روش کریسپ در اجرای پروژه‌های داده‌کاوی

کاری از دپارتمان داده‌کاوی گروه داده کاوی صدرا

www.dmining.ir

ممکن است برخی داده‌کاوی (Data Mining) را مجموعه‌ای از نرم‌افزارهای خودکار یا روش‌های ریاضی و آماری بدانند. در واقع داده‌کاوی یک فرآیند و متدولوژی است که به مدیران کمک می‌کند تا از داده‌های خام به اطلاعات ارزشمندی برسند که به بهبود تصمیم‌گیری‌های آنان منجر شود. یکی از متداول‌ترین فرآیندها برای انجام پروژه‌های داده‌کاوی، CRISP-DM (Cross-Industry Standard Process for Data Mining) نام دارد. در این مقاله به‌طور عمده به توضیح این روش خواهم پرداخت.

CRISP-DM

این استاندارد اولین بار در میانه دهه ۱۹۹۰ میلادی توسط گروهی از شرکت‌های اروپایی به‌عنوان روشی برای انجام پروژه‌های داده‌کاوی ارائه شد. شکل ۱ فرآیند یک پروژه داده‌کاوی را تحت این استاندارد نشان می‌دهد. این فرآیند شش مرحله‌ای از درک نیازهای اصلی کسب وکار شروع می‌شود و به ارائه راهکاری برای آن نیاز ختم می‌شود. اگرچه مراحل این فرآیند به دنبال یکدیگر می‌آیند اما در عمل رفت و برگشت‌های زیادی بین مراحل مختلف این فرآیند وجود دارد. کسانی که درگیر پروژه‌های داده‌کاوی بوده‌اند، به‌خوبی می‌دانند که کار کردن با داده نیازمند سعی و خطا و آزمایش کردن است.

شکل ۱:

گام اول: فهم کسب‌وکار

یکی از مراحل مهم یک پروژه داده‌کاوی فهم نیاز کسب وکار است. این کار با مطالعه و فهم دقیق نیازهای مدیریتی آغاز می‌شود. اهداف کسب‌وکار که انگیزه اصلی اجرای پروژه است باید به‌خوبی مشخص شوند. اهدافی مانند این‌که "ویژگی‌های مشترک مشتریانی که اخیراً از دست دادیم و از خدمات و محصولات شرکت‌های رقیب استفاده می‌کنند، چیست؟" یا "هر یک از مشتریان شرکت دارای چه ارزشی برای ما هستند؟" من همیشه توصیه می‌کنم بهتر است افرادی که دارای فهم خوبی از آن کسب‌وکار هستند در تمام مراحل همراه تیم پروژه داده‌کاوی باشند.

فهم کسب‌وکار و هدف اصلی اجرای پروژه، مشخص می‌کند چه داده‌هایی باید جمع‌آوری شوند، چگونه داده‌ها تحلیل شوند و چطور نتایج ارائه شوند. همچنین کمک می‌کند تا بودجه موردنیاز برای اجرا و زمان‌بندی پروژه تعیین گردد.

گام دوم: درک داده

با توجه به نیاز کسب‌وکار، مجموعه‌ای از داده‌ها که می‌توانیم از آن‌ها استفاده کنیم تا هدف آن پروژه محقق گردد، شناسایی می‌شوند. رعایت چند نکته در این مرحله ضروری است:

۱- تحلیل‌گر در مورد نوع داده‌هایی که نیاز دارد باید بسیار دقیق و شفاف باشد. برای مثال ممکن است که یک خرده‌فروش که به دنبال تحلیل رفتار خریداران زن که پوشاک فصلی می‌خرند است، داده‌هایی در مورد وضعیت جمعیت شناختی آنان، میزان خرید و ویژگی‌های اجتماعی-اقتصادی آنان جمع‌آوری کند.

۲- تحلیل‌گر باید با داده‌ها بخوبی ارتباط برقرار کند. او باید منابع جمع‌آوری داده را بشناسد؛ این‌که داده‌ها چگونه جمع‌آوری شده‌اند، در چه قالبی نگه‌داری می‌شوند، دستی جمع‌آوری می‌شوند یا به شکل خودکار، چه کسانی داده‌ها را جمع‌آوری می‌کنند، هر چند وقت یک‌بار داده‌ها به‌روزرسانی می‌شوند و مانند آن.

او همین‌طور باید تعریف دقیق متغیرهایی را که در داده‌ها وجود دارند، بداند. بر اساس تجربه شخصی می‌دانم که حتی ممکن است در داخل یک شرکت افراد مختلف تعریف واحدی از یک متغیر نداشته باشند. تحلیل‌گر باید بداند به‌طور دقیق هر متغیر چه معنی می‌دهد، آیا هم‌پوشانی بین آنچه اندازه‌گیری می‌شود وجود دارد، متغیرهای وابسته و مستقل را شناسایی کند و مانند آن.

۳- تحلیل‌گر باید تشخیص دهد کدام‌یک از متغیرها، کمّی (Quantitative) و کدامیک کیفی (Qualitative) است. متغیرهای کمّی به‌طور مستقیم با اعداد سنجیده می‌شوند. سطح درآمد ماهیانه هر فرد و یا میزان فروش هفتگی شرکت برحسب دلار مثال‌هایی از متغیرهای کمّی هستند. متغیرهای کیفی که متغیرهای رسته‌ای (Categorical) نیز نامیده می‌شوند، مستقیماً با مقادیر عددی سنجیده نمی‌شوند. این متغیرها به دو دسته اسمی (Nominal) و ترتیبی (Ordinal) تقسیم می‌شوند.
متغیرهای اسمی دارای مقادیر محدود و بدون ترتیب هستند. برای مثال جنسیت (زن و مرد بودن)، هوادار یک باشگاه ورزشی بودن با نبودن، استان محل زندگی و رشته تحصیلی نمونه‌هایی از متغیرهای اسمی هستند. متغیرهای ترتیبی دارای مقادیر محدود و بر اساس یک ترتیب هستند. سطح رضایت مشتری (که معمولاً بر اساس طیف لیکرت سنجیده می‌شود: از خیلی راضی تا خیلی ناراضی)، سطح تحصیلات (کاردانی، کارشناسی تا دکترا) نمونه‌هایی از متغیرهای ترتیبی هستند.
اهمیت فهم انواع متغیرها این است که روش‌های تحلیل این متغیرها از نظر آماری متفاوت است. همچنین متغیرهای کمّی را به‌طور مستقیم می‌توان تحلیل کرد ولی متغیرهای کیفی ابتدا باید به شکل عددی کدگذاری شوند تا بتوان آن‌ها را تحلیل کرد.

۴- تحلیل‌گر معمولاً در این گام شروع به بررسی اولیه داده‌ها می‌کند. در این مرحله معمولاً متغیرهای عددی بر اساس خلاصه‌های آماری مانند میانگین، کمینه/بیشینه، انحراف معیار، میانه و یا سایر کمیت‌های آماری موردعلاقه بررسی می‌شوند. در مورد متغیرهای رسته‌ای فرکانس و مد داده‌ها تحلیل می‌شوند. تحلیل‌های همبستگی، رسم نمودارهای پراکندگی، هیستوگرام و سایر روش‌ها برای نمایش گرافیکی داده‌ها در این مرحله بکار می‌روند تا تحلیل‌گر بتواند فهم بهتری نسبت به داده‌ها پیدا کند.

گام سوم: آماده‌سازی داده

هدف از این گام، آماده کردن داده‌ها برای فاز تحلیل با روش‌های داده‌کاوی است. بر اساس تجربه شخصی می‌دانم این فاز معمولاً بیشترین زمان را به خود صرف می‌کند. در برخی از پروژه‌ها ممکن است تا ۸۰ درصد زمان پروژه به مرحله آماده‌سازی داده اختصاص داده شود. علت این مسئله این است که در دنیای واقعی داده‌ها معمولاً آن‌طور که می‌خواهیم نیستند.

وجود المان‌های نامربوط، عدم وجود المان‌های موردعلاقه، خطا و داده‌های پرت(Outliers) ، ناسازگاری و مانند آن نیازمند این است که تحلیل‌گر زمان زیادی را برای آماده کردن داده‌ها بگذارد. در بسیاری از موارد پیش می‌آید که داده‌ها به شکل الکترونیکی ذخیره نشده‌اند و یا اگر شده‌اند نمی‌توان آن‌ها را مستقیم استفاده کرد. در یکی از پروژه‌ها بسیاری از داده‌ها در فایل‌های PDF توسط کارفرما ارائه شده بود. آماده کردن اعداد موجود در این فایل‌ها برای تحلیل کاری طاقت‌فرسا و زمان‌بر بود.

شکل ۲ نشان می‌دهد که در یک پروژه داده‌کاوی چه مراحلی باید طی گردد تا داده‌های دنیای کسب‌وکار برای تحلیل نهایی آماده شوند.

شکل ۲

در فاز درآمیختن داده (Data Consolidation) باید داده‌های مرتبط شناسایی و جمع‌آوری شوند، رکوردها و متغیرهای موردنیاز انتخاب و منابع داده با یکدیگر یکپارچه شوند. در بسیاری از موارد داده‌های کسب‌وکار از منابع مختلف به دست می‌آیند، برخی ممکن است از سیستم ثبت فروش به دست آیند، برخی دیگر از سیستم مدیریت انبار، برخی از طریق نظرسنجی و مانند آن. منظور از یکپارچه‌سازی داده این است که این داده‌ها بتوانند به شکلی کنار هم قرار گیرند که ارتباط آن‌ها مشخص‌شده و قابل‌تحلیل شوند.

در فاز پاک‌سازی داده (Data Cleaning)، داده‌های گم‌شده (Missing Values) که مقادیر آنان نامعلوم است شناسایی می‌گردند. روش‌های مختلفی برای برخورد با داده‌های گم‌شده وجود دارد. در برخی موارد ممکن است مقادیر بسیار محتمل برای آنان پیدا کنیم. در برخی موارد هم آنان را نادیده بگیریم و رکورد مربوط به آن را حذف کنیم. در این فاز داده‌های پرت باید شناسایی شوند. برخی موارد داده‌های پرت حذف می‌شوند چراکه ممکن است در اثر خطا در ورود داده به وجود آمده باشند. با داده‌های پرت باید با احتیاط رفتار کرد. در برخی حالات داده‌های پرت نشان‌دهنده رخدادهای منحصربه‌فرد هستند و بسیار می‌توانند جالب‌توجه باشند. همچنین ناسازگاری‌ها باید شناسایی شوند. برای مثال ممکن است مقادیر متفاوتی برای یک مورد، از دو منبع داده متفاوت به دست آید. در همه این موارد حضور خبرگان و کسانی که با کسب‌وکار آشنا هستند کمک می‌کند تا علت وجود این موارد شناسایی و در مورد نحوه برخورد با آن تصمیم‌گیری شود.

در فاز تبدیل داده (Data Transformation) ممکن است بخواهیم داده‌ها را نرمال کنیم. متغیرهای مختلف در مسئله ممکن است بازه متفاوتی از مقادیر به خود بگیرند. سطح درآمد سالیانه مقدار عددی بسیار بزرگ‌تری از میزان تجربه برحسب سال را به خود می‌گیرد. این مسئله ممکن است در مدل‌های ریاضی سوگیری ایجاد کند. به همین دلیل معمولاً مقادیر متغیرها را به‌گونه‌ای تغییر می‌دهند که نرمال شوند؛

برای مثال همه آن‌ها بین ۱- و ۱+ شوند. روش دیگر برای تبدیل داده، گسسته کردن داده‌های کمّی است. برای نمونه سطح درآمد که یک متغیر کمّی است به سه سطح بالا، متوسط و پایین تقسیم شود. اگرچه میزان دقت اندازه‌گیری افت پیدا می‌کند، ممکن است برای مسئله موردنظر همین سطح دقت کفایت کند. به‌این‌ترتیب از پیچیدگی محاسبات و یا دشواری ارائه نتایج برای مخاطب کاسته می‌شود. از سمت دیگر ممکن است بخواهیم داده‌های رسته‌ای را تجمیع کنیم. برای مثال در داده‌ها، محل زندگی مشتریان ۵۰ دسته مختلف را شامل می‌شود. ممکن است چنین حدی از دقت برای تحلیل لازم نباشد و اگر این نواحی به پنج منطقه کلی تقسیم شوند کفایت کند. در این فاز همچنین ممکن است بر اساس متغیرهای موجود، متغیر جدیدی تعریف شود تا فرآیند تحلیل را ساده‌سازی کند. برای مثال در مورد داده‌های اهدای عضو، در پایگاه داده اصلی گروه خونی گیرنده عضو و گروه خونی دهنده عضو ذکر شده است. تحلیل‌گر می‌تواند متغیر دو ارزشی (Binary) جدیدی تعریف کند که نشان دهد آیا گروه خونی گیرنده و دهنده عضو، باهم هماهنگ است یا خیر.

فاز نهایی، کاهش داده (Data Reduction) است. در داده‌کاوی تمایل داریم با داده‌های بزرگ کار کنیم اما خود این مسئله می‌تواند دشواری‌هایی ایجاد کند. لزوماً ممکن است همه داده‌ها موردنیاز نباشد. در یک پایگاه داده که داده‌ها دارای دو بعد هستند ستون‌ها)متغیرها) و سطرها (رکوردها)، تحلیل‌گر ممکن است ابعاد داده را کاهش دهد .یک روش، کاهش تعداد متغیرهاست. تکنیک‌های آماری مانند تحلیل مؤلفه‌های اصلی (Principal Component Analysis)، تحلیل همبستگی، آزمون کای دو (Chi-Square Test) و یا درخت تصمیم‌گیری (Decision Tree Induction) برای این منظور بکار می‌روند. در مورد تعداد رکوردها، برخی از منابع داده ممکن است شامل میلیون‌ها یا میلیاردها رکورد باشند. این مسئله می‌تواند توان محاسباتی را به شکل نمایی کاهش دهد. در این حالت به‌جای تحلیل همه داده‌ها می‌توان زیرمجموعه‌ای از آن را انتخاب کرد و تحلیل را روی آن انجام داد. تحلیل‌گر باید بسیار دقت کند که در این حالت نمونه به‌گونه‌ای انتخاب شود که منعکس‌کننده الگوها و روابط موجود در داده‌های اصلی باشد. در مورد داده‌هایی که چولگی (Skewness) دارند (به این معنی که یک زیرمجموعه از داده بخش زیادی از آن را تشکیل می‌دهد؛ مثلاً داده‌های فروشی که افراد زیر ۳۰ سال، ۹۰ درصد مشتریان را شامل می‌شوند) ممکن است نیاز باشد تا متعادل‌سازی صورت گیرد. مطالعات نشان داده مدل‌هایی که بر اساس داده‌های متعادل ساخته می‌شوند قدرت پیش‌بینی کنندگی بهتری دارند. یک روش افزایش نمونه‌گیری (Oversampling) از بخش‌هایی است که کمتر در داده‌ها حضور دارند.

گام چهارم: مدل‌سازی

در این گام، تحلیل‌گر ممکن است روش‌های مختلف داده‌کاوی را بر روی داده‌های آماده‌شده امتحان کند تا بتواند به هدف اصلی پروژه برسد. ساخت مدل یک فرآیند خطی نیست و رفت‌وبرگشت‌های زیادی وجود دارد. یک مدل بهینه در داده‌کاوی وجود ندارد و بسته به مسئله‌ای که تحلیل‌گر با آن مواجه است، روش‌های مختلف باید آزمایش شوند و خروجی آن‌ها باهم مقایسه گردند. در این مرحله احتمالاً لازم است به گام قبلی بازگشت و برای برخی از الگوریتم‌ها داده‌ها را به شکل دیگری آماده کرد.

بسته به نیاز کسب‌وکار، داده‌کاوی ممکن است باهدف پیش‌بینی(Prediction) ، پیدا کردن روابط (Association) و یا برای خوشه‌بندی (Clustering) استفاده گردد. در هر یک از این دسته‌ها الگوریتم‌های متفاوتی وجود دارند که بسته به شرایط یکی از آن‌ها یا ترکیبی از آنان استفاده می‌شوند.

گام پنجم: ارزیابی

در گام پنجم مدلی که توسعه یافته است بر اساس دقت و قابلیت عمومی‌سازی آن آزمایش می‌شود. در این مرحله باید ارزیابی شود که مدل تا چه حد می‌تواند به اهداف کسب‌وکار کمک کند. اگر زمان‌بندی و بودجه پروژه اجازه دهد بهتر است مدل در دنیای واقعی آزمایش شود. نتایج آزمایش کمک می‌کند تا مدل ارزیابی شود و شاید اطلاعات جدیدتری به دست آید که به کامل‌تر شدن مدل کمک کند.

این مرحله بسیار مهم و چالش‌برانگیز است .در این مرحله تیم پروژه باید نشان دهد که دانش به‌دست‌آمده از مدل می‌تواند الگوها و روابط جدیدی را به تصمیم‌گیر نشان دهد که با استفاده از آن ارزش جدیدی برای کسب‌وکار خلق می‌شود. این مانند حل کردن یک معما است. آنچه

از فرآیند داده‌کاوی به دست می‌آید تنها بخشی از یک کل است. مدیران و تحلیل‌گران باید نتایج را در فضای کلی آن کسب‌وکار مورد ارزیابی قرار دهند. در اینجا دانش کسب‌وکار کمک بسیاری به بررسی خروجی‌های مدل می‌کند.

مدیران کسب‌وکار معمولاً علاقه و دانش کافی برای آنکه درگیر تحلیل‌های پیچیده ریاضی شوند، ندارند. وظیفه تحلیل‌گر و تیم پروژه داده‌کاوی است تا با ابزارهای گرافیکی و استفاده از جداول ساده به بهترین شکل ممکن نتایج و الگوهای کشف‌شده در داده‌ها را به تصمیم گیران عرضه کنند.

گام ششم: استقرار

بسته به نوع پروژه، فاز استقرار می‌تواند متفاوت باشد. در برخی موارد ارائه گزارش از روند کار و خروجی تحلیل، پایان یک پروژه داده‌کاوی است. در سمت دیگر استقرار یک سیستم قابل تکرار که سازمان از آن بتواند برای مدت‌ها استفاده کند قرار دارد. در استقرار چنین سیستمی باید تحلیل‌گر نیز مشارکت داده شود تا فهم خود را به اجراکننده سیستم انتقال دهد.

مرحله استقرار می‌تواند شامل فعالیت‌های نگه‌داری نیز شود. در طول زمان محیط کسب‌وکار و نیازهای آن تغییر می‌کند و ممکن است مدل به‌دست‌آمده کارایی خود را از دست بدهد. طراحی یک استراتژی نگه‌داری مناسب می‌تواند کمک کند تا کسب‌وکار برای مدت طولانی به‌اشتباه از مدل داده‌کاوی استفاده نکند.

سخن پایانی

در پایان می‌خواهم تأکید کنم که مدیران نباید پروژه‌های داده‌کاوی را یک جعبه سیاه ببینند که از خروجی آن می‌توانند استفاده کنند. چنین رویکردی عموماً به شکست می‌خورد. مدیران باید از فرآیند داده‌کاوی آگاهی داشته باشند، در توسعه آن مشارکت فعال کنند و فهم خود را از کسب‌وکار به شکل سازنده‌ای به تیم پروژه منتقل کنند. این تعامل هم کمک می‌کند تا مدل بهتری ساخته شود و هم به مدیران کمک می‌کند تا به نتایج اطمینان بیشتری داشته باشند و در تصمیم‌گیری‌های خود از آن استفاده کنند.