۷	فصل دوم: مدل مرجع CRISP-DM
۲۲	۱ شناخت و درک کسب و کار
	۱-۱ تعیین اهداف فعالیت تجاری
	۲-۱ ارزیابی وضعیت
	۳-۱ تعیین اهداف داده کاوی
	۴-۱ ارائه طرح پروژه
۲۷	۲ شناسایی و درک داده ها
	۱-۲ جمع آوری اولیه داده ها
	۲-۲ توصیف داده ها
	۳-۲ کاوش داده ها
	۴-۲ بررسی کیفیت داده ها
۳۰	۳ آماده سازی داده ها
	۱-۳ انتخاب داده ها
	۲-۳ پاکسازی داده ها
	۳-۳ ساخت داده ها

۳-۴	یکپارچه سازی داده‌ها	
۳۴	۴ مدل‌سازی	
	۱-۴ انتخاب تکنیک مدل‌سازی	
	۲-۴ ایجاد طرح آزمون	
	۳-۴ ساخت مدل	
	۴-۴ ارزیابی مدل	
۳۷	۵ ارزیابی	
	۱-۵ ارزیابی نتایج	
	۲-۵ مرور فرآیند	
	۳-۵ تعیین مراحل بعدی	
۴۰	۶ گسترش و استقرار	
	۱-۶ طراحی برای گسترش و استقرار	
	۲-۶ طراحی برای نظارت و نگهداری	
	۳-۶ تهیه گزارش نهایی	
	۴-۶ بازبینی پروژه	
۴۳	■ فصل سوم: راهنمای کاربر CRISP-DM	
۴۴	۱ شناخت و درک کسب و کار	
	۱-۱ تعیین اهداف فعالیت تجاری	
	۲-۱ ارزیابی وضعیت	
	۳-۱ تعیین اهداف داده کاوی	
	۴-۱ ارائه طرح پروژه	
۵۳	۲ شناسایی و درک داده‌ها	
	۱-۲ جمع آوری اولیه داده	
	۲-۲ توصیف داده‌ها	
	۳-۲ کاوش داده‌ها	
	۴-۲ بررسی کیفیت داده‌ها	
۵۸	۳ آماده سازی داده‌ها	
	۱-۳ انتخاب داده‌ها	
	۲-۳ پاکسازی داده‌ها	
	۳-۳ ساخت داده‌ها	
	۴-۳ یکپارچه سازی داده‌ها	

۵-۳ فرمت داده‌ها

۶۴ ۴ مدل‌سازی

۱-۴ انتخاب تکنیک مدل‌سازی

۴-۲ ایجاد طرح آزمون

۳-۴ ساخت مدل

۴-۴ ارزیابی مدل

۶۸ ۵ ارزیابی

۱-۵ ارزیابی نتایج

۲-۵ مرور فرآیند

۳-۵ تعیین مراحل بعدی

۷۱ ۶ گسترش و استقرار

۱-۶ طراحی برای گسترش و استقرار

۲-۶ طراحی برای نظارت و نگهداری

۳-۶ تهیه گزارش نهایی

۴-۶ بازبینی پروژه

۷۵ ■ فصل چهارم: خروجی‌های CRISP-DM

۷۶ ۱ شناخت و درک کسب و کار

۷۷ ۲ شناسایی و درک داده‌ها

۷۸ ۳ آماده‌سازی داده‌ها

۷۹ ۴ مدل‌سازی

۸۲ ۶ گسترش و استقرار

۸۳ ۷ خلاصه وابستگی‌ها

۸۵ ■ فصل پنجم: پیوست

۸۶ ۱ واژه‌نامه و شرح اصطلاحات تخصصی

۸۷ ۲ انواع مسئله داده کاوی

۲-۱ توصیف و خلاصه سازی داده‌ها

۲-۲ خوشه بندی

۲-۳ توصیف مفاهیم

۲-۴ طبقه بندی

۲-۵ پیش بینی

۲-۶ تحلیل وابستگی

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 فرآیندهای 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 ترادیتی خود، تیمی را متشکل از مشاوران داده کاوی و متخصصان صنعت، برای رفع نیازمندی‌های آنان، به کار گرفته بود.

گرایش‌های بازار آن زمان حاکی از گسترش روزافزون داده کاوی بود؛ گسترشی که هیجان برانگیز بود و البته نگران‌کننده. 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 می‌رساندیم که داده کاوی می‌تواند 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 می‌تواند این موضوع را به ما و همه کاربران نشان دهد.

یک سال بعد کنسرسیومی تجاری تشکیل دادیم (با عنوان فرآیند استاندارد فراصنعتی 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 می‌بود، دریافتیم که برای ادامه کار می‌بایست از گروهی به اندازه کاربران و سایر سرویس دهندگان (مانند ارائه دهندگان انبار داده‌ها و مدیران مشاوره که علاقمند به داده کاوی بودند) انرژی بگیریم. این کار را با ایجاد گروه علاقمندان ویژه^۲ CRISP-DM (SIG) انجام دادیم. به این صورت که از طریق تبلیغات رادیویی از علاقمندان دعوت کردیم در کارگاه یک روزه‌ای که در آمستردام برگزار شد شرکت کنند. ما می‌خواستیم ایده‌های 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.

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 مشت‌تخیلات 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، علی‌رغم تفاوت‌های ظاهری به ویژه در تأیین حد و مرز فازها و مجموعه لغات مورد استفاده، وجوه قابل توجه مشترک در دیدگاه‌های 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 کارگاه‌هایی 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 ساخت‌امی باشد 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 گروه‌های خدماتی حرفه‌ای مانند SPSS و CRISP-DM، NCR را پذیرفته‌اند، و از آن برای حل مسائل تجاری و صنعتی مشتریان استفاده می‌کنند. در این مدت سرویس‌های خدماتی زیادی که CRISP-DM را پذیرفته‌اند با تحلیل‌گران مختلف مکرراً به آن به عنوان یک استاندارد غیر رسمی صنعت مراجعه کرده‌اند. 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 گسترش‌ها و بهسازی‌هایی که بعدها صورت می‌گیرد مطلوب و البته اجتناب‌ناپذیر است اما نسخه CRISP-DM 1.0 آنقدر ارزشمند

هست که نشر و گسترش داده شود.

CRISP-DM نه از طریق نظریه پردازی و کار آکادمیک بر روی اصول فنی شکل گرفته است و نه به وسیله گروه نخبه ای که پشت درهای بسته در مورد آن تحقیق کرده باشند. این دو مسیر به توسعه روش هایی منجر شده است که پیش از این امتحان شده اند اما به ندرت به یک استاندارد مورد اقبال عموم که عملی و موفق باشد منجر شده اند. دلیل موفقیت CRISP-DM استوار بودن بر مبنای تجربه های عملی و واقعی مرسوم است که پروژه های داده کاوی با استفاده از آن ها انجام شده اند. از این رو ما بی نهایت مرهون کاربرانی هستیم که در طول پروژه ایده ها و تلاش های خود را در اختیار ما قرار داده اند.

کنسرسیوم CRISP-DM

آگوست ۲۰۰۰

www.ketab.ir