بسمه تعالی

نکات مهم و کاربردی در داده کاوی

داده کاوی به بررسی و تجزیه و تحلیل مقادیر عظیمی از داده ها به منظور کشف الگوها و قوانین پنهان و معنی دار درون داده ها اطلاق می شود. هدف داده کاوی، قادر ساختن یک سازمان به بهبود تصمیم گیری ها وعملکردهای درونی و بیرونی از طریق درک بهتر محیط می باشد.

الف – مروری بر مفاهیم پایه داده کاوی

داده کاوی شامل شش عمل و وظیفه مهم است که می توان بسیاری از مسائل محیط اطراف خود را در قالب یکی از این شش عمل و وظیفه زیر گنجاند:

■ دسته بندی Classification

ما برای شناخت و برقراری رابطه دربارهٔ دنیا، بطور مداوم دسته بندی، طبقه بندی و درجه‌بندی می کنیم.(Categorization وRanking) بعبارت ساده دسته بندی شامل بررسی ویژگی های یک شی جدید و تخصیص آن به یکی از مجموعه های از قبل تعیین، شده می باشد. (مثالها : دسته بندی متقاضیان وام و اعتبار به عنوان کم خطر، متوسط و پرخطر- انتخاب محتویات یک صفحهٔ وب برای قرار دادن در شبکهٔ اینترنت - تشخیص مدعیان غیر واقعی دریافت خسارت از بیمه و ...)

■ تخمین Estimation

در تخمین، داده های ورودی در قالب متغیرهای ورودی مختلف به سیستم داده می شود و متغیرهای خروجی آن رقمی چون درآمد یا تراز کارت اعتباری می باشد. مدل های رگرسیون و شبکه های عصبی از جمله تکنیک های مناسب داده کاوی برای تخمین می باشند. (سایر مثالها : تخمین تعداد فرزندان در یک خانواده - تخمین درآمد کل یک خانواده - تخمین دوره عمر یک مشتری - تخمین احتمال پاسخ فردی خاص به یک پیشنهاد بیمهٔ عمر و ...)

■ پیش بینی Prediction

پیش بینی مانند دسته بندی یا تخمین است با این تفاوت که اطلاعات، مطابق برخی از رفتارهای پیش بینی شدهٔ آینده یا ارقام تخمین زدهٔ آینده دسته بندی می‌شوند. در عمل پیش بینی، تنها روش برای بررسی صحت دسته بندی، انتظار دیدن آینده است. (مثالها : پیش بینی مشتریانی که در ۶ ماه آینده، بازار محصول/ خدمات ما را ترک خواهند کرد - پیش بینی مشترکین تلفنی که متقاضی خدمات ویژه مانند مکالمهٔ سه جانبه یا پیغام گیر خواهند شد و ...)

نکات مهم و کاربردی
در داده کاوی

گروه بندی شباهت Affinity Grouping

عمل گروه بندی شباهت، برای تعیین ویژگی های همزمانی هستند که در وقوع یک پدیده رخ می دهند. بعبارت دیگرعمل گروه بندی شباهت احتمال وقوع و یا عدم وقوع همزمان ویژگی ها را تعیین می نماید. مثال معمول این موضوع تعیین کالاهایی است که با هم در یک چرخ دستی خرید در سوپر مارکت قرار می گیرند. چیزی که آن را تحلیل سبد بازار می نامیم Market Basket Analysis (MBA)

خوشه بندی Clustering

خوشه بندی به عمل تقسیم جمعیت ناهمگن به تعدادی از زیر مجموعه ها یا خوشه‌های همگن گفته می شود. وجه تمایز خوشه بندی از دسته بندی این است که خوشه بندی به دسته های از پیش تعیین شده تکیه ندارد. دسته بندی بر اساس یک مدل هر کدام از داده ها به دسته ای از پیش تعیین شده اختصاص می یابد.(مثل جنسیت، رنگ پوست و مثال هایی از این قبیل)، اما در خوشه بندی هیچ دستهٔ از پیش تعیین شده ای وجود ندارد و داده ها صرفاً براساس تشابه گروه بندی می شوند.

توصیف و نمایه سازی Profiling

گاهی اوقات هدف داده کاوی، تنها توصیف آن چیزی است که در یک پایگاه داده ای پیچیده درجریان است. نتایج نمایه سازی درک ما را از مردم، محصولات یا فرآیندهایی که داده ها را در مرحلهٔ اول تولید کرده اند افزایش می دهد. شکاف جنسیتی مشهور در سیاست آمریکا، مثالی از این دست است که چگونه این توصیف ساده که " تعداد زنان حامی حزب دموکرات بیش از مردان است"

ب- نکات مهم و کاربردی در داده کاوی

- داده‌کاوی روش یادگیری از اطلاعات گذشته برای اتخاذ تصمیمات بهتری در آینده است. این یادگیری نباید منجر به نتایج زیر شوند:
"یادگیری چیزهایی که درست نیستند" و "یادگیری چیزهایی که درست هستند اما مفید نیستند "

- مطمئن شوید که مجموعهٔ مدل، نشانگر جامعهٔ مرتبط باشد.(مجموعهٔ مدل، منتخبی از داده‌های تاریخی است که برای ایجاد مدل‌های داده‌کاوی استفاده می‌شود. برای اینکه استنتاج‌های ما از مجموعهٔ مدل معتبر باشند، باید مجموعه مدل نمایندهٔ جامعه‌ای باشد که مدل می‌خواهد آن را توصیف، دسته بندی یا امتیاز دهی کند. نمونه‌ای که به درستی منعکس‌کنندهٔ جامعهٔ اصلی نباشد، اریب است؛ استفاده از یک نمونهٔ اریب به عنوان مجموعهٔ مدل، موجب یادگیری چیزهایی می‌شود که درست نیستند.)

- ساده‌ترین رویکرد برای سهیم کردن داده‌ها در فرآیندهای تصمیم‌گیری یک سازمان، آزمون فرضیه است؛ هدف از آزمون فرضیه، اثبات یا رد نظرات پیش‌داوری شده است و ایـن مـورد تقریبـاً جـزئی از همـهٔ فعالیت هـای داده‌کاوی است. داده‌کـاوان معمولاً بین رویکردها رفت‌وبرگشت می‌کنند؛ ابتدا به کمک متخصصان کسب‌وکار و تجارت به ارائهٔ توضیحات احتمالی برای رفتار مشاهده شده می پردازند و بر اساس این فرضیه‌ها مشخص می‌کنند که چه داده‌هایی باید تحلیل شوند، سپس اجازه می دهند داده‌ها فرضیات جدیدی برای آزمودن مطرح کنند. آزمون فرضیه کاری است که دانشمندان و آمارشناسان، عمدهٔ زمان خود را صرف آن می‌کنند. فرضیه، یک توضیح پیشنهادی است که اعتبارش را می توان با تحلیل داده‌ها آزمود. این داده‌ها ممکن است به سادگی با مشاهده جمع‌آوری شوند یا مثل ارسال نامه‌های آزمایشی توسط آزمایش تولید شوند. بالاترین ارزش آزمون فرضیه، زمانی‌است که نشان دهد فرض‌های هدایت کنندهٔ استراتژی های سازمان در بازار، نادرست بوده‌اند.

- مهمترین گام تعریف مسئله و شناسایی Business Problem است. برای تبدیل یک مسئله کسب‌وکار به یک مسئله داده‌کاوی(Data Mining Problem)، باید آن را به یکی از شش فعالیت داده‌کاوی زیر تبدیل نمود:
  - دسته بندی
  - تخمین
  - پیش‌بینی
  - گروه بندی شباهت

- خوشه بندی
- توصیف و نمایه‌سازی

- سه فعالیت اول یعنی دسته بندی، تخمین و پیش‌بینی نمونه‌هایی از داده‌کاوی هدایت‌شده‌اند. گروه بندی شباهت و خوشه بندی، نمونه‌های داده‌کاوی غیرهدایت‌شده‌اند. نمایه‌سازی می‌تواند هم هدایت‌شده و هم غیرهدایت‌شده باشد. در داده‌کاوی هدایت‌شده همیشه یک متغیر هدف وجود دارد، موضوعی که باید دسته‌بندی، تخمین یا پیش‌بینی شود.

- کیفیت و صحت داده ها در داده کاوی و مدل سازی داده ها، بسیار حائز اهمیت هستند. بنابراین به فرآیندهای Data Cleansing بسیار توجه فرمایید. همچنین مطمئن شوید که داده ها دارای سطح جزئیات نادرستی نباشند.

- ابزارهای تصویرسازی داده ها، می توانند در طول بررسی های اولیهٔ مجموعهٔ داده ها، بسیار مفید باشند. بررسی هیستوگرام هر کدام از متغیرها در مجموعهٔ داده ها و پی بردن به آنچه هست و بیان می کند، برداشتن یک گام خوب است. دربارهٔ هر چیز جالب توجه، یادداشت برداری کنید و نحوه توزیع های صفات را بررسی کنید.

- به دست آوردن شمارگان و خلاصهٔ آماری هر موضوع و پدیده، تعداد مجزای ارقام به دست آمده توسط متغیرهای دسته ای، جدول بندی چند بعدی داده ها و اطلاعات، مقایسه ارقام با توصیفات، نرخ تغییرات و حتی میانگین ها در شناخت داده ها بسیار اهمیت دارد.

- با هدف اعتبار بخشی فرضیات، با استفاده از جدول بندی چند بعدی و ابزارهای تصویرسازی مانند نمودارهای پراکندگی، نمودارهای میله ای و نقشه ها، فرضیات موجود دربارهٔ داده ها را می توان معتبر ساخت.

- تخصصی ترین مرحله داده کاوی ایجاد مدل داده ای است. مجموعهٔ مدل، حاوی همهٔ داده هایی است که در فرآیند مدل سازی مورد استفاده قرار گرفته است. از برخی از داده ها در مجموعهٔ مدل برای یافتن الگوها استفاده می شود، از برخی دیگر از داده ها در مجموعهٔ مدل برای اثبات تداوم مدل استفاده می شود و برخی برای ارزیابی عملکرد مدل به کار گرفته می شود.

- توجه شود که هدف اولیه در روش شناسی داده کاوی، تهیهٔ مدل های پایدار است؛ مهم این است که مدل ها در آینده و در هر زمانی از سال به خوبی کار کنند. اگر داده های موجود در مجموعهٔ مدل همگی از یک زمان در سال گرفته شده باشند، احتمال وقوع بی ثباتی در مدل بیشتر است. حتی اگر

قرار است مدلی بر اساس داده های تنها سه ماه ساخته شود، ردیف های مختلف مجموعهٔ مدل باید از دوره های سه ماههٔ متفاوت استفاده کنند.

- از تکنیک های تحلیلی همچون تحلیل روندها(Trend)، تبدیل اعداد به نسبت ها، تبدیل اعداد به مقادیر بازه های گسسته و .... استفاده کنید.

- به ارزیابی مدل داده ای تهیه شده بسیار دقت کنید. ارزیابی مدل داده ها تعیین می کند که آیا مدل ها بخوبی کار می کنند یا نه. در ارزیابی یک مدل باید به اینگونه سؤالات پاسخ داده شود:
  - صحت و دقت مدل چقدر است؟
  - مدل تا چه حد داده های مشاهده شده را به خوبی توصیف می کند؟
  - به پیش بینی های مدل چقدر می توان اعتماد داشت؟
  - مدل تا چه حد قابل فهم است؟